Inferring The Binding Preferences of RNA-Binding Proteins

by

Hilal Kazan

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Copyright © 2012 by Hilal Kazan

Abstract

Inferring The Binding Preferences of RNA-Binding Proteins

Hilal Kazan

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

2012

Post-transcriptional regulation is carried out by RNA-binding proteins (RBPs) that bind to specific RNA molecules and control their processing, localization, stability and degradation. Experimental studies have successfully identified RNA targets associated with specific RBPs. However, because the locations of the binding sites within the targets are unknown and because RBPs recognize both sequence and structure elements in their binding sites, identification of RBP binding preferences from these data remains challenging.

The unifying theme of this thesis is to identify RBP binding preferences from experimental data. First, we propose a protocol to design a complex RNA pool that represents diverse sets of sequence and structure elements to be used in an *in vitro* assay to efficiently measure RBP binding preferences. This design has been implemented in the RNAcompete method, and applied genome-wide to human and Drosophila RBPs. We show that RNAcompete-derived motifs are consistent with established binding preferences.

We developed two computational models to learn binding preferences of RBPs from large-scale data. Our first model, RNAcontext uses a novel representation of secondary structure to infer both sequence and structure preferences of RBPs, and is optimized for use with *in vitro* binding data on short RNA sequences. We show that including structure information improves the prediction accuracy significantly. Our second model, MaLaRKey, extends RNAcontext to fit motif models to sequences of arbitrary length,

and to incorporate a richer set of structure features to better model *in vivo* RNA secondary structure. We demonstrate that MaLaRKey infers detailed binding models that accurately predict binding of full-length transcripts.

Acknowledgements

I would like to thank:

Quaid Morris, for all his support and guidance, for reminding me that research is fun and exciting in every single meeting we had, and simply for being the ideal supervisor.

Michael Brudno and *Brendan Frey*, for serving on my supervisory committee and for giving me useful feedback.

Harmen Bussemaker and *Philip Kim*, for serving on my examination committee.

Timothy Hughes, *Debashish Ray*, *Kate Cook* and *Esther Chan*, for collaboration on RNAcompete project.

Past and present members of the Morrislab including *Amit*, *Anna*, *Gerald*, *Hossein*, *Sara*, *Xiao*, and *Wei*, for valuable discussions, and *Shankar*, for accompanying me at coffee breaks.

Family and friends, for their support and for reminding me that life is not all about work.

Ender, for his support, love and encouragement, and for waiting for all these years patiently on the other side of the world.

Contents

1 Introduction **1**

2 Background and Literature Review **5**

2.1 RNA-binding proteins and their roles 5

2.2 Binding mechanisms of RBPs 7

2.3 Experimental methods to identify targets of RBPs 8

2.3.1 *In vitro* experiments 9

2.3.2 *In vivo* experiments 10

2.4 Limitations of current experimental methods 11

2.5 Computational identification of binding sites 11

2.6 Review of motif finding 12

2.6.1 Word frequency motif models 13

2.6.2 Affinity-based motif models 14

2.6.3 Estimating sequence affinity from word affinity 15

2.7 RNA secondary structure 18

2.8 Related work in RNA-motif finding 23

2.8.1 Hackermuller-Stadler model 23

2.8.2 Extension of MEME for RNA motifs: MEMERIS 23

2.8.3 Covariance Models 24

2.8.4 Extension of MatrixREDUCE for RNA motifs: StructRED 26

3 Engineering the RNAcompete array **27**

3.1 Experimental protocol 28

3.2 RNA pool design v1 30

3.2.1 Weakly structured probes 30

3.2.2 Stem-loops 38

3.3 RNA pool design v2 40

3.4 Normalization 44

3.5 Analysis of binding specificities from v1 data 44

3.6 Analysis of binding specificities from v2 data 46

3.7 Discussion 51

4 RNAcontext: motif models for RNA-binding proteins **53**

4.1 Methods 55

4.1.1 Structural annotation of RNA sequences 55

4.1.2 Motif model 56

4.1.3 Parameter estimation 59

4.2 Results 60

4.2.1 Dataset 60

4.2.2 Justification of choice of motif models for comparison 62

4.2.3 Fitting motif models 63

4.2.4 Performance evaluation 65

4.2.5 The predictive value of structural context 68

4.2.6 Position weight matrices provide good approximations of sequence binding preferences for six RBPs 69

4.2.7 The sequence and structure binding preferences for seven RBPs . 71

4.2.8 *In vivo* confirmation of RNAcontext motif for SF2/ASF 75

4.3 Discussion 75

5 MaLaRKey: detailed in vivo binding models **78**

5.1 Methods 79

5.1.1 Overview of input and output of MaLaRKey 79

5.1.2 Structural annotation of RNA sequences 80

5.1.3 Motif model 81

5.1.4 Feature design 82

5.1.5 Parameter estimation 83

5.2 Results 85

5.2.1 Comparison of RNAcontext and MaLaRKey 85

5.2.2 RIP-chip data 87

5.2.3 CLIP data 98

5.3 Discussion 104

6 Conclusions and Future Work **107**

6.1 Future directions 110

Bibliography **113**

List of Tables

4.1 The composition of Sets A and B in terms of relative proportions of stem-loops and weakly structured sequences 62
4.2 Details about the chosen models for RNAcontext, MEMERIS and MatrixREDUCE 66
4.3 Comparison of predictive accuracy of three motif finding models using both weakly structured and stem-loop sequences in the test set 67
4.4 Bootstrap analysis of relative AU-PRs 67

5.1 Comparison of predictive accuracy of RNAcontext and MaLaRKey on nine RBPs from RNAcompete-v1 data 87
5.2 The number of bound and unbound transcripts for 14 RBPs from RIP-chip data 89
5.3 P-values for Wilcoxon sign rank test on 100 AU-ROC and AU-PR pairs . 92
5.4 P-values for Wilcoxon sign rank test on 100 AU-ROC and AU-PR pairs . 101

List of Figures

2.1 Roles of RBPs in various post-transcriptional processes 6

2.2 X-ray crystal structures of two RBP-RNA interactions 8

2.3 An RNA secondary structure example 20

2.4 Comparison of MEME and MEMERIS results for the PIE Rfam (RF00460) data set 24

3.1 Outline of the RNAcompete method 29

3.2 Generation of a de Bruijn sequence over the RNA alphabet 32

3.3 Splitting strongly structured probes 34

3.4 Flow chart describing the RNA pool design for RNAcompete-v1 37

3.5 Flow chart describing the RNA pool design for RNAcompete-v2 43

3.6 Correlations between identical loop sequences in Set A and Set B 45

3.7 Correlations between robust average 7-mer scores 46

3.8 RNAcompete analysis of nine RBPs, representing four different classes of RNA-binding domains 47

3.9 Correlations between 7-mer scores from Set A and Set B for HuR and Vts1p 48

3.10 RNAcompete analysis of 15 RBPs from fly 49

3.11 RNAcompete analysis of 15 RBPs from human 50

4.1 Overview of RNAcontext method: the input, parameters and motif model 57

4.2 Change in predictive accuracy of RNAcontext due to the representation of RNA structural context 69
4.3 The predictive value of the fully-specified 7-mer model (i.e. the RNAcompete model) with respect to the motif models 70
4.4 Relative RNA structure preferences inferred by RNAcontext 72
4.5 Inferred sequence and structure binding preferences for seven RBPs . . . 73
4.6 Inferred sequence and structure binding preferences for FUSIP1 using all the sequences as the training set 74
4.7 Inferred sequence and structure binding preferences for SF2/ASF from *in vivo* data 76

5.1 Overview of MaLaRKey method: the input, parameters and motif model 83
5.2 Inferred sequence and structure preferences of Vts1p from RNAcompete-v1 data 88
5.3 Inferred sequence and structure preferences of yeast RBPs from RIP-chip data 91
5.4 Distribution of structure parameters inferred from bootstrapped training sets 95
5.5 Distribution of structure parameters inferred from bootstrapped training sets 96
5.6 Mean profile values for the top k-mer occurrences in bound and unbound sequences for RBPs a) Ypl184 b) Scp160 97
5.7 Inferred sequence and structure preferences of Vts1p from RIP-chip data 98
5.8 Inferred sequence and structure preferences of seven RBPs from CLIP data 100
5.9 Distribution of structure parameters inferred from bootstrapped training sets 102
5.10 Mean profile values for the top k-mer occurrences in bound and unbound sequences for RBPs a) Khd1 b) TIAL1. 103

Chapter 1

Introduction

Gene expression is regulated at multiple levels to precisely control the concentration of macromolecular components according to the physiological needs of the cell(s). Most of the research in the last decades has focused on the first layer of gene regulation, namely the transcriptional control mediated by transcription factors that bind to DNA and promote the initiation of transcription. Weak correlation observed between the number of transcripts and protein products of a gene indicates that the control of the post-transcriptional steps has a substantial regulatory impact as well [1, 2]. Post-transcriptional regulation (PTR) is the control of gene expression at the mRNA level, between transcription and translation. PTR controls all aspects of RNA processing including splicing, stability, localization and degradation, and is mediated by numerous factors with major players being the RNA-binding proteins (RBPs).

RBPs contain one or more RNA-binding domains that recognize specific sequence and structure elements in short regions of RNA. Eukaryotic genomes encode for hundreds of RBPs, and each RBP binds to a set of unique targets. Combinatorial binding of multiple RBPs to functionally related mRNAs suggests the existence of a multi-dimensional network that dynamically coordinates the fates of mRNAs in response to various conditions [3–8].

The importance of RBPs is further emphasized by their involvement in a variety of diseases including developmental and cellular defects, as well as cancer [9]. For example, Fragile X syndrome is caused by the loss of function mutation in the 5'UTR of the RBP Fmr1 [10]. Besides the mutations of RBPs, mutations in cis-acting elements that correspond to RBP binding sites can also lead to disease. Most prominent are point-mutations that cause splicing defects [11]. A striking example is the mutation in the exonic splicing signal of the SMN2 gene that disrupts the binding of the splicing factor SF2/ASF, which in turn causes spinal muscular atrophy [12].

Detailed understanding of the role of RBPs in disease, and predicting the role of disease-associated polymorphisms, requires the identification of cis-acting elements in the transcripts to which RBPs bind. However, targets of most RBPs remain uncharacterized. Experimental methods developed in the past decade promise to fill this gap by expanding our knowledge of RBP binding targets using *in vivo* or *in vitro* techniques. *In vivo* methods have the potential to identify the genome-wide targets of the RBPs; however, some of these methods require RBP-specific antibodies, and are restricted in the cell types that they can query binding. *In vitro* methods isolate the RBP-RNA interactions in non-biological conditions; though, they have the advantage of providing a complete characterization of the sequences and/or structures that an RBP can, or cannot bind. Also, they can determine how disease-associated polymorphisms will impact the binding of RBPs. However, the power of these methods strongly depends on the design of complex RNA libraries.

Experimental methods provide noisy measurements of RBP affinity to several sequences, some of which contain one or more binding sites at unknown locations. Identification of binding preferences from such data requires the design of computational models that can simultaneously infer the location and representations of the binding sites. Several computational models have been developed addressing a similar problem: identification of binding preferences of transcription factors (TFs). However, RNA sec-

ondary structure, a key factor that is recognized by RBPs but not TFs, is ignored in these models. As such, novel RBP-specific computational models are needed to elucidate RBP binding preferences.

This thesis consists of two parts that relate to the overall goal of identifying RBP binding preferences from experimental data. In the first part, we aim to design an optimized RNA pool for use in an *in vitro* tool for rapid characterization of RBP binding sites. In particular, our pool design allows us to efficiently query the entire space of possible k-mers (i.e. "k" base RNA sequences) as well as selected elements of RNA secondary structure. An important contribution of our design is the consideration of RNA secondary structure, which is ignored in most of the current pool designs. Our design has been experimentally implemented and applied to several RBPs across different organisms. The analysis of the output data shows that the method is able to systematically identify the sequence and structure preferences of RBPs.

In the second part of this thesis, we develop computational models to elucidate RBP binding preferences from experimental data. Existing models that are used to identify RBP binding preferences either ignore RNA secondary structure completely, or use limited representations of it. To address this absence, we introduce a novel representation of RNA secondary structure that takes into account the distribution of all possible secondary structures that a sequence can assume. This representation allows us to query general (i.e. paired or unpaired) or detailed (i.e. in a hairpin loop) types of structural contexts that a base is in. We incorporate this representation in a motif model, and show that including structure information improves the prediction accuracy for several RBPs. We also extend our initial model to query a much richer set of secondary structure features, and show how to construct detailed models of RBP binding.

This thesis is organized into six chapters as follows:

- Chapter 2 provides background on RBPs and experimental methods to query RBP targets. This chapter also gives a summary of existing motif models to identify RBP

binding preferences, introduces RNA secondary structure and the computational methods to predict it.

- Chapter 3 describes RNAcompete, an *in vitro* experimental method for rapid analysis of RBP recognition specificities with a pool containing diverse elements of sequence and structure. In particular, we explain how we designed a complex pool that conforms to several constraints defined by the experimental protocol.

- Chapter 4 describes RNAcontext, our initial work on identification of sequence and structure preferences from large-scale binding data. We introduce a novel representation of RNA secondary structure, and show that incorporating structural information improves the prediction accuracy for several RBPs.

- Chapter 5 describes MaLaRKey, extension of RNAcontext that incorporates a much richer set of secondary structure features, and that uses a motif scoring model suitable for input sequences of arbitrary length. We show that detailed models of RBP binding can be constructed with the use of these features.

- Chapter 6 summarizes our findings and provides some possible future directions for this work.

Chapter 2

Background and Literature Review

This chapter is intended to cover the main concepts required to understand later chapters, and to examine the relevant literature in the field. First, we give an introduction to RNA-binding proteins (RBPs) and the mechanisms by which RBPs recognize their targets. Since identification of RBP targets is crucial for understanding their function, several experimental methods have been developed to identify RBP targets. We continue with a summary of these methods, and also discuss their limitations. We explain why computational models are still needed to identify RBP binding preferences. Then, we provide background on RNA secondary structure: an important factor to consider in modeling RBP binding. Finally, we review existing computational models to identify binding preferences of RBPs.

2.1 RNA-binding proteins and their roles

In order to relay the information stored in DNA, portions of DNA (i.e. genes) are copied or transcribed to RNA molecules called messenger RNAs (mRNAs). Unlike DNA, which resides in the nucleus, mRNAs can move around the cell and serve as genetic messengers by carrying instructions that are necessary for protein synthesis (i.e. translation).

RNA binding proteins (RBPs) are key factors in regulation of gene expression at the

mRNA level, also called post-transcriptional regulation. These proteins bind to target mRNAs and control multiple aspects of RNA processing including splicing, stability, localization and degradation (Figure 2.1) [13]. Individual RBPs typically have hundreds of targets, and the combinatorial binding of multiple RBPs to mRNAs suggests the existence of a multi-dimensional network that dynamically coordinates the fates of mRNAs in response to various conditions [3–8].

Not surprisingly, alterations in the activity of RBPs are associated with many human diseases such as fragile X syndrome, muscular atrophies, breast and lung cancer [9]. However, the biological mechanisms behind these associations are still unclear as many RBPs have uncharacterized targets (and consequently functions). In the next section, we discuss how RBPs recognize their binding sites.

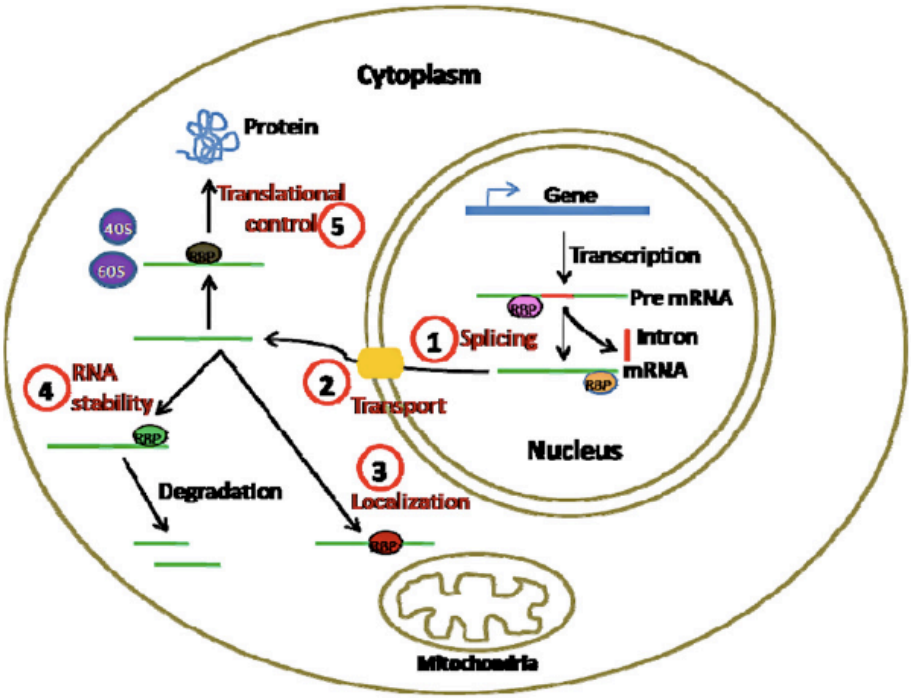


Figure 2.1: *Roles of RBPs in various post-transcriptional processes.* RBPs have critical roles in numerous cellular processes including splicing, transport, localization, stability and translation (shown with numbers 1-5) that take place in the nucleus and cytoplasm. Figure taken from [14].

2.2 Binding mechanisms of RBPs

RBPs bind to their targets through one or more RNA-binding domains (RBDs). There are about 40 different types of RBDs that are found to be involved in RNA binding [15]. Some of these domains are quite abundant and found in hundreds of proteins such as RNA recognition motif (RRM) and K homology domain (KH), whereas some others are seen only in specific organisms. The diverse activities of RBPs would suggest a similar diversity in their RBDs; however, the architecture of RBPs is more simple. Most RBPs are constructed from few individual domains that often recognize short RNA stretches with weak affinity. The versatility in affinity and specificity is achieved by combining multiple domains in variable number of copies and with different arrangements.

Unlike DNA-binding proteins, which recognize the sequence content of their binding sites, many RBPs recognize both the sequence and the secondary structure of their binding sites. This difference is because of the different helical configurations adopted by DNA and RNA. DNA assumes a B-form helix where the major groove is wide and easily accessible by proteins. In contrast to the B-form helix, the A-form helix formed by RNA has a major groove that is too deep and narrow for proteins to access and recognize the bases therein. As such, most RBPs recognize single-stranded regions or openings in double-stranded regions where the major groove has been widened by bulges or hairpins [16]. For example, an HIV-1 protein TAT binds to three-nucleotide (UCU) bulge loops with high affinity [17]. Puf3, a yeast RBP, contains a Pumilio-homology domain (PUM-HD or PUF) consisting of 8 PUF repeats that each recognize a single nucleotide of the motif UGUACAUA in single-stranded context (Figure 2.2-a) [18]. Another yeast RBP, Vts1p, binds to hairpins with the loop consensus sequence CNGG (N stands for any base) by means of its SAM domain (Figure 2.2-b)[19]. Staufen is a Drosophila RBP that binds to double-stranded regions without any sequence preference [20]. These examples show that some RBPs bind single-stranded RNA by direct readout of the primary sequence, whereas others recognize primarily the structure of the RNA or both the sequence and

the structure.

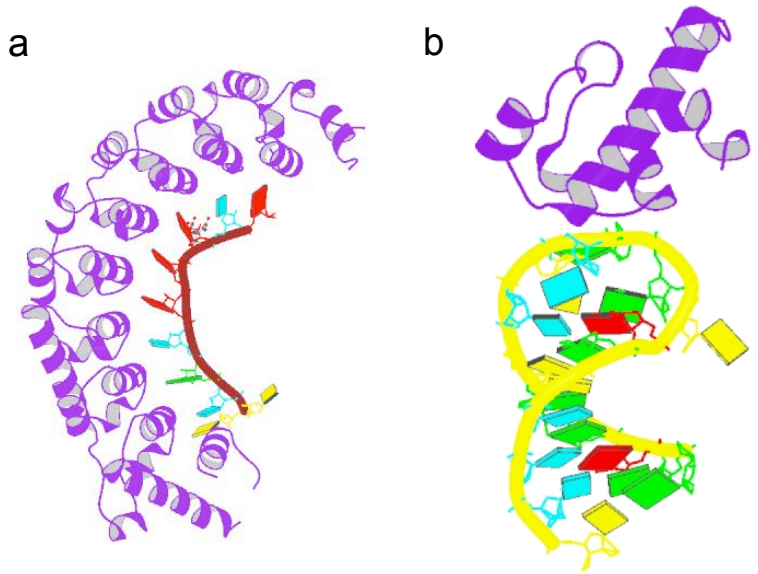


Figure 2.2: *X-ray crystal structures of two RBP-RNA interactions.* a) The interaction of Puf3 repeats (purple) with the RNA sequence CCUGUAAAUA (PDB 3K49)[18]. b) Structure of Vts1p SAM domain (purple) contacting to the RNA sequence CUGGC in a hairpin loop (PDB 2B6G) [19].

The most common families of RBDs, RRMs and KHs, bind a variety of sequence and structural contexts, and there is no known simple relationship between their amino acid sequence and the RNA sequences they recognize. Recognition codes are lacking for the vast majority of other RBDs as well. For this reason, several experimental methods have been developed to investigate the binding specificities of RBPs. In the next section, we give an overview of those that are relevant to this thesis.

2.3 Experimental methods to identify targets of RBPs

Experimental methods for detection of RBP binding sites can be classified into low-throughput and high-throughput techniques. Also, depending on where the experiment is carried out, these methods could be referred as *in vivo* (i.e. within the cell) or *in vitro* (i.e. not within the cell but in some controlled environment) studies. *In vivo* methods

have the advantage to query RBP-RNA interactions in biological conditions; however, they may not identify direct interactions, and require prior knowledge on the experimental conditions that the RBP is active. Also, some *in vivo* assays require engineering cells to generate RBP levels that are much higher than normal and/or require analysing RBP targets in cell cultures that are derived from the organism but are not representative of the organism (e.g. HeLa cells). *In vitro* methods identify RBP targets in non-biological conditions, but this also allows querying non-genomic sequences such as variants of the wild-type binding sites and testing a wide range of interesting conditions (e.g. salts, pH) that would be incompatible with traditional *in vivo* selection methods.

2.3.1 *In vitro* experiments

Selection of ligands by exponential enrichment

Selection of ligands by exponential enrichment (SELEX) is a low-throughput method for *in vitro* detection of RBP targets [21]. High-affinity binding sequences are selected from a randomized pool through multiple rounds of selection, purification and amplication. The resulting RNAs are cloned and sequenced, providing a set of high-affinity targets. Recently, SELEX is combined with massively parallel sequencing to analyze binding preferences of proteins in a high-throughput manner [22]. One disadvantage of SELEX assay is that it reveals only the highest affinity RNA targets, which do not necessarily reflect the physiological targets.

RNAcompete

RNAcompete is a high-throughput *in vitro* method that estimates relative binding affinities of selected RBPs to each RNA sequence in a defined population [23]. The experimental protocol consists of three steps: (i) generation of an RNA pool comprising each k-mer in a variety of structural contexts; (ii) a single pulldown of the RNAs bound to a tagged RBP of interest; and (iii) microarray analysis of the relative enrichment of each RNA in the bound fraction relative to the initial pool. The relative recovery of each

RNA in the pulldown RNA population as compared to the total pool signal provides a measure of binding preference. Chapter 3 describes the RNAcompete method in greater detail.

2.3.2 *In vivo* experiments

RNA immunoprecipitation

Genome-wide immunoprecipitation studies allow identification of *in vivo* targets of RBPs by querying cellular RNA. RNA immunoprecipitation (RIP) is a high-throughput method that purifies RBP-mRNA complexes from cellular extracts and identifies protein-bound mRNAs using either a microarray (RIP-chip) or high-throughput sequencing (RIP-seq) [24, 25]. Targets identified with this method should be interpreted carefully since the absence of a cross-linking step may allow RBPs to dissociate from their *in vivo* targets and re-associate with higher affinity targets after cell-lysis [26]. Therefore, this method may not measure the true *in vivo* interactions.

Cross linking and immunoprecipation

Cross linking and immunoprecipation (CLIP or HITS-CLIP) is proposed to eliminate the re-association artifact by incorporating a ultraviolet light (UV) cross-linking step before immunoprecipitation. This allows a more stringent washing procedure to reduce contaminants and eliminate interactions that occur after cell lysis [27]. Another advantage of CLIP is the detection of binding sites with a resolution as few as a few hundred nucleotides as opposed to some RIP methods that can provide information on only full-length transcripts. This is achieved by the ribonuclease digestion step that liberates the binding sites from full transcripts. PAR-CLIP, which stands for photoactivatable-ribonucleoside enhanced crosslinking and immunoprecipitation, is a recent variant of CLIP technique where cross-linked sites are enriched with a thymidine to cytidine transition [28] and this is utilized for locating the binding sites in higher resolution. iCLIP, another CLIP-based technique, utilizes the fact that reverse transcriptase stops at cross-linked sites to

determine binding sites in higher resolution [29].

2.4 Limitations of current experimental methods

Recent developments in experimental methods are quite promising; however, it is also important to realize their inherent limitations. To begin with, one of the most important limitations is the possibility of identifying false-positive targets. This could be due to non-specific binding and would be revealed as high-background effect in SELEX or RNAcompete. Similarly, for CLIP experiments, false-positive targets could be identified due to contamination with non-cross-linked sites. In RIP, false-positive targets might stem from the re-association of the RBP with *in vitro* targets after cell lysis. Another source of inaccuracy common to any of the mentioned experimental methods is the additional noise introduced by the downstream method to identify bound sequences (i.e. microarray or sequencing). Even when we ignore these limitations, identification of binding sites is still difficult because the locations of the binding sites within these sequences are not known. CLIP techniques can pinpoint binding sites within a couple hundred nucleotides resolution, whereas methods like RIP-Chip (and most RIP-seq methods) can only provide binding information for full-length transcripts. *In vitro* experimental methods (e.g. SELEX, RNAcompete) usually query short oligos (i.e. less than 50nts); however, the location of the binding site within the oligo is still unknown. Altogether, these limitations make the task of determining RBP binding sites challenging. In the next section, we discuss how these challenges are being addressed by computational methods.

2.5 Computational identification of binding sites

As described in the previous section, experimental methods provide noisy measurements of RBP affinity to several sequences, some of which contain one or more binding sites at unknown locations. Hence, the input to computational models is a set of unaligned

sequences and their associated observations. Computational models can be classified into two groups based on how they treat the input. Models in the first group define a subset of the sequences as bound (or unbound) based on a threshold and discard the rest of the data. Binding sites can then be determined using only bound sequences or using both bound and unbound sequences. Models in the second group adhere to the continuum of binding principle, and utilize all the available sequences with associated quantitative measurements.

Once the input is defined, the rest of the task amounts to finding patterns that explain the input data. Namely, identified patterns should be over-represented in the sequences that are likely to be bound. Since the locations of these patterns are not known, this task requires simultaneously inferring the location and representations of the binding sites. The former challenge can be handled either by treating the location of the binding site as a hidden variable or by scanning across all possible starting positions. For the latter, representations that can model the degeneracy of the binding sites have to be considered. Typically, motif models fit parameters for all possible bases at every position in the site. Also, unlike DNA motif finding models, models for RNA motifs have to consider the secondary structure of the binding sites.

In the next section, we present an extensive discussion of different representations of binding sites. Most of the discussed models are developed for DNA motif finding; however, they are still relevant as they form the basis for RNA motif models.

2.6 Review of motif finding

There are two major classes of motif model representations for sequence preferences of nucleic acid-binding (i.e. DNA-binding or RNA-binding) proteins (NBPs). The models in the first class (hereafter "word frequency (motif) models") represent a probability distribution over nucleic acid words. Models in the second class (hereafter "affinity-

based (motif) models") parameterize a function that assigns the relative binding affinity (or log binding affinity) of the protein for nucleic acid words.

Typically, algorithms that learn word frequency models take as input a set of sequences enriched for NBP binding sites and fit their models using maximum likelihood or penalized maximum likelihood methods. Though recently methods that use different point estimates of the parameter posterior have been developed [30, 31]. In contrast, most algorithms that learn affinity-based models directly also take as input a numerical value associated with each sequence. This value is interpreted as a measurement of the binding affinity of the NBP for the sequence. Typically, these types of models require examples of both bound and unbound sequences. These types of affinity data have become increasingly common [23, 32, 33] as have algorithms (including RNAcontext, see Chapter 4) to fit motifs using these values [34–39]. However, it is also possible to transform a word frequency model into an affinity-based model, as we describe below.

2.6.1 Word frequency motif models

A key step in fitting word frequency motif models is using the input sequences to generate counts of words that represent NBP binding sites; motif models are then fit to these counts. Often generating these word counts, or expected word counts, is done as part of procedure that iterates between attempting to locate the NBP binding sites within the input sequences and refining the fit of the motif model [31, 40].

Formally, let $\{\#\text{P-}s_1, \#\text{P-}s_2, \ldots, \#\text{P-}s_N\}$ be the counts (or expected counts) of how often various words $\{s_1, s_2, \ldots, s_N\}$ appear as binding sites of protein P in the input set. Let $Pr(s;\Theta)$ be a word frequency motif model where $Pr(s;\Theta)$ is a probability distribution over words s parameterized by Θ. Often, but not always, the support of $Pr(s;\Theta)$ is nucleic acid words of a fixed, pre-defined length K. The parameters Θ can

be learned by optimizing the fit to the empirical word probabilities, i.e.,

$$Pr(s_i;\Theta) \approx \frac{\#\text{P-}s_i}{\sum_{j=1}^{N} \#\text{P-}s_j}, \forall i. \tag{2.1}$$

Until recently, NBP binding preferences were estimated from a relatively small number of binding sites defined *in vivo* or through low-throughput *in vitro* selection procedures. This small number of observations was insufficient to reliably estimate $Pr(s_i;\Theta)$ for each word, and as such word probabilities for fixed-length motif models were estimated using a product-multinomial model, commonly known as a position frequency matrix (PFM), in which the distributions of bases at each position in the word are independent. In the PFM, $Pr(s_i;\Theta) = \prod_{k=1}^{K} \Theta_{k,s_i(k)}$ where $s_i(k)$ indicates the k-th base in word s_i. This model only contains a small number of parameters, $3K$, and its maximum likelihood estimate is easily found by setting $\Theta_{k,s_i(k)} = f_{k,s_i(k)}$ where $f_{k,s_i(k)}$ is the frequency of $s_i(k)$ at position k. However, because PFM models are inaccurate representations of transcription factor (TF) binding affinity [41, 42], a variety of more complex probability distributions have been developed to model interactions between bases [43–47] or variable-spacing between binding sites of obligatory heterodimers like bZIP or bHLH proteins [48].

2.6.2 Affinity-based motif models

These models are based on physical principles of protein-ligand interactions. In particular, consider the equilibrium reaction of binding of a protein P to the DNA or RNA word s:

$$\text{P} + \text{s} \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} \text{P-}s \tag{2.2}$$

where k_{on} and k_{off} represents the protein binding and dissociation rates respectively. The binding affinity of the protein for s can be expressed in terms of its equilibrium

constant $K_a(s)$:

$$K_a(s) = \frac{[\text{P-}s]}{[\text{P}][s]} = \frac{k_{on}}{k_{off}} \tag{2.3}$$

where [P], $[s]$, [P-s] correspond to the concentrations of the unbound protein, unbound s, and the protein in complex with s, respectively. Affinity-based motif finding methods fit parameters Ω of their motif models $W(s;\Omega)$ by trying to match their affinity estimates for a given word s those implied by the input, i.e. $W(s;\Omega) \approx K_a(s)$.

Note that, in addition to $K_a(s)$, motif models in this class have also been designed to estimate a number of other measures of binding affinity, e.g. the dissociation constant $K_d(s) = K_a(s)^{-1}$, the log binding affinity $\log K_a(s)$, or the relative binding affinity $CK_a(s)$ up to an unknown constant C that is independent of s [36, 49]. The popular position weight matrix (PWM) [50, 51] or the position-specific affinity matrix (PSAM) [52] are examples of these types of models. Note also, that one can derive an estimate of relative binding affinity $W(s_i;\Theta)$ implied by a word frequency model, $Pr(s_i;\Theta)$ by dividing the latter by $[s_i]$, i.e.

$$W(s_i;\Theta) = \frac{Pr(s_i;\Theta)}{[s_i]} \approx \frac{\left(\#\text{P-}s_i \Big/ \sum_{j=1}^{N} \#\text{P-}s_j\right)}{[s_i]} \propto K_a(s_i) \tag{2.4}$$

For example, MotifRegressor [34] uses this approach, approximating $[s_i]$ with a third-order Markov model trained on background sequence, to translate word frequency models fit by MDScan [53] into affinity-based models suitable for scoring sequences.

2.6.3 Estimating sequence affinity from word affinity

As mentioned in the previous sections, rarely do the sequences input into motif finding algorithms consist of delineated binding sites. Furthermore, the input sequences "enriched" for binding sites can contain more than one binding site, or possibly none at all. As such, an important component of any motif finding procedure is a sequence scoring

function that takes as input the probabilities or affinities assigned to each word by the motif model and outputs a "score" for the entire sequence that reflects the number of likely NBP binding sites therein and their strength.

Word frequency motif models are usually paired with probabilistic generative models of sequences. The "score" computed by these generative models for an arbitrary sequence is the probability of generating the sequence under the model. MEME is a widely used DNA motif finding method of this kind [40]. A mixture model is used to generate a sequence from a combination of a motif model and a background model, where the former is parameterized by a multinomial distribution and the latter is parameterized with a vector of letter frequencies. Locations of the motifs are treated as hidden variables and a uniform prior distribution is often used for possible motif start positions. Though, these priors can be changed to incorporate other knowledge about the location of the binding site (see next paragraph), or to incorporate information about RNA accessibility (see Section 2.8.2). These hidden variables and the parameters of the motif model are simultaneously estimated using Expectation Maximization (EM) algorithm [54].

Further refinements to generative models of sequences employ, for example, hidden Markov models to model steric hindrances that prevent overlapping binding sites and/or to model clustering of binding sites within cis-regulatory modules. A good summary of recent work in this area is provided in [55]. One advantage to this approach is that it is easy to incorporate competition for NBP binding sites from, for example nucleosomes [56] or internal RNA secondary structure [57], by assessing a prior probability on possible NBP binding sites according to the strength of competition for the site. A disadvantage to this approach is that the physical interpretation of these generative probabilities becomes difficult when there are multiple binding sites within a sequence.

There remains some controversy about the best approach for scoring sequences using affinity-based motif models. Early algorithms (e.g. MotifRegressor [34], MatrixREDUCE [36]) used the sum of the affinities of each word in the sequence as an estimate of the

NBP affinity for the entire sequence. One criticism of this approach is that the number of proteins bound to the sequence (also called the "occupancy" of the sequence) also depends on the number of proteins initially available for binding. So, if the initial protein concentration is low and the sequence has many high affinity binding sites, the actual occupancy of the sequence can be much lower than its potential occupancy implied by the estimated affinity. To address these concerns, some sequence scoring functions use affinity-based motif models to compute a function $N(s)$, the "occupancy" of word s [58–60], that also considers the initial concentration of proteins available for binding. This occupancy, which represents the proportion of words s bound by the protein, can be expressed as follows:

$$N(s) = \frac{[\text{P-}s]}{[\text{P-}s] + [s]}. \tag{2.5}$$

Note that dividing both sides with [P-s] gives:

$$N(s) = \frac{1}{1 + \frac{[s]}{[\text{P-}s]}} \tag{2.6}$$

Now we can rewrite the occupancy in terms of $K_d(S)$ since $\frac{[S]}{[\text{P-}s]}$ is equal to $\frac{K_d(S)}{[\text{P}]}$:

$$N(s) = \frac{1}{1 + \frac{K_d(s)}{[\text{P}]}} \tag{2.7}$$

which is equal to:

$$N(s) = \frac{1}{1 + \exp(\log K_d(s)) - \log[\text{P}])} \tag{2.8}$$

So, given an affinity-based motif model $W(s;\Omega)$ of binding affinity $K_a(s) = K_d(s)^{-1}$, measured under the same conditions, one can calculate an estimate $\hat{N}(s)$ of occupancy as follows:

$$\hat{N}(s) = \frac{1}{1 + \exp(-\log W(s;\Omega) - \log[\text{P}])}. \tag{2.9}$$

When fitting an affinity-based model using occupancy-based scoring, one can represent the often unknown constant log[P] with a bias β and train it while fitting the model. Note also, that one can adapt an existing affinity-based motif model $\tilde{W}$ for an NBP to predict occupancy under different experimental conditions (i.e. a change in temperature) by also introducing a scale α, so that the final model becomes:

$$\hat{N}(s) = \frac{1}{1 + \exp(-\alpha \log \tilde{W}(s;\Omega) - \beta)}. \tag{2.10}$$

As we will describe in Chapter 4, in RNAcontext, we replace $\alpha \log \tilde{W}(s;\Omega) \approx \log K_d(S)$ with $\sum_{k=1}^{K} \Theta_{k,s(k)}$ where Θ represents the model parameters. So that we write equation 2.10 as follows:

$$N^{\text{seq}}(s) = \sigma(\sum_{k=1}^{K} \Theta_{k,s(k)} + \text{bias}) \tag{2.11}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function.

Occupancy-based sequence scoring functions include those that simply sum the occupancy of all words in the sequence [58], those that calculate the probability that at least one site in the sequence is bound using the "noisy-OR" function [60], and more complex schemes that consider competitive and cooperative binding [60, 61].

2.7 RNA secondary structure

A single-stranded RNA molecule can fold back onto itself forming a secondary structure. After the formation of secondary structure, higher order interactions can occur to form the tertiary structure. The secondary or tertiary structure of an RNA can be determined using experimental techniques including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryo-electron microscopy. However, these methods are difficult and time-consuming. Also, structures of many RNAs cannot be characterized by these methods because of their conformational flexibility or large size. Recently, advances

in RNA structure analysis by chemical probing increased the accuracy and the throughput of probing experiments [62–64]. The idea is to treat the RNA of interest (or RNA pool) with a chemical reagent that leaves an imprint on the RNA molecule (or ensemble of RNA molecules). These imprints can then be used to reveal which nucleotides are in double-stranded (i.e. paired) or single-stranded (i.e. unpaired) conformation. These experiments, however, only report partial information of the structure.

Computational prediction methods have provided a valuable alternative to the discovery of RNA secondary structure. Most of the existing computational methods only aim to predict secondary structure since less is known about higher-order interactions that form the tertiary structure. This simplification is supported by the fact that the interactions that govern the secondary structure are stronger than tertiary interactions, and can be determined independently [65].

RNA secondary structure can be represented formally by assigning an index to each base in the RNA sequence. RNA molecule has a chemical orientation: the left end of the sequence is the 5' end and the right end of the sequence is the 3' end. The bases of an RNA molecule can be indexed starting from the 5' end and proceeding towards the 3'end. If we assume that the indexing starts from 1 and the length of the sequence is N, a secondary structure S can be defined as a set of pairs $i - j$, $1 \leq i < j \leq N$ where each base is only paired with at most one other base. Most common base pairs are the Watson-Crick pairings A-U, G-C and the wobble pairing G-U. An example secondary structure is shown in Figure 2.3. A stack of base pairs is called a stem. Unpaired regions that are enclosed by one or more base pairs are called loops. Loops can be annotated as hairpin loops, internal or bulge loops, or multi-loops depending on the number of closing base pairs. A pseudoknot, which is not shown in the figure, is a structural motif that involves non-nested, crossing base pairs. Most of the existing models ignore pseudoknots because they make the computational task much more difficult due to the violation of nestedness.

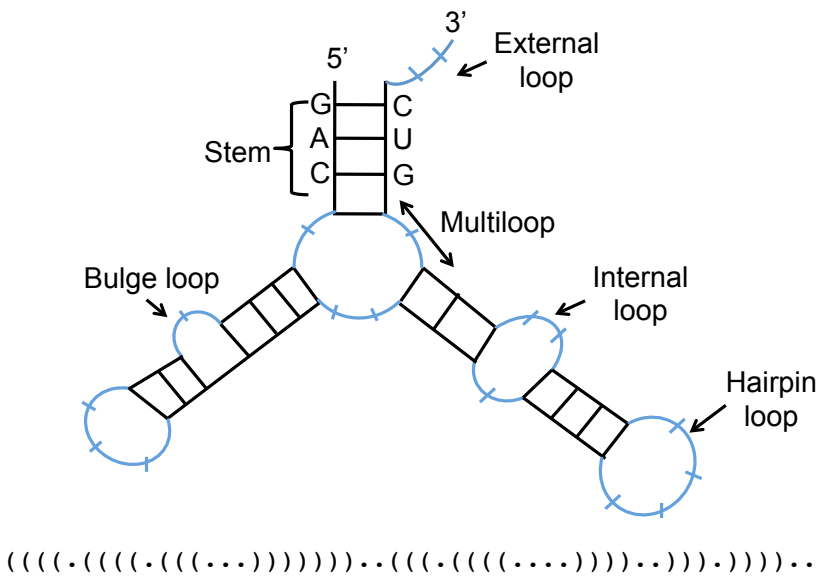


Figure 2.3: *An RNA secondary structure example.* One or more consecutive base pairs form a stem. Unpaired bases form various kinds of loops (shown in blue) which are classified according to the number of neighboring stems; hairpin loop, when there is a single closing base pair, bulge or internal loop, when there are two closing base pairs, multi-loop when there are more than two closing base pairs and, external loop if the unpaired bases are at the ends of the sequence. The same structure is shown with a dot-parenthesis notation on the bottom of the figure. In this notation, a matching pair of parentheses indicates a base pair and a dot denotes an unpaired base.

Algorithms for RNA secondary structure prediction fall into three general categories: (i) comparative methods that infer the secondary structure common to aligned sequences across conserved species; (ii) thermodynamic methods that predict the structure with the minimum free energy (MFE); and (iii) probabilistic models that use parameters learned from known structures to predict unknown structures. The comparative approach is the most accurate method when several aligned homologous RNA sequences are available [66]. The idea is to find base pairs that covary across the aligned sequences (e.g. compensatory mutations). Most of these methods use aligned sequences to infer a consensus secondary structure using a combination of energy-based scores and covariation scores [67, 68]. These methods are limited to data sets where the sequences are similar enough to get a reliable alignment yet divergent enough to find covarying base pairs. If this is not the case, "Sankoff algorithm" can be used to simultaneously align and infer a consensus

structure; however, this algorithm requires extreme amounts of memory and time [69]. Current approaches that employ simultaneous folding and alignment either limit the size or shape of substructures, (SLASH [70], Dynalign [71], pmcomp [72]) or exploit the sparsity in probable base pairings and aligned nucleotides (RAF [73]).

The second class of methods use a set of thermodynamic parameters that predict the folding free energy of a given structure and a dynamic programming algorithm to find the structure with minimum free energy [74, 75]. However, since thermodynamic parameters have substantial uncertainities and an RNA molecule can fold into multiple structures during its lifetime, the predicted MFE structure may not be the real, functional structure. To address the latter concern, instead of a single structure, some methods consider the distribution of possible structures. RNAshapes is one such method that can be used to enumerate secondary structures within a certain energy range of the MFE [76]. The probability of an individual secondary structure s can then be calculated as $e^{-E(s)/RT}/Z$ where $E(s)$ is the free energy of s, R is the gas constant, T is the absolute temperature, and Z is the partition function that represents the Boltzmann ensemble of all possible structures. RNAshapes can also be used to enumerate "shapes" (i.e. secondary structure motifs) that represent the topology of the secondary structure at various levels of specificity. For example, the structure `..((((((.((((....))))).....)))))).((((((....))))))...` can be represented with the bracket notation [ ] at the most abstract level and with _ [_ [ ] _] _ [ ] _ at the least abstract level, where [ ] denotes a stem and _ denotes unpaired regions. The topological properties of the ensemble can be characterized by calculating the probability of individual shapes, which is defined as the sum of the probabilities of all structures that fall into that shape. Sfold represents the distribution of possible structures by sampling structures from the Boltzmann ensemble of RNA secondary structures (i.e. partition function) [77]. RNAfold is another method of this kind [78–80], and computes probabilities for every possible base pair from the partition function. RNAplfold is a local folding

variant of RNAfold, and only base pairs within a certain span are considered possible [81]. In this approach, average base pair probabilities are calculated by averaging over all windows of a certain length that contain the pair. RNAplfold is shown to perform more accurate than the classic global folding algorithms in predicting mRNA secondary structure [82]. This result is in agreement with previous studies supporting local folding of mRNA [83, 84].

Models in the last category, probabilistic models, use parameters estimated from known RNA structures to predict secondary structure. CONTRAfold is the first probabilistic method that significantly outperforms existing physics-based approaches [85]. Contextfold is based on the same idea but fits a much larger set of features [86]. SimFold combines both thermodynamic parameters and probabilistically estimated parameters [87]. These methods are becoming more powerful due to the increasing availability of known RNA structures. Also, a number of methods have been proposed to combine the probing data with thermodynamic parameters or comparative analysis [88–90]. These methods improve the accuracy over computational predictions alone; however, currently they are limited to existing probing data.

The set of sequences that are queried for identification of RBP targets are quite diverse. *In vitro* experiments query random or specifically designed oligos and *in vivo* methods query arbitrary sets of transcripts. Hence, comparative methods are not suitable to determine the secondary structure of these sequences. In the rest of this thesis, we use thermodynamic methods that consider the ensemble of all possible structures to predict secondary structure as these are applicable to all the data sets (i.e. *in vitro* and *in vivo* for any organism) that we will consider.

2.8 Related work in RNA-motif finding

In section 2.6 we gave an overview of motif representations. In this section, we focus on methods that are specifically designed for RNA motifs.

2.8.1 Hackermuller-Stadler model

Hackermuller et al. developed the first quantitative model to measure the effect of RNA secondary structure on RBP-RNA binding [91]. This model predicts that the binding affinity of an RBP for a RNA sequence containing a single potential binding site is equal to the product of the probability that the site is in the preferred context and the affinity of the RBP for the site in the preferred context. The probability of the binding site to be in some particular structural context is calculated by considering the ensemble of possible secondary structures that the RNA sequence can form. They demonstrated that this model is a good fit to *in vitro* affinity data for HuR to sequences with differing levels of accessibility of the HuR binding site [92]. However, this model cannot be used for *de novo* motif finding as it assumes that the preferences of the RBP are already known.

2.8.2 Extension of MEME for RNA motifs: MEMERIS

MEMERIS [57] is an RNA motif model that extends MEME by integrating RNA secondary structure information to the motif search. As a preprocessing step, it precomputes for each word (i.e. kmer) the probability that the word is in single stranded context and then uses these values as priors on possible motif start positions. This adaptation changes the search so that motifs that are enriched in single stranded regions are preferentially found (See Figure 2.4 for an example). In fact, the score assigned to a k-mer with this model precisely implements the Hackermuller-Stadler model. It has been demonstrated that MEMERIS performs better than MEME in RNA motif finding on both artificial and *in vitro* data sets.

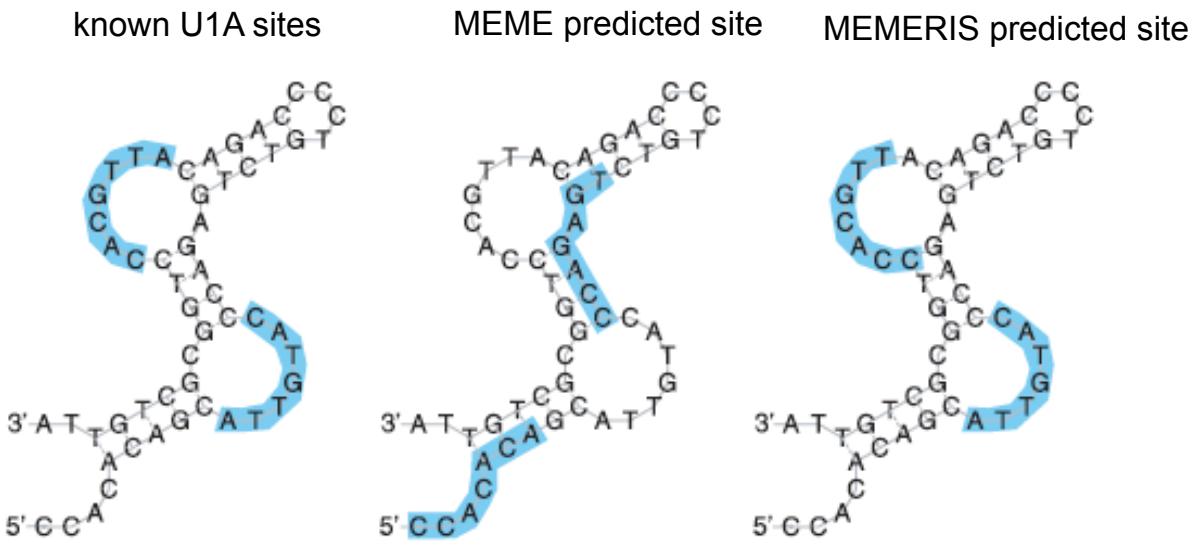


Figure 2.4: *Comparison of MEME and MEMERIS results for the PIE Rfam (RF00460) data set.* The figure on the left shows the consensus sequence and structure of the PIE RNA. The U1A protein binds the single-stranded sequences in the two asymmetrical internal loops in a cooperative manner. MEME identifies double stranded motifs that do not overlap with the real binding sites. On the other hand, MEMERIS is able to identify the correct binding sites as it preferentially searches for single stranded regions. (Figure taken from [57])

MEMERIS is shown to work well on *in vitro* data; however, global folding of RNAs (as computed by RNAfold) might limit its use on *in vivo* data. Moreover, MEMERIS cannot differentiate between different kinds of single stranded regions (e.g. hairpin loop or external loop).

2.8.3 Covariance Models

Covariance models (CMs) are probabilistic models that are originally proposed to model RNA families (e.g. tRNA, group I introns) and are also applicable for problems such as multiple RNA sequence alignment and consensus RNA structure prediction for a set of sequences [93]. CMs can be thought of as a generalization of Hidden Markov Models (HMMs) for modeling both primary sequence consensus and secondary structure of an RNA multiple sequence alignment. CMs are implemented in the software package COVE [93]. COVE would be the method of choice when the input sequences are expected to

have a good alignment and their consensus secondary structure is global. However, it may not be well suited for identification of RBP binding sites as most RBPs have binding sites ranging between 2-10 nts [15]. In other words, the problem of finding RBP sites is akin to searching for local RNA motifs in a set of long sequences (i.e. full length mRNAs or UTRs) with low sequence similarity and with variable flanking regions. CMfinder [94] is introduced to overcome some of the disadvantages of CMs for finding local motifs. It is a combination of the DNA-motif finding algorithm MEME and CM-based RNA analysis tool COVE. Original CMs consider only the existence of possible pairings but not their thermodynamic stability, whereas CMfinder initially chooses motif candidates with thermodynamically stable structures and initializes a CM from these candidates. CMfinder had promising results in discovering RNA motifs from 19 families of noncoding RNAs from the Rfam database. However, CMfinder is not suitable for finding RBP motifs for several reasons. During the selection of candidate motifs, absence rather than the presence of stable secondary structure must be preferred as most RBPs prefer their binding sites to be in single stranded context. Also, the authors of CMfinder report that the candidate selection process becomes harder with longer flanking regions due to the larger number of stable local structures. Lastly, the smallest allowed motif size on CMfinder website is 15 [95]; however, most RBPs have much shorter binding sites.

RNApromo [96] is a recent CM-based model for finding local motifs. It first finds the structural elements that appear in a maximal number of input RNAs to initialize a CM-model, and then improves this model using the EM algorithm. RNApromo correctly predicts three human RNA motifs: HFD, IRE, SECIS. The method is further tested on predicting motifs from mRNA targets of the RBPs Puf5, Puf4, Puf3, Pub1 and SAM68. All predicted motifs from these data sets appear in a hairpin loop and some of the bases of the binding site are paired. This contradicts with solved co-crystal structures for Puf3p and Puf4p, which show that both proteins bind unstructured RNA [18, 97].

CM-based models have good results in discovering RNA motifs from Rfam families.

They are also the only possible ways to model RBP binding sites with complex structure (e.g. ADAR proteins [98]). However, it should be noted that CMs have a reported tendency to overpredict secondary structure [99]. Indeed, as mentioned above, CM-based motif models of Puf3p, Puf4p, and HuR predict that they preferentially bind RNA hairpins and contradict structural, *in vitro* and *in vivo* evidence that they bind unstructured single-stranded RNA.

2.8.4 Extension of MatrixREDUCE for RNA motifs: StructRED

Foat and Stormo introduced the StructRED method that extends MatrixREDUCE to find RNA cis-regulatory elements in mRNAs [100]. MatrixREDUCE is a biophysical (i.e. affinity-based) model of protein-DNA interaction that fits a PSAM whose elements predict relative affinity for each nucleotide. MatrixREDUCE uses all the available data rather than a subset pre-defined as bound or unbound. StructRED is highly similar to MatrixREDUCE except that it models loop sequences of a certain length to explain quantitative measurements of mRNAs. Stem loops are formed through Watson-Crick base pairs, but their thermodynamic stability is not taken into account. StructRED correctly recovers the known binding preferences of Vts1p in yeast and Smaug in flies, and discovers a number of RNA regulatory elements in human and flies. However, modeling only stem-loops restricts its use in finding preferences of single stranded RNA binding RBPs.

Chapter 3

Engineering the RNAcompete array

In this chapter, we present the RNAcompete method for rapid characterization of binding specificities of RBPs *in vitro* [23]. RNAcompete consists of three main steps: (i) generation of a custom-designed RNA pool containing short sequences; (ii) a single binding reaction to identify the RNAs bound by the tagged RBP of interest; and (iii) analysis of the microarray data to determine binding preferences of the RBP. Our contributions within this framework are to design the RNA pool and to analyze the microarray data to identify the binding specificities of the RBP.

There are two versions of the RNAcompete method which we will call RNAcompete-v1 [23] and RNAcompete-v2 (unpublished) hereafter. Besides the differences in experimental procedure, one of the key differences between the two versions is the representation of secondary structure elements in the RNA pool. RNAcompete-v1 (Agilent array design: AMADID # 022053) was an initial design to evaluate the protocol, and the pool contained both probes that are designed to fold into weakly structured sequences and probes that are designed to fold into stem-loops. However, in analyzing RNAcompete-v1 data on nine RBPs, we found very little binding to stem-loops and we could determine the structural preference of the stem-loop binding protein Vts1p using only weakly structured sequences. Therefore, to increase the coverage of all possible k-mers in the array, we redesigned the

pool so that all probes are likely to be weakly structured. Our collaborators [1] performed RNAcompete-v2 experiments on a large set of fly and human RBPs and our analysis of this data revealed binding preferences of many RBPs with unknown binding specificities.

The rest of the chapter is organized as follows. First, we give an overview of the experimental protocol in order to justify the design constraints that it imposes on the probe design. We then describe the design of the RNA pool (i.e. probe set) for RNAcompete-v1 and v2. Since many of the steps are common in between, we will only mention the differences from v1 when describing RNAcompete-v2 pool design. We then describe the normalization procedure of the microarray data. Finally, we will explain the computational analysis of normalized intensities to identify binding motifs, and show that the array design is successful.

3.1 Experimental protocol

RNAcompete experiment consists of the following steps: synthesis of the RNA pool, a single pulldown reaction with the RBP of interest, detection of bound RNA on the array (Figure 3.1). RNA pool is generated using a custom designed Agilent array. The Agilent 244K array is a microarray that contains 244,000 separate spots. Each spot contains millions of short oligonucleotides with the identical sequences and each sequence is attached to the microarray surface via a linker on its 5' end. Each of the 244K sequences starts with the same primer sequence and the remaining 27-35nt are complementary to one of the sequences in the RNA probe set, and so will bind to it by hybridization. We use the microarray as a template to generate an RNA pool as follows. First, primer extension is performed to make the probe sequences on the array double stranded. Then, a common double stranded DNA (dsDNA) linker is ligated to all double-stranded probes on the array. Next, single stranded DNA (ssDNA) containing the linker are stripped from

[1] Debashish Ray, Kate Cook, Timothy R Hughes

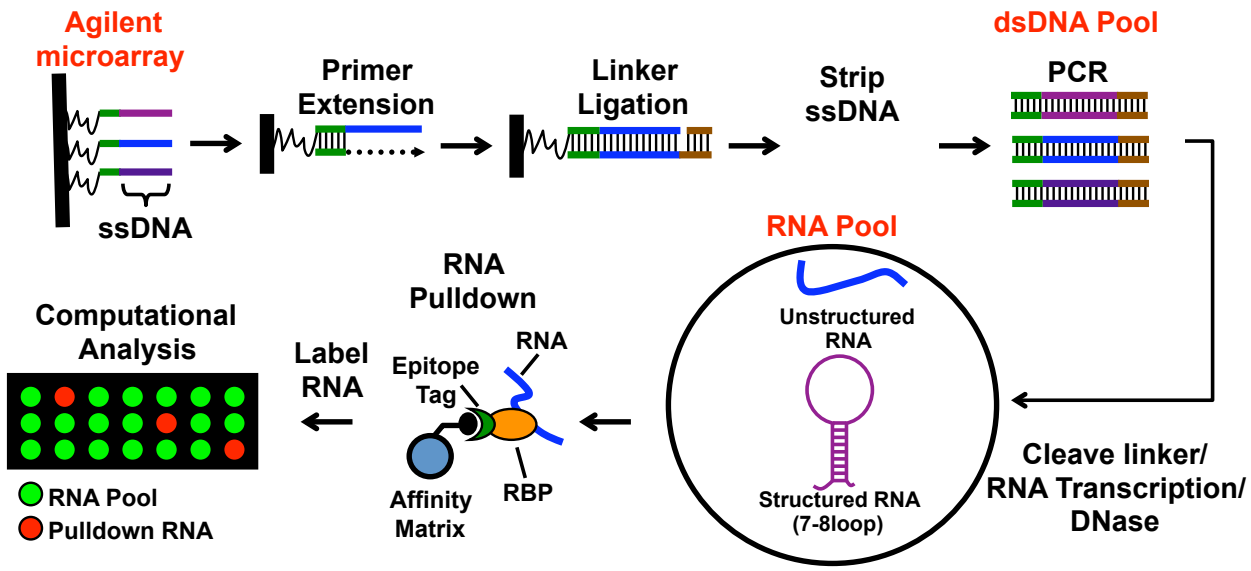


Figure 3.1: *Outline of the RNAcompete method.* Figure modified from [23].

the array and amplified using polymerase chain reaction (PCR). The linkers could be removed from the probes since they contain the Type IIS (Sap1) restriction site. Finally, the resulting dsDNA pool is used as a template for *in vitro* T7 RNA polymerase-mediated RNA transcription to generate the RNA pool. In the pulldown step, GST-tagged RBP of interest is incubated with a 75-fold molar excess of RNA pool. The recovered pulldown RNA is directly labelled with Cy5 and co-hybridized with Cy3 labelled pool RNA to an unused Agilent 244K microarray with the same design as used to generate the RNA pool. The relative recovery of each RNA in the pulldown is then measured and used as an estimate of binding affinity.

The experimental protocol imposed the following constraints for the pool design: (i) probes can only share a certain amount of primary sequence as otherwise cross-hybridization will occur on the microarray; (ii) probes cannot have long regions that are reverse-complementary as this would lead to RNA-RNA interactions in the pool; (iii) each probe must have the T7 initiator sequence (i.e. AGA or AGG) to get transcribed to RNA;

and (iv) probes cannot contain the Sap1 binding sites (i.e. GCUCUUC/GAAGAGC) as this would result in cutting at wrong positions. In the next section, we explain how we designed the RNA pool subject to these constraints.

3.2 RNA pool design v1

The pool contains two independent replicate designs (Set A and Set B) with different probe sequences, but each containing a nearly equal representation of sequence and structure elements. This allows us to evaluate the success of the experiment and also the performance of computational motif models on this data by training on one set of sequences and evaluating the performance on the other set. Each of the replicate designs is formed from a combination of two types of probes: those that are linear, non-folding sequences or fold into weak structures; or those fold into a single stem-loop. As required by the experimental design, we ensured that the probes do not contain the Sap1 restriction sites by removing or modifying (i.e. changing single bases) the probes that contain these sites.

3.2.1 Weakly structured probes

The sequences in this group are designed to satisfy two constraints: (i) each 8-mer is represented at least 16 times; and (ii) the probe sequences have to fold into weakly structures as defined based on their predicted free energies. We satisfied the first constraint using de Bruijn sequences, a well known approach for representing all possible k-mers on an array [101]. Here, we first give a brief introduction to de Bruijn sequences. We then describe how to generate them using linear feedback shift registers (LFSRs) by following the procedure from a previous study [101]. Lastly, we explain how we ensure that the resulting probe sequences fold into weak structures.

de Bruijn sequence

A de Bruijn sequence of order k is a circular string containing every k-mer from an alphabet Σ exactly once, and thus it is of length $|\Sigma|^k$. De Bruijn sequences represent all possible k-mers with a minimum length and also ensure that the distribution of n-mers, for all $n < k$, are uniform. We can compare de Bruijn sequences against randomly designed sequences. Namely, we can think of random sequence as obtained by sampling from the set of all possible k-mers with replacement. Let m be the number of all possible k-mers. Then, every k-mer has a probability $1 - (1 - 1/m)^m$ of being selected at least once within a sample of size m. The limit of this expression for large values of m is $1 - 1/e \approx 63.2$. Therefore, on average, only a 63.2% of all possible k-mers will be present in a random sequence, some of them appearing multiple times. This argument ignores the constraint that the k-mers have to overlap by assuming that they can be sampled independently. However, in our empirical tests, we found that it is a good approximation. Namely, when we generated five probe sets from random sequences, 36.8% of all possible 10-mers were missing on average.

There exist many de Bruijn sequences for a particular k and alphabet Σ [102]. We chose to use linear feedback shift registers (LFSRs) to generate de Bruijn sequences because of their desirable pseudo-randomness properties [103]. As background, let GF(p,n) represent a Galois field with order p^n where p is prime. GF(p,n) is a set containing p^n elements that is closed over multiplication and addition operations where arithmetic is performed modulo p. Elements of GF(p,n) can be generated using the roots of a primitive polynomial over that field. A primitive polynomial over GF(p,n) is an irreducible polynomial of degree n with coefficients taken from GF(p,n). A polynomial $p(x)$ over GF(p,n) is irreducible if it is not divisible by any polynomial over GF(p,n) of degree n', $0 < n' < n$.

In order to construct de Bruijn sequences with LFSRs, the size of the alphabet has to be a prime power. To satisfy this constraint for generating a de Bruijn sequence of

order k over the RNA alphabet, we first generated a de Bruijn sequence of order 2k over the Boolean field $A = \{0, 1\}$. That is, we recursively generated $S = (s_1 s_2 \ldots s_{2k-1})$ from the preceding 2k elements by the equation: $s_i = \theta_{k-1} s_{i-1} + \theta_{k-2} s_{i-2} + ... + \theta_0 s_{i-2k} \forall i$. If the coefficients θ is chosen so that the polynomial $\sum_{i=0}^{k-1} \theta_i x_i$ is primitive , the sequence S generated by this recursive equation will have periodicity $2^{2k} - 1$ and will contain every subsequence of length $2k$ over the Boolean alphabet except the sub-sequence that consists of 2k 0s. We inserted an additional 0 into one of the subsequences of $2k - 1$ 0s in S to convert it a de Bruijn sequence. Finally, we traversed S twice considering both reading frames (Figure 3.2) and transformed pairs of letters to the RNA alphabet A, C, G, U using the following embedding: $A \leftrightarrow (0, 0), C \leftrightarrow (0, 1), G \leftrightarrow (1, 0), U \leftrightarrow (1, 1)$.

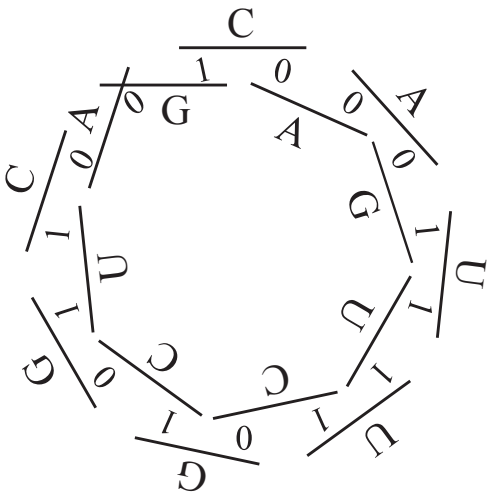


Figure 3.2: *Generation of a de Bruijn sequence over the RNA alphabet.* A de Bruijn sequence over the Boolean alphabet $\Sigma = \{0, 1\}$ can be transformed into an RNA sequence using the following embedding: $A \leftrightarrow (0, 0), C \leftrightarrow (0, 1), G \leftrightarrow (1, 0), U \leftrightarrow (1, 1)$ and considering both reading frames. Figure modified from [101].

We repeated the same process twice for each of the sets (i.e. Set A and Set B). Two primitive polynomials corresponding to LFSRs we used are [1]:

- $x^{20} + x^{19} + x^{18} + x^{17} + x^{16} + x^{15} + x^{13} + x^{11} + x^9 + x^8 + x^7 + x^6 + x^4 + 1$

[1]Primitive polynomials were taken from http://fchabaud.free.fr/English/Poly/)

- $x^{20}+x^{19}+x^{18}+x^{17}+x^{16}+x^{15}+x^{14}+x^{13}+x^{10}+x^{9}+x^{8}+x^{7}+x^{6}+x^{5}+x^{4}+x^{3}+x^{2}+x+1$

We then partitioned the two de Bruijn sequences into subsequences (i.e. probes or oligos) with sliding windows of 35nts and overlapping by 9nts to prevent losing any 10-mers. After this process, each probe set contained 40,330 oligos of length 35.

Checking the secondary structure of probes

Because our aim is to design a set of probe sequences that, when transcribed, would give rise to RNA oligos that were as linear (i.e. unstructured) as possible, we first identified those sequences that would give rise to RNA oligos with stable secondary structure. To do this, we used the existing RNA secondary structure prediction program RNAshapes [76] to enumerate all secondary structures that had predicted free energies within 70% of the minimum free energy (MFE structure) using the following call: RNAshapes -s -c 70.0 -r -M 30 -t 1 -o 2. Generating secondary structures within a larger percent of the MFE substantially slows down the program and does not increase the coverage of the full ensemble significantly. For each enumerated secondary structure, RNAshapes also outputs the probability of occurrence within the full ensemble. Using these probabilities, we classified as "strongly structured" all RNA oligos for which the sum of probabilities of all secondary structures with predicted free energies less than -3 kcal/mol is greater than 0.6. The free energy value -3 kcal /mol can be illustrated with an example: one of the predicted structures of the probe sequence AGAUAUGUAACUAGGUAUAAGACA***CGC***CACCG***GUG***GAG is `.....(((.....)))....` Note that I have italicized the paired bases in the RNA sequence for clarity. So, the constraint of -3 kcal / mol does not lead to highly structured sequences.

Designing weakly structured RNA sequences is difficult because 6 pairs (i.e. AU, UA, CG, GC, UG, GU) out of 16 can form a canonical base pair. To illustrate this empirically, we have generated five random probe sets where each probe set consists of 40,330 overlapping oligos of length 35nt. On average, 72% of oligos were strongly

structured. Interestingly, probes generated from de Bruijn sequence were less structured than those generated from random sequence: there were 26,429 (65%) and 26,385 (65%) strongly structured sequences in Set A and Set B, respectively. As we will describe in the following four subsections, we recombined subsequences of the strongly structured probes to produce less structured probes that had the same 8-mer content.

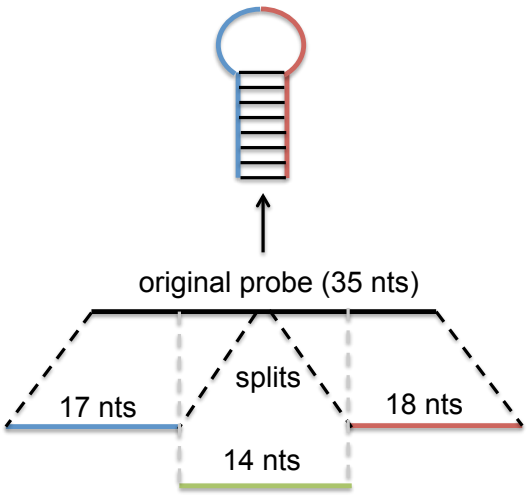


Figure 3.3: *Splitting strongly structured probes.* Probes that fold into strong secondary structures are split approximately in half (blue and red fragments). Fragments that belong to different strongly structured probes are then recombined to produce less structured probes. The green segment that spans the breakpoint (7 bases on each side) also needs to be recombined in order to recover the lost 8-mers in the middle.

Rearranging fragments of strongly structured sequences

We first split each of the strongly structured probe sequences into two fragments of length 17nts and 18nts which we will reassemble with other probe fragments (Figure 3.3). We split probes in half rather than at the point that is most disruptive to their structure, because it provides consistent lengths for the fragments, thus simplifying their reassembly, and typically disrupts their structure well enough to satisfy our design criteria. Compared to generating random probes, splitting the probes (generated from de Bruijn sequence) into two fragments of length 17-nt and 18-nt and then reassembling these together with the fragment in the middle (14 nts) allows us to capture all of the 10-mer content and gives

the flexibility to make unstructured probes while only increases the amount of sequence used by $(17 + 18 + 14)/35 = 40\%$. In comparison, when we increased the number of randomly generated probes by 40% (i.e. 40,330 to 56,462), 24.6% of all possible 10-mers were still missing on average.

After splitting, our next step was to pair the fragments belonging to different strongly structured sequences which, when attached, were weakly structured. We deemed a newly formed sequence (of length $34 = 17+17$, $35 = 17+18$ or $36 = 18+18$) weakly structured if the sum of probabilities of the predicted structured with free energies less than -2.5 kcal/mol was less than 0.5. Note that the criterion for weakly structuredness is slightly more strict compared to the initial structure check. We paired the fragments of strongly structured sequences using a greedy search, at the end of these search, we were left with 5,292 and 6,247 strongly structured sequences in Set A and B respectively.

Recovering lost 8-mers

This process of splitting and recombining probe sequences maintains most of the 8-mer content, except for the seven 8-mers spanning the middle of each of the original probes. We attempted to recover these lost 8-mers by forming additional probe sequences (of length 28) by merging the 14 bases fragments of the strongly structured probes that span the breakpoint (7 bases on either side). Again, we used a greedy algorithm to search for pairs of fragments that, when appended produced weakly structured sequences (as defined by the criterion mentioned above). Before merging, the two sets had 24,575 and 25,176 missing 14mers. After the merge, there remained 4,463 and 5,152 missing 14mers in Set A and Set B, respectively, which could not be combined into weakly structured sequences.

Merging remaining missing 14-mers with fragments of still structured sequences

We then attempted to merge the 14-mers with the previously unmergeable 17- and 18-nt fragments of the original set of strongly structured sequences to produce 31-nt (i.e. $14 + 17$) or 32-nt (i.e. $14 + 18$) weakly structured sequences. After this final redesign attempt, there remained 695 missing 14mers, 1,968 still structured sequences of length 35 and 2,880 still structured fragments of length 17 or 18 in Set A and 708 missing 14mers, 2,484 still structured sequences of length 35 and 3,082 still structured fragments of length 17 or 18 in Set B. We removed these sequences completely from consideration, determined the 8-mer distribution and attempted to design probes from scratch in order ensure that each 8-mer was represented at least 16 times.

Adding missing 8mers

We found that we needed to add 6,118 and 6,887 8-mers in total (often multiple copies of the same 8-mer) in order to ensure that each 8-mer was represented at least 16 times. We started by appending four missing 8-mers by forming a new sequence of length 32 provided that they are weakly structured. After finding all possible merges of four 8-mers, we tried to merge combinations of three missing 8-mers to produce a sequence of length 24 which we then padded out to a 27-mer (minimum allowed probe length) using three random bases in such a way that the probes remained weakly structured. Similarly after finding all possible merges of three 8-mers, we tried to merge two missing 8-mers and fill out the sequence so that it was 27-nt long and weakly structured. Finally, for the 8-mers that could not be merged, we attempted to design a linear sequence (of length 27) containing this missing 8-mer using the program RNAinverse [80] with the constraint that no other bases in the sequence outside of the 8-mer could pair. After the steps described above, all possible 8-mers were represented at least 16 times in each set and the sets had 52,259 and 52,219 weakly structured sequences, respectively.

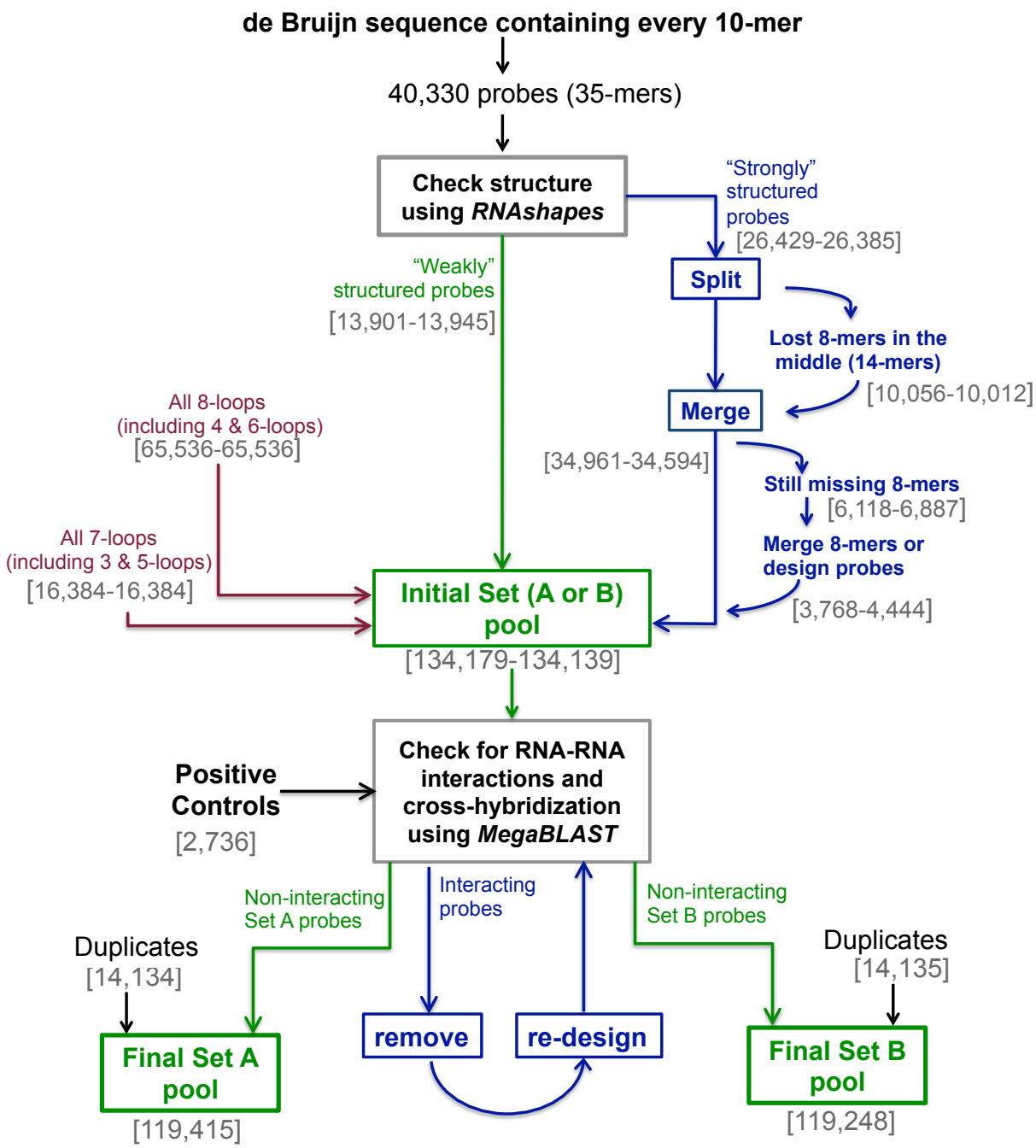


Figure 3.4: *Flow chart describing the RNA pool design for RNAcompete-v1.* The numbers in square brackets show the number of probes or n-mers during that step for Set A and Set B, respectively. Initial Set A or Set B pool consists of all possible 8-loops (65K), all possible 7-loops (16K), weakly structured de Bruijn probes (14K), and probes that are formed by merging fragments of strongly structured probes (35K), merging missing 8-mers or designing new probes to represent missing 8-mers (4K). Note that the numbers below "Lost 8-mers in the middle (14-mers)" are less than the number of strongly structured probes because they show the number of 14-mers that could be merged. Once the initial Set A or B pool is formed, MegaBLAST is used to check for cross-hybridization and RNA-RNA interaction. The probes that have matches to other probes are either removed or redesigned. Lastly, some of the probes in Set A and Set B are duplicated (14K), and control probes that contain the known binding sites for a set of RBPs (3K) are added to the final design.

3.2.2 Stem-loops

The stem loop probes sets within Set A and Set B were each designed to contain all possible 7-mers and 8-mers within a loop under the constraints that (i) the stems were all distinct, in order to minimize microarray cross-hybridization between different probes; (ii) the free energy of each stem was within a defined range; (iii) no stem was dominated by a particular base; and iv) that the expected stem-loop structure was very likely to form.

In this step, we generated two sets of stem-loop sequences each of which contained all possible 7mer and 8mer loops. First, we generated a set of possible stem sequence that satisfied a set of criteria designed to ensure that the stem was stable but unlikely to cross-hybridize either with other RNA oligos in the library (i.e. RNA-RNA interactions in the pool) or with other microarray probes. We generated all possible 10-base pair (bp) long stems in which the first and last pair were 5' G-C 3' to ensure tight bonds at either end of the stem. The remaining 8-bps were selected from the set of Watson-Crick and wobble pairs (i.e. A-U, U-A, G-C, C-G, G-U and U-G) such that the number of A-U and U-A base pairs is less than 5, the number of G-U and U-G base pairs is less than 3, the number of A-U, U-A, G-U and U-G base pairs is less than 6, and the number of G-C and C-G base pairs is less than 5. To this set, we applied the additional constraint that the stems have to have stacking free energy between -13kcal/mol and -20kcal/mol. Stems with lower free energies had G-C rich sequences and/or G-repeats. Because strong secondary structure and G-repeats are known to cause problems at the PCR stage, we opted to keep the stems with "strong" probability for stem formation but not "too" strong. Enumeration of all possible sequences satisfying these criteria led to 253,473 unique stems. We then matched all possible loops of 7-mers and 8-mers with the set of stems in such a way to avoid obvious complementary between the loop sequence and the stem sequence. In order to check for other structural features, we estimated the base pair probabilities for each probe sequence using SFOLD [77] . Using this information,

we identified the sequences that fold into structures with interfering bulges or internal loops in their stems. In order to replace these problematic stems, we generated a new stem set with the same constraints as described above except that the stacking free energy range is between -20.1kcal/mol and -20.8kcal/mol and then attempted to pair these new stems with the problematic loop sequences. Finally, RNAinverse [80] is used to design stem-loop sequences for the loops of the sequences that still fold into undesired structures after the stem replacement. At the end of this design process, we prepended the T7 initiator sequence AGA to the 5 end of each probe. We also added a set of control probes containing likely binding sites for each of the RBPs with known sequence binding preferences.

Removal of cross-hybridizing sequences

We then combined the probes in Set A and Set B into the same pool, and applied a series of BLAST-based filtering and redesign steps to remove probes likely to cross-hybridize. During this stage, our goal was to retain all control sequences, all loop sequences of size 7 or smaller, and ensure that each 8-mer was represented at least 12 times in unstructured probes in each of Set A and Set B. As such, we prioritized the removal of loops of size 8 which did not generate smaller loops (i.e. we would deem GAAAAAAC a loop of size 6). To this end, we ran MegaBLAST (version 2.2-18 with command line parameters -W 12 -D 3 - g -S 3) to identify to identify 81,774 sequences that had a match of at least 14 consecutive bases, or a match of at least 17 bases with at least 12 consecutive bases, to other sequences in either the forward orientation (which could lead to array cross-hybridization) or reverse orientation (which could lead to RNA-RNA interactions in the pool). We first removed the 28,532 8-base loop sequences among these matched sequences. We then used a greedy algorithm to remove the minimum number of probe sequences necessary to ensure that there were no more matches, except that when a non 8-base loop sequence aligned with an unstructured sequence, we would remove the loop

sequence and that we would only remove control probes when they aligned with other control probes. Through this process, we removed 20,137 loops with 7 or fewer bases, and 6,495 unstructured sequences. We attempted to replace these sequences as follows: (i) by changing the AGA T7 initiator to AGG; (ii) by swapping or replacing stems; and (iii) by replacing single bases in the middle of matching fragments of the unstructured sequences. We then removed any newly designed sequences still containing at least 14 consecutive base matches, checked the 8-mer distribution and added new sequences to represent the missing 8-mers. The final pool design contained 213,130 unique sequences. We duplicated 28,269 non-control probes so that the array contained 241,399 spots: 119,415 probes in Set A, 119,248 probes in Set B and 2,736 RBP binding control probes. Each of the sets contained at least 12 copies of all possible 8-mers, 59% of all stem-loops of length 8 and 99% of all stem-loops of length 7 and smaller.

3.3 RNA pool design v2

The RNA pool for v2 is very similar to v1 except that there are no stem-loops, and the procedure for converting strongly structured sequences to weakly structured sequences is simpler. By removing the stem-loop sequences, much more space was available for unstructured probes, as such, we started with a de Bruijn sequence of order 11 (generated using LFSRs with the primitive polynomial $x^{22}+x^{21}+x^{20}+x^{19}+x^{18}+x^{17}+x^{16}+x^{15}+x^{13}+x^{12}+x^{11}+x^{10}+x^{9}+x^{4}+x^{3}+x^{2}$) and then partitioned it into overlapping windows of length 35-nt as described before. This resulted in 167,773 probes. For each probe, we chose to prepend the T7 initiator (AGA or AGG) that forms a less strongly structured probe of length 38nt. The definition of structuredness, as before, is based on RNAshapes predictions (with arguments: -s -c 70.0 -r -M 30 -t 1 -o 2). If the sum of the probabilities of structures with free energies less than -2.5kcal/mol is larger than 0.5, then that probe is classified as strongly structured. Based on this, there were 130,936 strongly structured

probes and 36,837 weakly structured probes. The percentage of strongly structured sequences is higher than the v1 design (79% compared to 65%) because the definition for strongly structuredness is more strict and prepending the T7 initiator forms longer probes that have greater potential to form stable structures. In the v1 design, the T7 initiator was prepended after the structuredness assessment.

We tried a number of strategies to make the strongly structured probes less structured so that every 9-mer is represented in weakly structured context at least 16 times. First, we split each of the strongly structured probes into two equal fragments of length 19nt. We fixed the prefixes ($[1-19]$) of the probes and tried swapping the suffixes ($[20-38]$). This succeeded in forming 98,602 weakly structure probes, leaving 32,334 strongly structured. Then, we recombined the fragments $[4-19]$ and $[20-38]$ from two strongly structured probes, and prepended the T7 initiator sequence (i.e. the sequence [1-3]) that results in a less structured probe. This step produced 8,260 weakly structured probes. Finally, we merged 16-mers that span the breakpoints of strongly structured probes (8 bases on either side). We were able to merge 107,070 16-mers that resulted in 53,535 weakly structured probes. We combined all the weakly structured probes and calculated the distribution of 9-mer occurrences. For 65,723 9-mers (often multiple copies of the same 8-mer) that were represented less than 16 times, we attempted to increase the number of occurrences by merging four 9-mers or three 9-mers. For the 9-mers that did not form a weakly structured probe when merged, we designed probes that each contain one missing 9-mer using RNAinverse [80]. The final probe set contained 214,948 weakly structured probes after this step.

Similar to the v1 design, we would like to have two replicate sets for evaluation purposes. Therefore, we divided the probe set into two sets (i.e. Set A and Set B) such that the distribution of 9-mer occurrences is balanced across two sets. To do this, we first randomly assigned each probe to Set A or Set B, and then identified the 9-mers that have unbalanced distributions between Set A and Set B. For each such 9-mer, we

tried to find probes that would give a more balanced distribution when swapped. After this greedy swapping step, Set A had 105,527 probes and Set B had 106,558 probes. We ensured that each 9-mer appears at least 8 times in any of the sets by adding probes (3804 for Set A and 3538 for Set B) that are formed by merging three 9-mers.

We ran MegaBLAST (version 2.2.20 with command line parameters -W 12 -D 3 -g -S 3) in order to identify matches with at least 14 consecutive bases, or with at least 17 bases with at least 12 consecutive bases, to other sequences in either the forward or reverse-complement orientation. Some probes can match to many other probes because the same set of 9-mers tend to get merged in the same probe when we combine three or four 9-mers. We removed the probes that have matches to at least four other probes. For probes with less than 4 matches, we disrupted the matches by replacing two bases in the middle of matching subsequences. Among the 15 (excluding the original probe from 16 possible cases) modified probes, we kept the ones that are weakly structured. We also checked for matches between the set of modified probes and the original probe set, and removed the modified probes that have matches to the original probe set. Then, we checked the distribution of 9-mers and designed probes to add missing 9-mers either by merging three 9-mers or designing a probe for a single 9-mer (using RNAinverse) when merging was not possible. We continued with running MegaBLAST with the addition of designed probes. During this iterative process, we ensured that the Sap1 binding sites did not appear in newly added probes. We fixed the probe set once each 9-mer is represented at least 8 copies in each set. There were 109,642 probes in Set A and 110,348 in Set B. Since we had more space in the array we duplicated some of the probes and ended up with 120,326 probes in Set A and 121,031 probes in Set B. Lastly, we added 22 control sequences which are known targets for a set of RBPs (i.e. RBM4, SLM2, SF2, HuR, Nucleolin, Gtf3a, Egr1). The final Set A and Set B each contained at least 8 copies of each 9-mer, 33 copies of each 8-mer, and 155 copies of each 7-mer. There remained 2,858 strongly structured probes (containing 9-mers that are self-structured) in the final design.

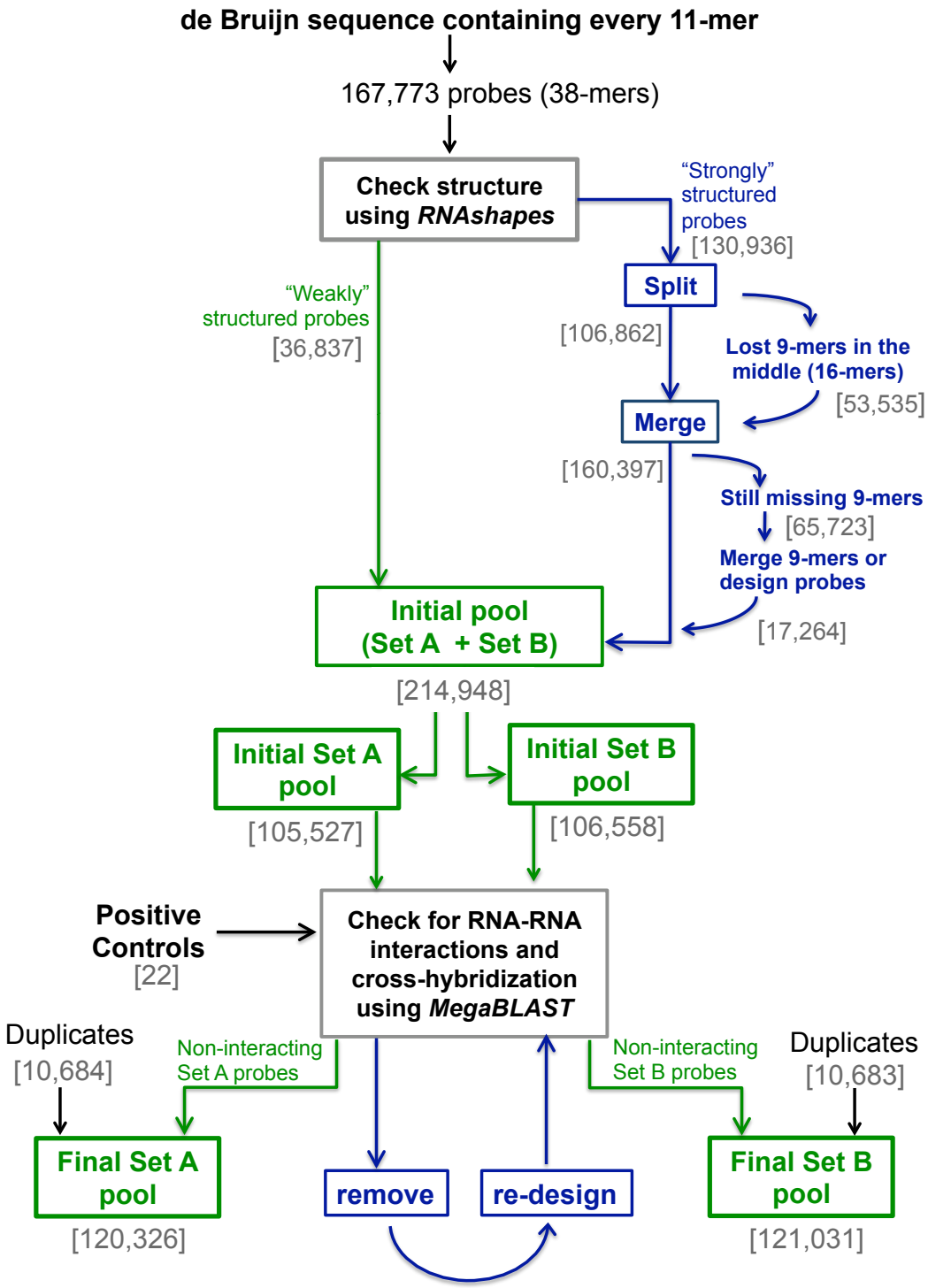


Figure 3.5: *Flow chart describing the RNA pool design for RNAcompete-v2.* The numbers in square brackets show the number of probes or n-mers during that step for Set A and Set B, respectively. Initial pool (Set A + Set B) consists of weakly structured de Bruijn probes (37K), and probes that are formed by merging fragments of strongly structured probes (160K), merging missing 9-mers or designing new probes to represent missing 9-mers (17K). Initial pool is then divided into two sets (Set A and Set B) that each contains at least 8 copies of all possible 9-mers. Next, MegaBLAST is used to check for cross-hybridization and RNA-RNA interaction. The probes that have matches to other probes are either removed or redesigned. Lastly, some of the probes in Set A and Set B are duplicated (10K), and control probes that contain known binding sites (22) are added to the final design.

3.4 Normalization

The normalization procedures for v1 and v2 were similar to widely used methods. First, each batch of experiments was represented with a matrix where rows correspond to probes and columns are the pulldown intensities of the RBPs in that batch. The v1 normalization involved removing low-quality spots due to problems related to image analysis, running variance stabilization normalization to calibrate multiple arrays[104], adjusting pool and pulldown channels with an offset and multiplicative constant, and removing the spots with low intensity in either pool or pulldown channels. Similarly, the v2 normalization involved removing low-quality spots due to spatial trends or other problems related to image analysis, applying quantile normalization to calibrate multiple arrays, and removing effects due to background signal or nonspecific binding by subtracting the row medians from each probe intensity.

3.5 Analysis of binding specificities from v1 data

We analyzed the binding specificities of an RBP by estimating its affinity to every possible 7-mer. The affinity for a 7-mer is estimated as the trimmed average (excluding the top and bottom quartiles) of the intensities of the probes containing that 7-mer. We performed the 7-mer analysis within each of the replicate designs and evaluated the success of an experiment by comparing 7-mer scores across the two sets.

We evaluated RNAcompete-v1 using nine RBPs (GEO record GSE15769) representing four different classes of RNA-binding domains (RRM, K Homology (KH), SAM and cold-shock domain). We first checked whether there is any correlation between stem-loops with identical loop sequences (embedded in different stems) for each protein. Except for Vts1p, we did not find any correlation between Set A and Set B stem-loops containing the same loop sequence (Figure 3.6). This is consistent with the fact that Vts1p is the only protein (among the nine RBPs) that is known to bind to loop sequences within

an RNA stem-loop structure. Surprisingly, Vts1p also bound to weakly structured sequences; however, we found out that these probes can fold into short stem-loops with the Vts1p binding site in the loop region. We therefore decided to include only weakly structured sequences to calculate 7-mer scores. Figure 3.7 shows the 7-mer scores across Set A and B for the nine RBPs. For all the proteins, Set A and Set B 7-mer scores were well correlated and the motifs were readily apparent among the top-scoring sequences. Figure 3.8 compares the previously reported Position Weight Matrix (PWM) models (column 3) with PWMs learned from RNAcompete data (work done by Esther T Chan) and with the top five scoring 7-mers. For six of the proteins that are tested (HuR, PTB, SF2/ASF, U1A, Vts1p and FUSIP1), top scoring 7-mers correspond to known binding preferences [105–110]. We identified novel binding preferences for three RBPs with unknown binding specificities. In summary, RBM4 preferred GC-rich sequences, SLM2 bound to sequences with UAA motif, and YB1 preferred sequences with combinations of G, C and U.

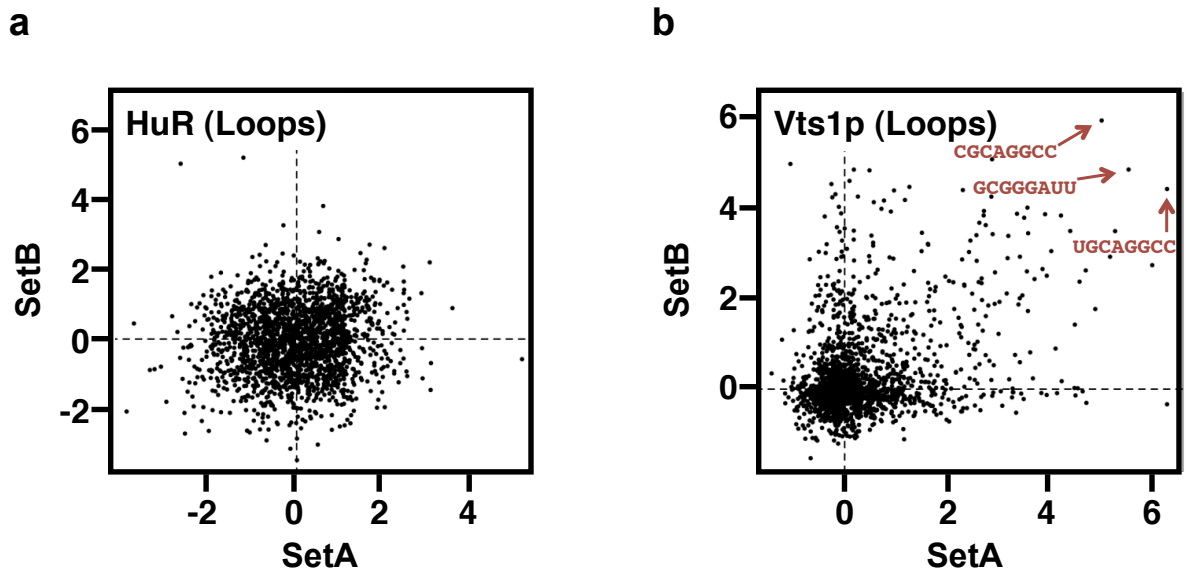


Figure 3.6: *Correlations between identical loop sequences in Set A and Set B.* Correlations for HuR that does not bind to loop sequences (a) Vts1p that binds to CNGG motif in a hairpin loop (b), displayed as Z-scores. Figure modified from [23].

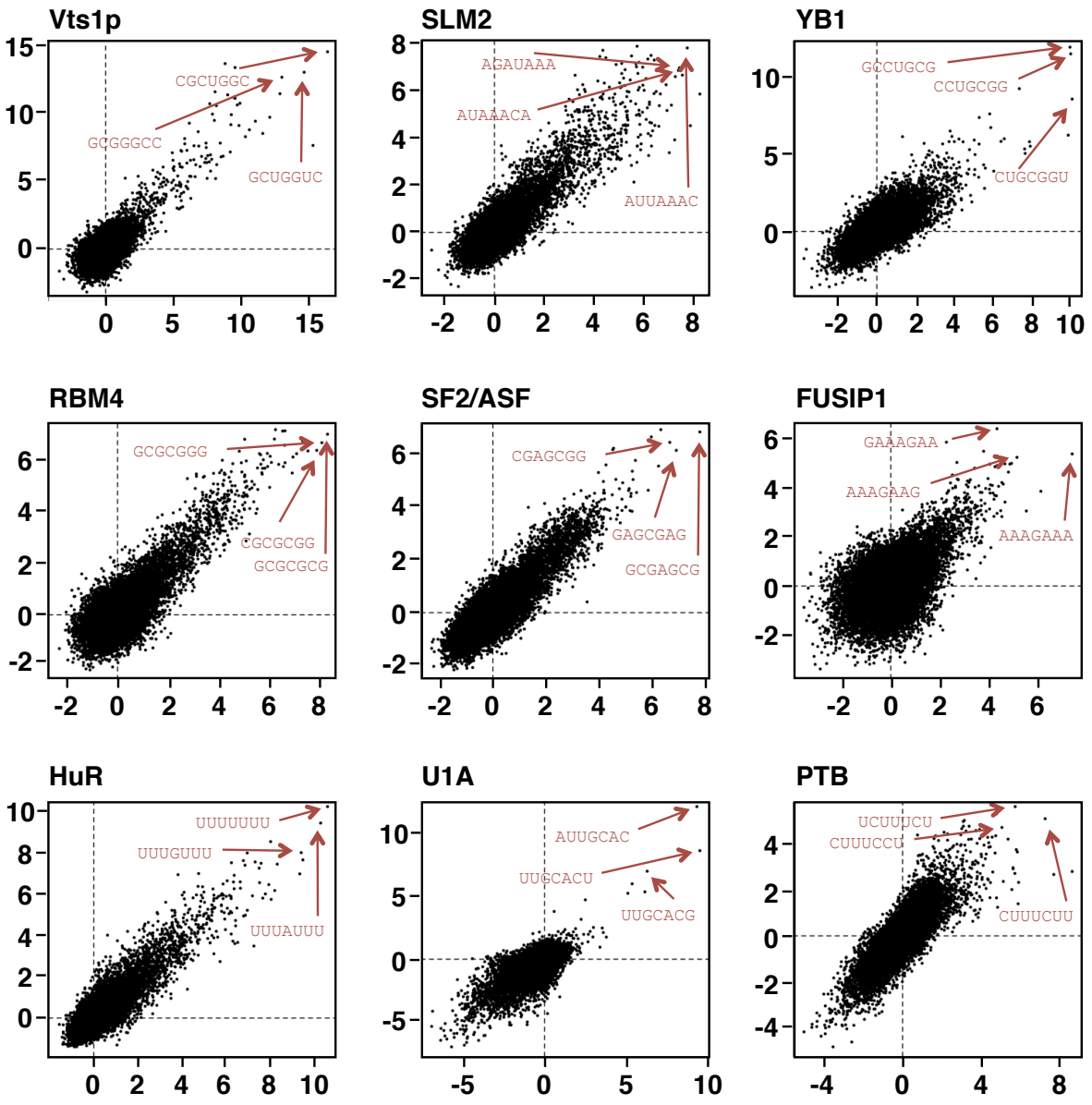


Figure 3.7: *Correlations between robust average 7-mer scores (i.e. excluding the top and bottom quartiles) from independent microarray probe sets (Set A and Set B) displayed as Z-scores (i.e. both axes have a median of zero and standard deviation of one).* Figure modified from [23].

3.6 Analysis of binding specificities from v2 data

Similar to the analysis of v1 data, we calculated scores for every 7-mer as the trimmed average (excluding top and bottom five percentiles) of intensities of the probes containing that k-mer. Unlike v1 analysis, we ignored probes with negative intensities by

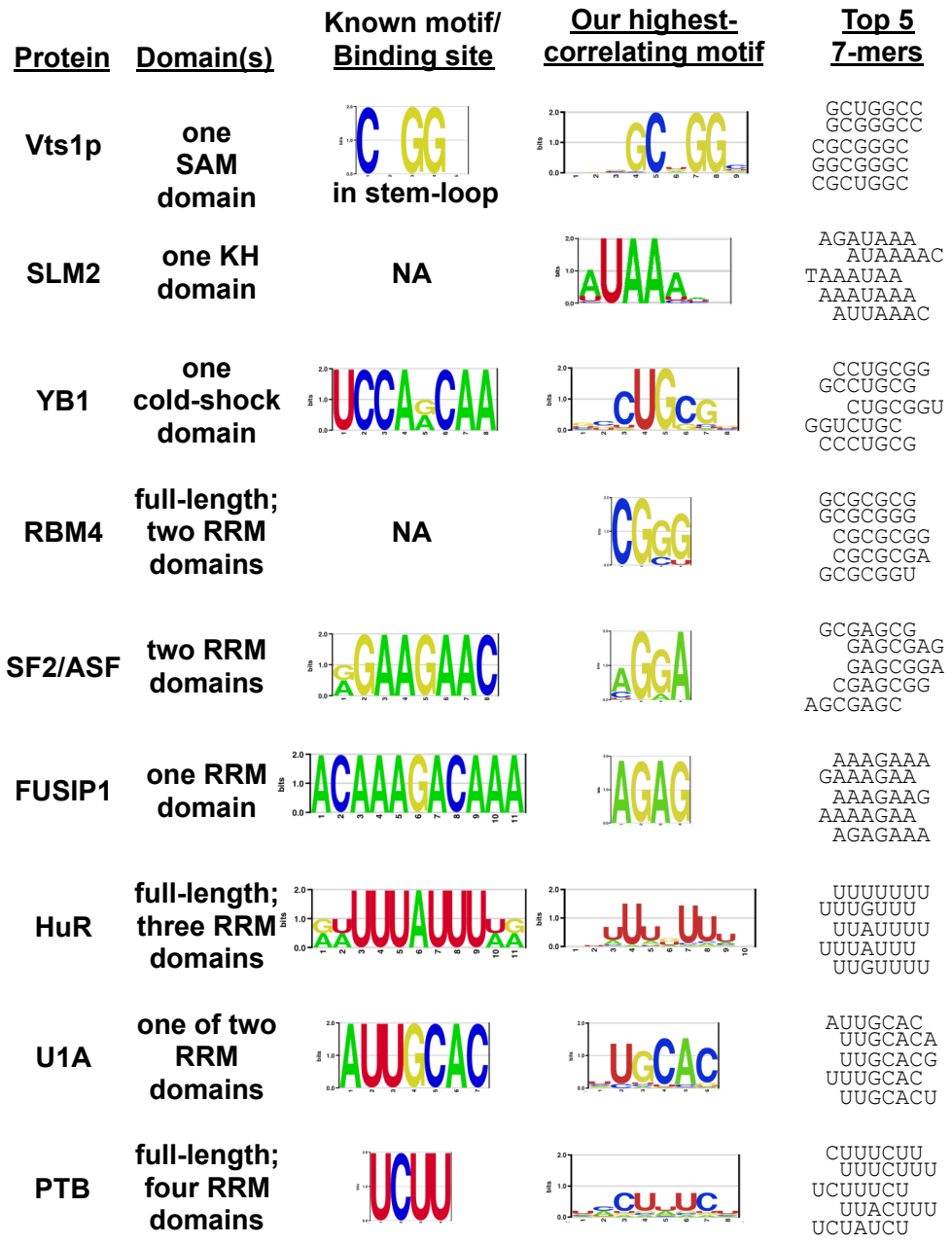


Figure 3.8: *RNAcompete analysis of nine RBPs, representing four different classes of RNA binding domains.* Previously reported consensus or PWMs models are shown for comparison. The entries under Our motifs are the highest scoring of the two possibilities (A vs. B) or (B vs. A). Figure modified from [23].

collapsing them to zero since we think that these represent the background signal. We began by checking whether RNAcompete-v2 can recover the known binding preferences

of two well-characterized proteins HuR and Vts1p. In particular, it is important to check whether Vts1p's structure preferences can still be recovered with the removal of stem-loops from the pool design. Figure 3.9 shows that there is a strong correlation between 7-mer scores calculated from Set A and Set B probes. Also, the top scoring 7-mers contain the core motif CNGG motif flanked by sequences that can basepair, confirming that RNAcompete-v2 recovers Vts1p's sequence and structure preferences. We further evaluated RNAcompete-v2 on a set of RBPs from fly and human. Figures 3.10 and 3.11 show the 7-mer scores across Set A and Set B for 15 fly and human RBPs, respectively. Strong correlations between Set A and Set B 7-mer scores, and correspondence of motifs to known binding sites for several proteins (i.e. Lark, Pum, msi, MSI1, IGFBP2, PABPC4, hnRNPC, QKI, TIA1, SAM68, SRSF2-dSR, lin28, Vts1p) demonstrate that RNAcompete-v2 is able to determine the binding specificities of diverse RBPs.

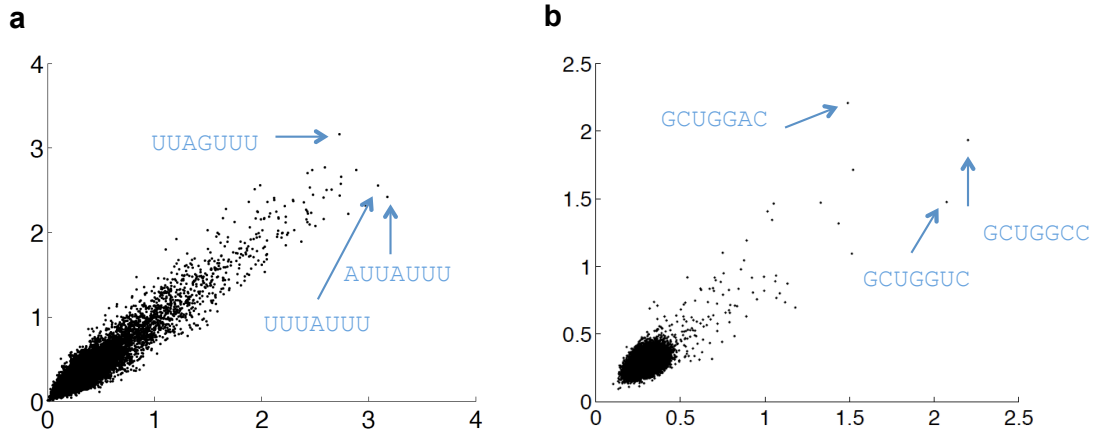


Figure 3.9: *Correlations between 7-mer scores from Set A and Set B for HuR (a) and Vts1p (b)* The top three 7-mers are shown with arrows.

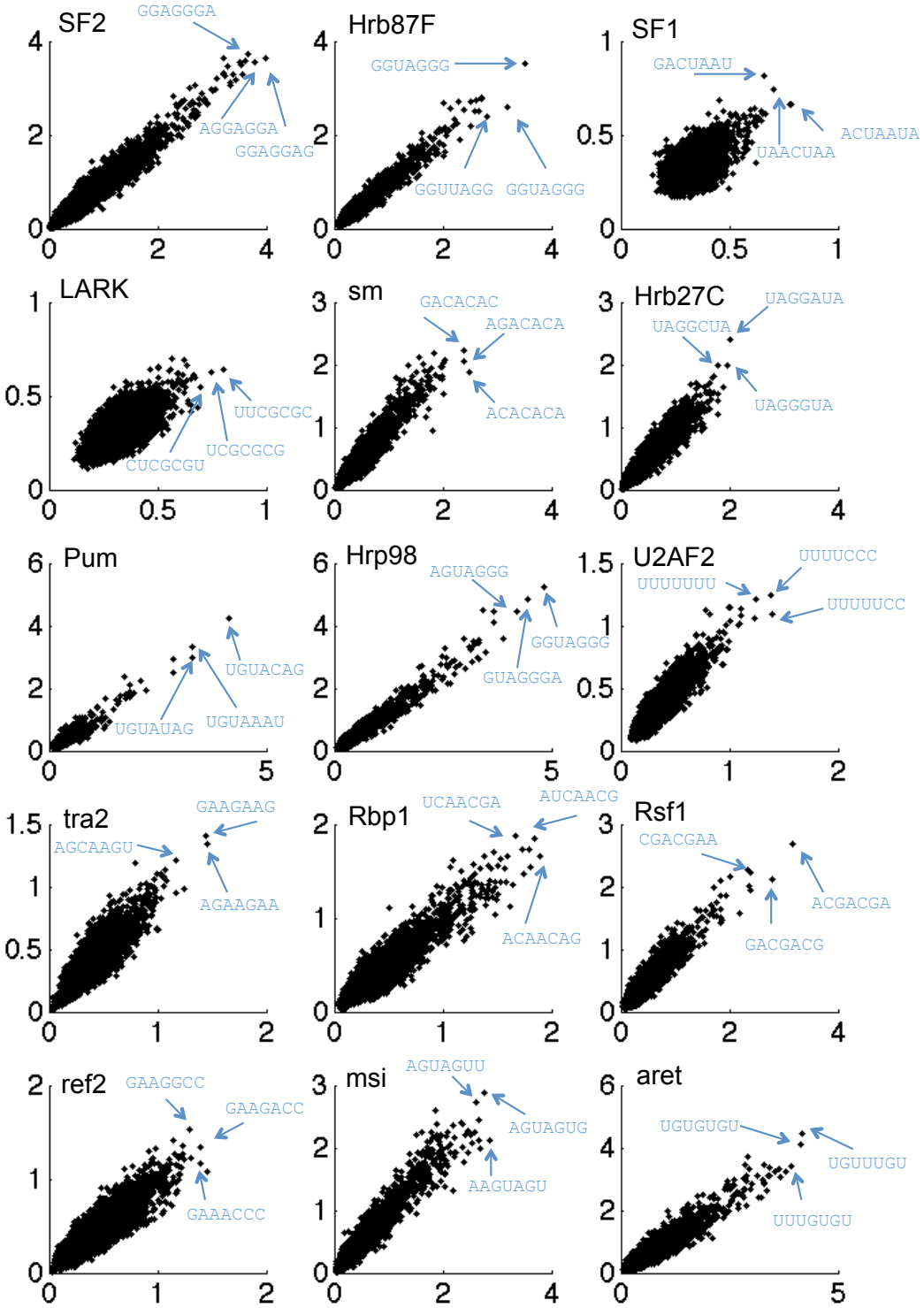


Figure 3.10: *RNAcompete analysis of 15 RBPs from fly.* Correlations between robust average 7-mer scores (i.e. excluding the top and bottom quartiles) from independent microarray probe sets (Set A and Set B) are displayed. The top three 7-mers are shown with arrows.

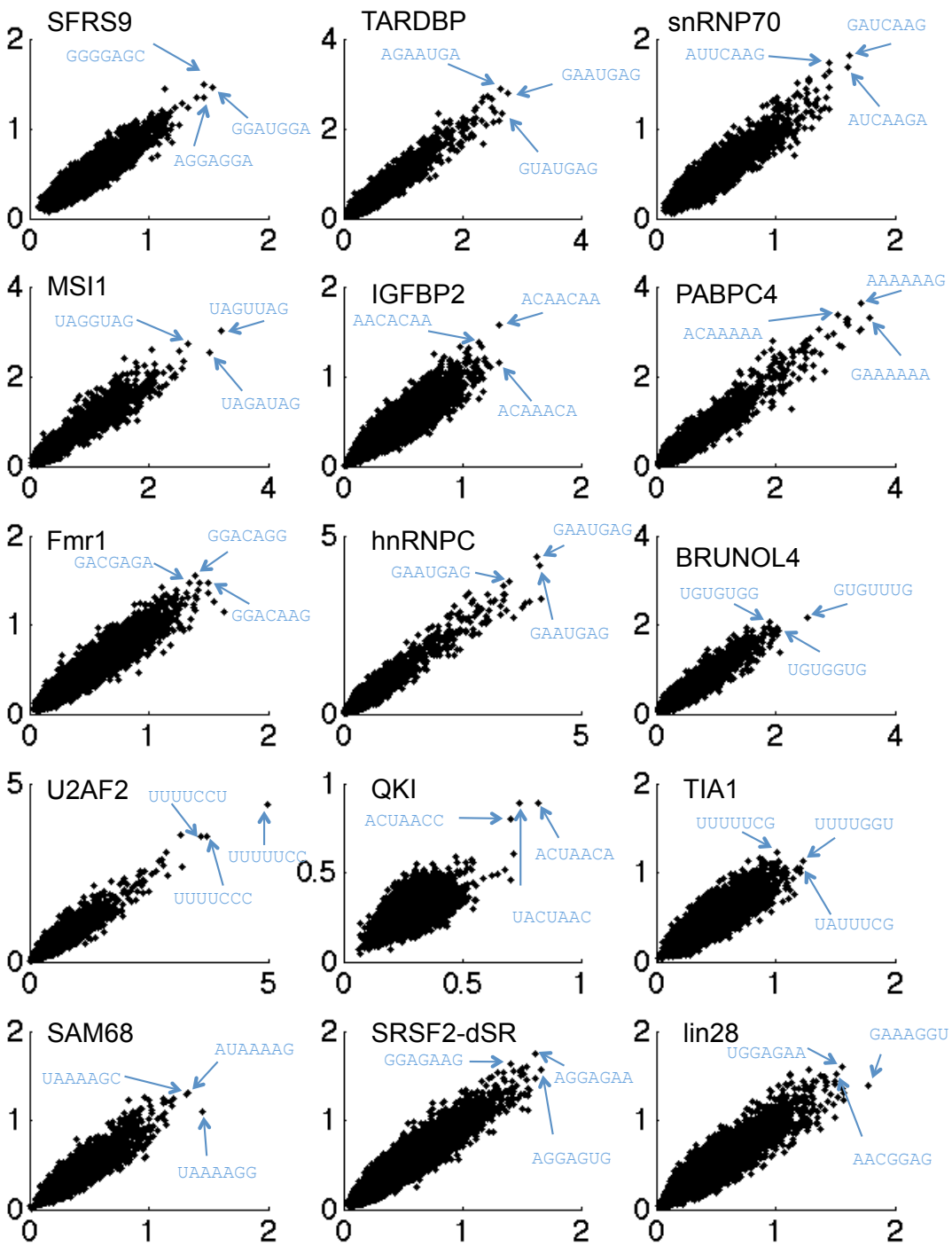


Figure 3.11: *RNAcompete analysis of 15 RBPs from human.* Correlations between robust average 7-mer scores (i.e. excluding the top and bottom quartiles) from independent microarray probe sets (Set A and Set B) are displayed. The top three 7-mers are shown with arrows.

3.7 Discussion

Determining the RNA-binding specificities of RBPs is critical to identify their targets and functions. In this chapter, we introduced RNAcompete, an *in vitro* method to rapidly characterize the binding specificities of RBPs against a custom RNA pool. One of the main contributions of RNAcompete is its careful pool design that represents diverse elements of sequence and structure. To evaluate the experiment and to increase the robustness, the pool (i.e.v1) contained two internal duplicates that each represents all possible 10-mers and all possible loops of length 8 or smaller using unique probe sequences. To compactly represent all possible 10-mers on the array, we utilized de Bruijn sequences, a well-known approach that has been previously used in protein binding microarrays [101].

A large proportion of probes generated from de Bruijn sequences were predicted to fold into strong structures. Since most characterized RBPs prefer their binding sites to be in unstructured context, we aimed to represent all possible 10-mers in unstructured (or weakly structured) context. Our strategy to produce less structured probes was to design new probes by rearranging fragments of strongly structured probes. This strategy allowed us to represent all possible 10-mers in weakly structured context using only $50K$ probes. We confirmed that this is superior to a random design strategy: probes generated from a random sequence of equivalent length only contained 76% of all possible 10-mers, and a large percentage of these probes were strongly structured. We put substantial effort to avoid cross-hybridization and RNA-RNA interactions by removing or redesigning probes, and this resulted in reduction of desired coverage.

RNAcompete-v1 design was evaluated on nine RBPs encompassing four classes of RNA-binding domains. For all nine proteins tested, motifs were apparent among the top-scoring sequences, and individual 7-mer scores correlated well between Set A and Set B. Motifs deduced from top scoring 7-mers recovered the known binding preferences of six of the proteins, and identified novel binding preferences for three uncharacterized RBPs.

In analyzing RNAcompete-v1 data, except for Vts1p, we did not observe a correlation for stem-loop sequences between the duplicate designs, suggesting that none of these additional RBPs have a preference for stem-loops. Also, we were able to recover Vts1p's preferences using weakly structured sequences alone. Therefore, we decided to remove stem-loops from the array design in the improved version of RNAcompete (i.e. v2). We used the space generated by the removal of stem-loops to increase the coverage of all possible k-mers in weakly structured context. The initial design for RNAcompete-v2 pool initially contained all possible 11-mers. After satisfying several constraints of the protocol (e.g. removing cross-hybridization), and splitting the total probe set to two replicate designs, there remained at least 8 copies of all possible 9-mers in weakly structured context. RNAcompete-v2 was evaluated on several RBPs from fly, human and yeast. Similar to v1 results, we observed a high correlation between 7-mer scores from Set A and Set B, and top scoring 7-mers correspond to known binding preferences for several control RBPs. Moreover, Vts1p's preference for stem-loops was recovered with the v2 design that only contained weakly structured sequences.

Our results from RNAcompete-v1 and v2 data indicate that RNAcompete is a valuable tool to study RBP-RNA interactions. Other strengths of RNAcompete are that the system is relatively fast, and that arbitrary binding preferences can be queried without any prior knowledge. The complex pool design provides a proof-of-principle to represent arbitrary sequence and structure elements on a microarray. We envision that much more complex defined pools can be generated and coupled to a deep sequencing output to avoid problems of cross-hybridization.

Chapter 4

RNAcontext: motif models for RNA-binding proteins

Several motif finding methods have been introduced to identify the binding preferences of transcription factors. These methods can also be used for finding RBP sequence preferences. However, this approach can give misleading results when an RBP has non-trivial structural preferences. For example, Vts1p is a yeast RBP that preferentially binds CNGG loop sequences within RNA hairpins [111], however, this binding preference can be difficult to detect without consideration of this structural preference (e.g., [6]). RBP motif finding can made more reliable by training structure-naive algorithms only on RNA sequence likely to be in the preferred context [57, 100]. For example, Foat and Stormo [100] could reliably extract the Vts1p sequence binding preferences from *in vivo* binding data by using only loop sequences (from likely hairpin loops) to train the MatrixREDUCE [36] motif finding algorithm. Similarly, the MEMERIS [57] algorithm adapts the MEME [40] motif finding algorithm to search for RNA motifs enriched in single-stranded regions by assessing a prior on each word according to its structural accessibility. MEMERIS

[0]The work in this chapter previously appeared in the following publication:
Kazan et. al. (2010) RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins. PLoS Comp Biol 6(7):e1000832

predicts binding sites more accurately than MEME for a number of proteins, including the mammalian stem-loop binding RBP U1A. However, applying this strategy only allows a single, pre-defined structural preference to be queried. Ideally, an RBP motif finding method should consider multiple possible structural contexts simultaneously, and detect the relative preferences of a particular RBP for each.

Covariance models (CMs) [93] are RNA motif models often used for modeling families of ncRNAs (e.g., [112]) and have the capacity, in theory, to represent both the sequence and (arbitrary) structure preferences of RBPs. However, CMs have a reported tendency to overpredict secondary structure [99]. Indeed, recent CM-based motif models of Puf3p, Puf4p, and HuR [96, 113] predict they preferentially bind RNA hairpins and contradict structural, *in vitro* and *in vivo* evidence [23, 97, 108, 114], that they bind unstructured single-stranded RNA.

In this chapter, we present a new strategy for modeling RBP binding sites that learns both the sequence and structure binding preferences of an RBP. Our method assumes that the primary role of RNA secondary structure in RBP binding is to establish a structural context (e.g., loop or unstructured) for the RNA sequence recognized by the RBP. As such, we annotate each nucleotide in terms of its secondary structural context (e.g., paired, in a hairpin loop or bulge). Cognizant of the fact that a given RNA sequence can have multiple, distinct stable secondary structures, this annotation takes the form of a distribution over all its possible contexts. These distributions are estimated using computational algorithms for RNA folding. Compared to CMs, this simpler secondary structure representation is much more suitable to query RBP structure preferences including the preference to single-strandedness. Our new model can be discriminatively trained (as [36, 59, 115]) thus facilitating its use with either binding affinity data or sets of bound sequences.

We apply RNAcontext to several RNA-binding affinity datasets (obtained with RNAcompete), demonstrating that it can infer the RBP structure and sequence-binding pref-

erences with greater accuracy than other motif-finding methods. RNAcontext recovers previously reported sequence and structure binding preferences for well-charactered RBPs including Vts1p, HuR, and PTB and predicts new structure binding preferences for FUSIP1, SF2/ASF, SLM2, and RBM4.

4.1 Methods

We now present our approach for discovering RNA sequence and structure binding preferences of RBPs. This section is organized as follows: we first describe how we annotate an RNA sequence in terms of its structural context. Then, we discuss the details and the mathematical formulation of our motif model. Lastly, we describe our procedure for fitting the RNAcontext motif model.

4.1.1 Structural annotation of RNA sequences

We use computational models to predict the RNA secondary structure though our algorithm can use experimentally determined RNA secondary structures when they are available [89]. Instead of focusing on the single minimum free energy structure, which is often not representative of the full ensemble of possible structures [77], we consider the ensemble of secondary structures that the RNA can form.

In the experiments reported here, we used Sfold [77] to estimate the marginal distribution at each nucleotide over structural contexts (e.g. paired, unpaired, hairpin loop) for each position of the sequence by sampling a large number of structures for the sequence according to the Boltzmann distribution. We annotated each base in each structure using our context annotation alphabet (described below) and then we set the structural context distribution (hereafter called the annotation profile) to be the empirical annotation frequencies for that base across these samples. In all experiments described herein we used 1,000 samples.

Our motif model can use any annotation alphabet. However, in the experiments described herein, we only use the alphabet P, L, U, M indicating that the nucleotide is paired (P), in a hairpin loop (L), or in an unstructured (or external) region (U). The last annotation, M, stands for miscellaneous because we combine the remaining unpaired contexts (i.e. the nucleotide is in a bulge, internal loop or multiloop). This group of structural contexts can capture most known RBP structural context preferences.

4.1.2 Motif model

Figure 5.1 shows an overview of our method. A set of sequences together with Sfold predicted structure annotation profiles serve as input to the model. Each input RNA molecule is scored using the sequence and structure parameters. Formally, let $S = \{s^1, s^2, \ldots, s^N\}$ represent the input set of sequences and let $P = \{p^1, p^2, \ldots, p^N\}$ be a set of real-valued matrices that represent the annotation profiles of the corresponding sequences. We use A to represent the alphabet which is composed of the structure features and associate each annotation in A with one of the rows of p^i. The columns of p^i correspond to the positions in sequence s^i and are discrete probability distributions over the annotations in the alphabet A.

Let $\Theta = \{\Phi, \Gamma, \beta_s, \beta_p, K\}$ represent the model parameters where K is the width of the binding site, Φ is a Position Weight Matrix (PWM) of sequence features with dimensions $4 \times K$, Γ is a vector of structure annotation parameters with one element for each letter in the alphabet A. For instance if $A = \{\mathtt{P}, \mathtt{L}, \mathtt{M}, \mathtt{U}\}$ then Γ will consist of parameters (Γ_P, Γ_L, Γ_M, Γ_U) for the structure annotations P, L, M and U, respectively. Lastly, β_s and β_p stands for the bias terms in sequence affinity model and structural context model respectively.

We use Θ to assign a score, $f(s, p, \Theta)$, to a sequence s and its corresponding annotation profile p. For an RBP with a binding site of width K, following [60], we define $f(s, p, \Theta)$ as the probability that at least one of its subsequences of length K (which we call K-mers)

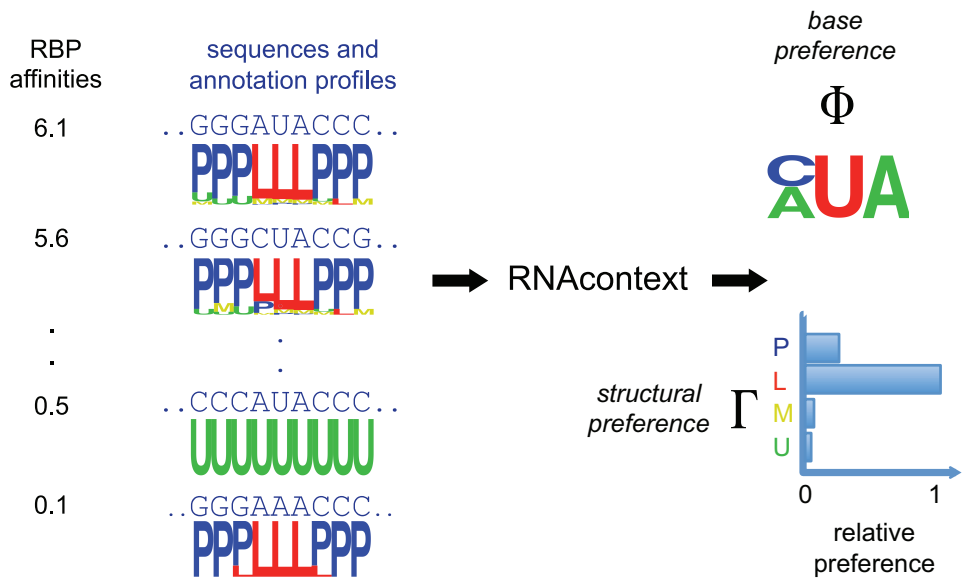


Figure 4.1: *Overview of RNAcontext method: the input, parameters and motif model.* The input to RNAcontext consists of a set of sequences together with their associated structure annotation profiles (estimated using Sfold) and RNA-binding affinity estimates for the given RBP. The motif model has sequence parameters (Φ) and structure parameters (Γ) where the former describes the inferred base preferences (as a Position Weight Matrix) and the latter describes the relative structural preferences of the RBP to different structural contexts. Shown is a toy example, where the sequences with highest binding affinities have *AUA* or *CUA* in hairpin loop context and the sequences with lowest binding affinities either lack the sequence motif or contain the sequence motif in another structural context. By learning a motif model that predicts the input affinities, RNAcontext would infer the sequence and RNA structure preferences as shown on the right part of the figure.

is bound by the RBP, that is:

$$f(s, p, \Theta) = 1 - \prod_{t=1}^{|s|-K+1} 1 - N(s_{t:t+K}, p_{t:t+K}, \Theta) \tag{4.1}$$

where $N(s_{t:t+K}, p_{t:t+K}, \Theta)$ is an estimate of the probability that the K-mer with base content $s_{t:t+K}$ and with structural context defined by the probability profile matrix $p_{t:t+K}$ is bound. Here, $s_{t:t+K}$ indicates the subsequence of s between (t)-th element and $(t+K)$-th element, inclusive, and $p_{t:t+K}$ is a matrix whose columns are the annotation distributions for each of the bases between (t)-th and $(t+K)$-th position. We set $N(s_{t:t+K}, p_{t,t+K}, \Theta)$ to be the product between a term that depends only its base content, $N^{seq}(s_{t:t+K}, \Theta)$,

and one that depends only upon its structural context $C(p_{t,t+K}, \Theta)$, i.e.:

$$N(s_{t:t+K}, p_{t:t+K}, \Theta) = N^{seq}(s_{t:t+K}, \Theta) \times C(p_{t:t+K}, \Theta). \tag{4.2}$$

We interpret the term $N^{seq}(s, \Theta)$ as an estimate of the probability that the RBP will bind $s_{t:t+k}$ in the ideal structural context. We use a standard biophysical model [58–60] to define $N^{seq}(s, \Theta)$:

$$N^{seq}(s, \Theta) = \sigma(\beta_s + \sum_{k=1}^{K} \Phi_{s_k,k}) \tag{4.3}$$

where $\sigma(x) = (1 + \exp(-x))^{-1}$ is the well-known logistic function. The logistic function takes value $\sigma(0) = 0.5$ at $x = 0$ where it is an approximately linear function of x, but it quickly saturates toward 0 for negative x and 1 for positive x. We also model the structural context term using a logistic function of the sum of the structure parameters weighted by corresponding profile values plus a bias term β_p:

$$C(p, \Theta) = \sigma(\beta_p + (\sum_{a \in A} \Gamma_a \times \sum_{k=1}^{K} p_{a,k})) \tag{4.4}$$

where $p_{a,k}$ represents the probability that the base at position k of s has structural annotation a. In a preferred structural context, as represented by an annotation a associated with large positive values of Γ_a, the score $N(s, p, \Theta)$ for a K-mer s approximately equals $N^{seq}(s, \Theta)$ and is thus determined by the base content s. Whereas in a highly disfavored structural context, as represented by highly negative values of Γ_a, $C(p, \Theta) \approx 0$ and therefore the score $N(s, p, \Theta) \approx 0$ regardless of s because $N^{seq}(s, \Theta)$ is bounded above by 1 for all s. So, the context term licenses binding in favored structured contexts.

In the following section, we describe how to estimate the parameters of our motif model from binding data. However, in theory, our motif model has the flexibility to represent many different modes of RBP binding. For example, the binding preferences of RBPs, like HuR and Vts1p, that bind their preferred sequences within a specific structural

context, unstructured (U) [92] and hairpin (H) [111] respectively, can be represented by setting Φ to match their sequence binding preferences and Γ to have negative elements except for the elements of Γ that corresponds to their preferred structural context (either Γ_U or Γ_H respectively). The binding preferences of RBPs, like U1A, that have multiple preferred contexts (e.g., hairpin loops [105] or unstructured ssRNA [116]) can be captured by setting Γ_H and Γ_U to large positive values. RBPs, like Staufen, that bind dsRNA without obvious sequence preferences [117], can be represented by setting the elements of Φ to constant values, and setting Γ_P to a large positive value. Similarly, RBPs without strong structure preferences can be represented by setting the elements of Γ to zero and setting β_p to a large positive value. Our model thus extends previous efforts that model RBP binding preferences [91] by associating each RBP with a single preferred structured context which is required for binding.

In the next section, we describe how we can estimate the sequence and structure preferences of new RBPs by training our model using RBP binding (bound or unbound) or RBP binding affinity data (quantitative) for short RNA sequences.

4.1.3 Parameter estimation

We learn Θ by using our model to attempt to reproduce the observed affinity data $R = \{r^1, r^2, \ldots, r^N\}$ given the associated sequences $S = \{s^1, s^2, \ldots, s^N\}$. In particular, we model the affinity r^i of a sequence s_i as a linear function of the sequence score $f(s^i, \Theta)$ with unknown slope α and y-intercept b and search for settings of Θ, α, and b that minimize the sum of the squared differences between the measured affinity r^i and our predicted affinities $\hat{r}^i = \alpha \times f(s^i, \Theta) + b$. When we only know whether or not a given sequence is bound we use $r^i = 1$ for all bound sequences and $r^i = -1$ for sequences not bound. This formulation leads to the following least squares cost function, $E(\Theta, \alpha, b)$, that we attempt to minimize with respect to α, b, and Θ using the limited memory

BroydenFletcherGoldfarbShanno (L-BFGS) method [118]:

$$E(\Theta, \alpha, b) = \sum_{i=1}^{N} (r^i - \hat{r}^i)^2 + \lambda(\sum_{\Phi_k \in \Phi} (\Phi_k)^2 + \sum_{\Gamma_k \in \Gamma} (\Gamma_k)^2) \tag{4.5}$$

Here, we have added a regularization term scaled by a small constant λ to avoid overfitting and to avoid indeterminancy (i.e. due to the bias terms inside the logistic function for sequence and structure model and due to the scaling factor α) thus ensuring a unique global minimum. We use the same value of this constant in all experiments. We use the bound constraints feature of the L-BFGS-B package to constrain α to take positive values so that the estimated affinity increases as a function of the sequence score.

Different initializations can generate different results when running RNAcontext. For the experiments reported here, we used ten different initialization for each motif width. For motif lengths, $K + 1$, longer than the minimum length, two of these initial settings are generated by taking the optimal Φ matrix learned for K and adding a column of zeros to its left and right sides, respectively. The elements of Φ matrix for the other initializations are randomly sampled uniformly between -0.05 and 0.05. In all cases, the other parameters (Γ, β_s, β_p, α, b) are randomly sampled uniformly between -0.05 and 0.05.

4.2 Results

4.2.1 Dataset

We evaluated our motif model on RNAcompete-derived datasets (see Chapter 3 and [23]) comprised of the measured binding preferences of nine RBPs (i.e. HuR, Vts1p, PTB, FUSIP1, U1A, SF2/ASF, SLM2, RBM4 and YB1) to a pool of 213,130 unique short (29- to 38-nt) RNA sequences (the RNAcompete-v1 design described in the previous chapter). RNAcompete estimates an RBP's binding affinity for each sequence in an

RNA pool based on the relative enrichment of that RNA sequence in the bound fraction versus the total RNA pool (as measured by transformed microarray intensity ratios).

The RNA pool can be divided into two separate sets, Set A and Set B, that each individually satisfy the following constraints: (i) each loop of length 3 to 7 (inclusive) is represented on at least one sequence flanked by RNA stems of 10 bases; and (ii) a population of "weakly structured RNAs" wherein each possible 7-mer is represented in at least 64 different sequences that have high folding free energy, and therefore are linear or form weak secondary structures. We call the group satisfying the first constraint the *stem-loop* sequences. This group also contains 60% of all possible 8-base loop sequences. We call the sequences satisfying the second constraint the *weakly structured* sequences. There is no overlap between the stem-loop and weakly structured sequences.

So in summary, there are two different groups of stem-loops, one in Set A and one in Set B, and similarly, two different groups of weakly structured sequences. It is important to note two things. First, though we attempted to design these sequences to be linear or hairpins, there are many unintended structures represented in the pool. For example, some of the stem-loops contain bulge or internal loops and some of the weakly structured sequences contain stem-loops. Second, no two sequences within the pool share a common subsequence more than 14 nt long.

The division of the RNA sequence pool into Set A and Set B provides a natural strategy for evaluating our motif models using two-fold cross-validation: we train our algorithm on one of the two sets and test its predictive power on the other set. This strategy provides us with two independent measurements of performance on non-overlapping training sets. Table 4.1 contains more information on the sizes and compositions of the sequences used for training and testing. The categorizations "Positive", "Negative", and "Other" that appear in this table are described below. Note that because of the removal of some probes due to RNAcompete quality controls, some affinity data is not available for some of the sequences, so the numbers in the table do not add up to 213,130 for each

Proteins	Set A						Set B					
	positive		negative		other		positive		negative		other	
	weak	stem-loop	weak	stem-loop	weak	stem-loop	weak	stem-loop	weak	stem-loop	weak	stem-loop
Vts1p	289	260	16,863	3409	11,390	9,107	282	239	16,697	3373	11,136	8,885
SLM2	520	1	18,373	9,984	19,109	14,819	566	5	18,143	10,100	18,879	14,671
YB1	555	59	15,985	11,838	29,317	21,858	519	62	15,701	11,784	29,177	21,905
RBM4	726	6	16,762	9,798	17,140	9,231	792	14	16,655	9,659	16,841	9,127
SF2	523	71	17,961	12,616	22,713	11,922	529	92	17,867	12,282	22,378	12,017
FUSIP1	201	436	21,181	4,148	9,918	15,149	234	426	20,986	4,096	9,928	15,185
HuR	608	1	16,379	13,422	24,877	4,348	586	0	15,944	13,332	25,095	4,405
U1A	310	322	8,232	325	6,752	1,740	277	328	8,250	337	6,665	1,738
PTB	467	62	21,315	6,908	18,040	15,356	448	72	21,309	6,820	17,779	15,332

Table 4.1: *The composition of Sets A and B in terms of relative proportions of stem-loops and weakly structured sequences among their Positive, Negative and Other groups.* The input sets are partitioned into these three groups according to their RNAcompete-measured affinities. The sequences with affinities above a threshold are defined as *Positive*; the sequences with affinities below the median affinities over all the sequences in the given set are defined as *Negative* and the remaining sequences are placed in the *Other* group. Within each group, the number of weakly structured sequences and stem-loops are displayed. For RNAcontext and MatrixREDUCE all the sequences in *Positive*, *Negative* and *Other* categories are used for training whereas when running MEMERIS, only *Positive* sequences are used for training. The test sets are comprised of all sequences in the *Positive* and *Negative* groups.

RBP.

4.2.2 Justification of choice of motif models for comparison

We evaluate RNAcontext against two other motif finding methods: MEMERIS [57] and MatrixREDUCE [36]. MEMERIS and RNAcontext use similar approaches to model the structural context of an RNA binding site except that MEMERIS only models a single structural context where RNAcontext considers multiple contexts simultaneously. In contrast, MatrixREDUCE does not consider the structural context of RBP binding sites and therefore can help determine the value of considering structural context in RNA motif finding. Additionally, MatrixREDUCE outperforms many standard DNA motif finding algorithms on a similar experimental assay [37] and therefore provides a strong algorithm to benchmark to compare RNAcontext and MEMERIS against. Also, like RNAcontext, MatrixREDUCE learns its motif model by trying to predict RNA

sequence affinity whereas MEMERIS searches for motif models enriched in a set of bound sequences.

4.2.3 Fitting motif models

In this subsection we describe our protocol for using the training data to fit the MEMERIS, MatrixREDUCE and RNAcontext motif models. Note that for all three methods, we fit all parameters, including those of the motif models and any free parameters (like motif width), using the training data. One of the free parameters that we consider for each method is whether it is better to train their motif model on the whole training set, or a defined subset of the training set. All of the free parameters that we consider for each method are described below. For every setting of the free parameters, we fit one motif model. The "best" motif model for each method was selected based on its ability to correctly classify "Positive" and "Negative" RNA sequences in the training set, as defined in the next paragraph. The final result of training is a single motif model for each method that we then evaluate on the test set.

The parameters of some motif models are fit using subsets of the training set because: (i) MatrixREDUCE does not model RNA secondary structure and it is possible that its performance would degrade when trained on stem-loop sequences (most of whose bases are paired); and (ii) MEMERIS takes as input a set of "bound" sequences that contain RBP binding sites. For MEMERIS, "bound" sequences are selected using a manual cutoff that captures the right tail of the distribution of the RNAcompete affinity estimates. We used a different cutoff for each RBP and each training set and the number of bound sequences ranged between 234 and 792 for the RBPs analyzed. Additionally, we used these bound sequence as the "Positive" sequences for Area Under the Precision-Recall Curve (AU-PR). For the "Negative" sequences required by the AU-PR calculation, we used those with estimated affinities below the median affinity of the training set. Any sequence not deemed a "Positive" or "Negative" is labelled as "Other" in Table 4.1.

We score each motif model's performance by using it to estimate RNA-binding affinities for the "Positive" and "Negative" sequences and then evaluating classification accuracy using the AU-PR. Because each algorithm models RBP binding preferences in a slightly different manner, in this section, we also describe how we estimate RNA-binding affinity for each sequence using the motif models for each algorithm.

For each method, we trained two sets of motif models. One set of models was fit using the full training set which consists of all RNA sequences in the training set for MatrixREDUCE and RNAcontext and all bound RNA sequences in the training set for MEMERIS. The other set of models was fit using only the weakly structured sequences in the training set (i.e. removing the stem-loops).

We consider a wide range of combinations of free parameters for MEMERIS. In particular, we tried all possible combinations of the following free parameter choices: the EF and PU options for measurement of single-strandedness; OOPS, ZOOPS and TCM options for the expected number of motifs per sequence; motif lengths between 4 and 12 nts (inclusive); different values for the pseudocount parameter (i.e. 0.1, 1 and 3); and selecting the training set using a permissive cutoff (i.e. the bound sequences) or a stringent cutoff (i.e. the top half of bound sequences). The final option means that we consider four different subsets of the training set for each setting of the other free parameters (i.e. permissive/full, stringent/full, permissive/weak, stringent/weak). In total, we fit 648 different motif models for MEMERIS for each training set. We estimate affinity for each RNA sequence using a MEMERIS Position Frequency Matrix (PFM) motif model by following an approach similar to that used by MotifRegressor [34]. Namely, we calculated the foreground probability of a K-mer under the product-multinomial distribution defined by the PFM and calculated the background probability using a third-order Markov model trained on either the full training set (or test set, as appropriate). As explained in Chapter 2, the ratio of the foreground and background probabilities is an estimate of the relative affinity of the RBP for that K-mer. For some RBPs, when it led to a performance

increase, we also multiplied this affinity by the probability that the site was accessible, as determined using the optimized settings of the EF/PU and pseudocount parameters for that training set. To estimate the affinity of the entire sequence, we summed its k-mer relative affinities. Note that we also tried MAST [119] to score the sequences using MEMERIS's motif models but test set performance decreased (data not shown).

We used MatrixREDUCE to generate single motifs with widths ranging from 4 to 12 by setting `max_motif` to 1. The MatrixREDUCE program automatically selects the appropriate motif width, so we only needed to choose between two different MatrixREDUCE motifs on each training set (one trained on the full set and the other only on the weakly structured sequences). Note that MatrixREDUCE's PSAM motif model directly estimates relative binding affinity of the RBP for each k-mer, so to estimate RNA sequence affinity, we summed PSAM scores for each constituent k-mer.

We ran RNAcontext with motif widths ranging from 4 to 12, thus creating 18 motif models per training set, and used equation (4.1) to score RNA sequences using these models.

For all three methods, for each training set, we used the AU-PR on training set "Positives" and "Negatives", to select the best single model among the fitted models. The free parameters settings for the selected models are in Table 4.2.

4.2.4 Performance evaluation

RNAcontext achieved higher average AU-PR values than MEMERIS and MatrixREDUCE on all of the nine RBPs analyzed (Table 4.3). It also had significantly higher AU-PRs than either method on 15 of the 18 test sets encompassing seven of the nine RBPs (the largest P-value was $P = 10^{-94}$, Wilcoxon's sign-rank test on the AU-PR values of 1,000 bootstrap samples; See Table 4.4 for the complete results of bootstrap analysis).

The improvement in AU-PR of RNAcontext compared with MatrixREDUCE is largest for proteins whose preferred structural context is less common in the RNA pool, reflecting

Proteins	RNAcontext			MEMERIS					MatrixREDUCE		
	Set	MW-A	MW-B	Set	MW-A	Other-A	MW-B	Other-B	Set	MW-A	MW-B
Vts1p	full	7	7	full	7	EF ZOOPS 0.1 *	7	EF ZOOPS 0.1 *	full	8	9
SLM2	weak	8	8	full	11	PU ZOOPS 0.1	5	PU OOPS 3	weak	8	9
YB1	full	8	8	weak	8	EF ZOOPS 1	12	EF ZOOPS 1	full	8	8
RBM4	full	6	8	weak	4	EF OOPS 3	4	PU ZOOPS 3	full	7	7
SF2	weak	5	5	weak	4	PU OOPS 1 *	4	EF OOPS 3	full	4	5
FUSIP1	full	12	11	full	6	EF TCM 0.1 *	6	EF TCM 0.1	full	8	8
HuR	weak	9	9	full	10	PU OOPS 0.1	12	PU OOPS 1	full	8	8
U1A	weak	9	9	weak	12	EF ZOOPS 0.1	8	EF ZOOPS 0.1 *	full	8	9
PTB	weak	7	7	weak	12	EF OOPS 3	12	EF OOPS 1	weak	7	5

Table 4.2: *Details about the chosen models for RNAcontext, MEMERIS and MatrixREDUCE* Optimal free parameter settings for RNAcontext, MEMERIS and MatrixREDUCE. The column *Set* describes the training set and contains either *weak* or *full* where *weak* indicates that both motifs were trained on the weakly structured sequences and *full* indicates that full set of sequences was used. The columns, *MW-A* and *MW-B*, show the selected motif length for the test sets A and B. There is an extra *other* column for MEMERIS which shows the other free parameters that are chosen. Namely, *EF* and *PU* are two different ways to measure single-strandedness of a region. *OOPS* (exactly one motif occurrence per sequence), *ZOOPS* (zero or one motif occurrence per sequence), *TCM* (zero or more motif occurrence per sequence) are options (*-mod*) that indicate the expected number of motifs per sequence. The values in the next column (i.e. 0.1 or 1) are the chosen pseudocount parameters among the available values $\{0.1, 1, 3\}$. The lower the pseudocount value, the more impact the single-strandedness of the binding site has in the model. Two different thresholds were used to define the input to MEMERIS. The rows containing * indicate that the more stringent threshold was selected. The last three columns contain the selected free parameter settings for MatrixREDUCE.

the fact these are the hardest binding sites for MatrixREDUCE to predict. For example, RNAcontext performs much better than MatrixREDUCE on Vts1p which binds to CNGG in the loop of an RNA stem-loop. This sequence appears frequently outside of a loop context in the RNA pool. We also see large improvements for RBM4 that binds to CG containing sequences in an unpaired context, likely because these sequences often appear in stems. In contrast, HuR's binding site is U-rich and, as such, is rarely paired in either the training or test set. In this circumstance, MatrixREDUCE's lack of a structural model does little harm to its performance.

Although MEMERIS has higher average AU-PR than MatrixREDUCE for stem-loop binding proteins Vts1p and U1A, reflecting the value of its model of structural context,

Proteins	**RNAcontext**	**MEMERIS**	**MatrixREDUCE**
RBM4	**0.91***	0.43	0.63
FUSIP1	**0.53***	0.31	0.32
Vts1p	**0.65***	0.58	0.56
YB1	**0.17***	0.07	0.11
SLM2	**0.81***	0.49	0.77
SF2	**0.70***	0.50	0.66
U1A	**0.30**	0.27	0.21
HuR	**0.96***	0.74	0.94
PTB	**0.69**	0.26	0.67

Table 4.3: *Comparison of predictive accuracy of three motif finding models using both weakly structured and stem-loop sequences in the test set* The values show the average AU-PR across two test sets and bold values indicate the best performing method. Rows are sorted by decreasing relative gain of RNAcontext to the best of MEMERIS & MatrixREDUCE. For all the methods, displayed values were calculated using the single best motif model for each method chosen based on the two training set performance. According to the Wilcoxon's sign rank test performed on paired AU-PR values across 1,000 bootstrap samples from the test set results of the three methods, all differences between RNAcontext AU-PR and that of the other algorithms are statistically significant (the largest P-value is 1.9×10^{-95}) except for the differences on PTB and U1A.

Proteins	MatrixREDUCE		MEMERIS	
	A	B	A	B
RBM4	1000/0	1000/0	1000/0	1000/0
FUSIP1	1000/0	1000/0	1000/0	1000/0
Vts1p	990/10	999/1	996/4	959/41
YB1	1000/0	995/5	1000/0	1000/0
SLM2	974/26	803/197	1000/0	1000/0
SF2	983/117	875/125	1000/0	1000/0
U1A	1000/0	504/496*	999/1	295/705
HuR	978/22	996/4	1000/0	1000/0
PTB	213/787	961/39	1000/0	1000/0

Table 4.4: *Bootstrap analysis of relative AU-PRs* Each entry represents the number of times RNAcontext has a larger/smaller AU-PR value compared to AU-PR values of MatrixREDUCE & MEMERIS on 1,000 bootstrap samples from the test set results (shown in Table 1). * indicates that difference is not significant according to Wilcoxon's sign rank test.

its average AU-PR was otherwise worse than that of MatrixREDUCE and always worse than that of RNAcontext. This is likely due to its inability to make use of the affinity data associated with each sequence. One consequence of this is that it can only trained on a small subset of the data. Some of the loss in AU-PR on the test set may also be due to overfitting because of the large number of parameter combinations that needed to be considered.

4.2.5 The predictive value of structural context

Having established that RNAcontext can capture RBP binding preferences better than comparable motif models that either do not model RNA secondary structure (MatrixREDUCE), or use a limited representation (MEMERIS), we then attempted to confirm that the added predictive value was due to the incorporation of structural context, rather than differences in how we estimate sequence affinity.

To do this, we compared our model based on the $\{\mathtt{P}, \mathtt{L}, \mathtt{U}, \mathtt{M}\}$ structural annotation alphabet to a simplified version where we removed the structure model. That is, we used RNAcontext motif models selected in Section 1.2.3 to score the sequences in the full test set. For the simplified version of RNAcontext, we only used the sequence parameters to score the test sequences. As in previous sections, we compared test set AU-PRs to evaluate the two models. To assess the significance of difference in AU-PR, we used 95% confidence interval of the difference estimated from 1,000 bootstrap samples. Figure 4.2 shows these differences for nine RBPs on the two cross-validation test sets. Using structural context lead to a significant improvement in AU-PR for all of the RBPs. In some cases, the difference was dramatic, particularly for RBM4, FUSIP1, U1A and Vts1p.

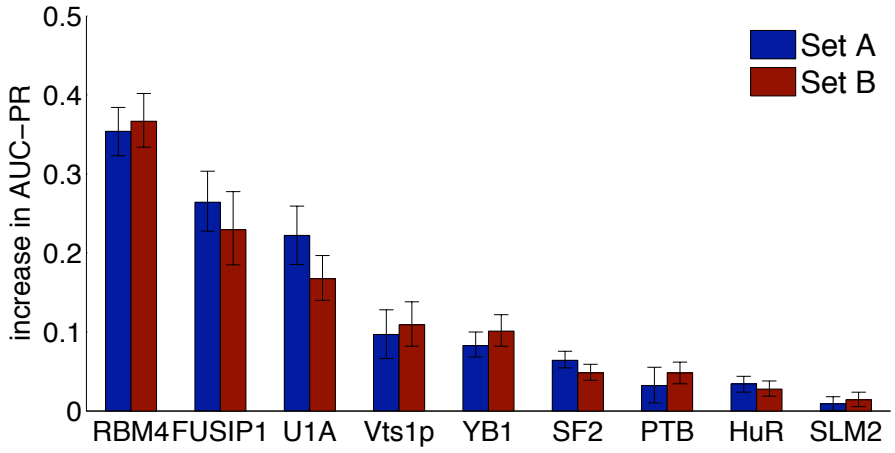


Figure 4.2: *Change in predictive accuracy of RNAcontext due to the representation of RNA structural context.* Bar graph shows the increase in AU-PR of RNAcontext with $\{\mathtt{P}, \mathtt{L}, \mathtt{U}, \mathtt{M}\}$ alphabet compared with the model without the structure model for each of nine RBPs using the two test sets. Error bars show 95% confidence interval of the difference estimated from 1,000 bootstrap samples of the test set.

4.2.6 Position weight matrices provide good approximations of sequence binding preferences for six RBPs

We then sought to assess the accuracy of PWM approximations of RNA-sequence binding preferences by comparing the predictive power of inferred 7-mers affinities to that of the three PWM-based models. We trained a "fully-specified 7-mer model" that estimates the binding affinity of an RBP for every 7-mer by taking a trimmed average of the transformed intensity ratios of the weakly-structured sequences that contain the 7-mer in the training set (see [23] for more details of this model). We then used these estimated affinities to assign a score to RNA sequences longer than seven nucleotides, by taking the mean of the affinities of each 7-mer in each sequence in the test set. We also trained and evaluated RNAcompete, MatrixREDUCE and MEMERIS motif models as previously described except that we always restricted the training and test sets to the weakly-structured sequences. We used only the weakly-structured sequences in this comparison so that

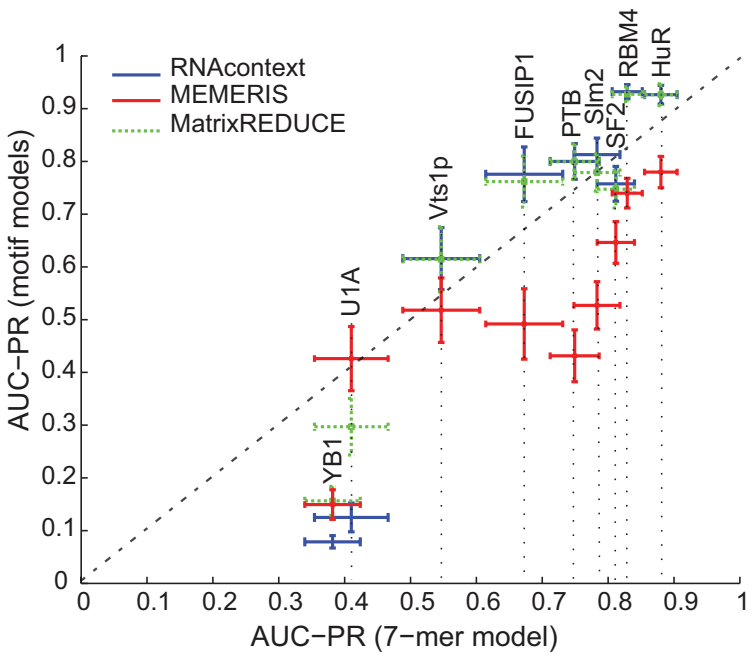


Figure 4.3: *The predictive value of the fully-specified 7-mer model (i.e. the RNAcompete model) with respect to the motif models.* This scatter plot compares the performance of the RNAcompete model to other motif models where the training and test sets are formed using only weakly structured sequences. The x-axis shows the AU-PRs of RNAcompete and the y-axis shows the AU-PRs of the three motif models: RNAcontext (blue), MEMERIS (red) and MatrixREDUCE (green). Each point corresponds to the mean AU-PR of 1,000 bootstrap samples, averaged across the two test sets. The error bars indicate the 95% bootstrap confidence interval.

we could more readily evaluate the ability of PWM models to assess sequence binding preferences separately from each method's ability to capture RBP structure binding preferences. Figure 4.3 compares the 7-mer model against the three methods with respect to average AU-PR on the test sets. PWM-based motif models perform significantly better than the 7-mer model for every RBP except U1A (and only on test set A), YB1, and SF2/ASF (the Wilcoxon sign-rank P-values for the best PWM motif model are all less than 10^{-22}). Notice that because MatrixREDUCE performs significantly better than the RNAcompete method for five of the nine RBPs, this performance gain can not be explained by the incorporation of structural context in RNAcontext.

4.2.7 The sequence and structure binding preferences for seven RBPs

Having established that RNAcontext accurately predicts the *in vitro* affinity for seven of the nine RBPs (with the exception of YB-1 and U1A), we applied RNAcontext to the entire dataset to make the best possible prediction for their binding preferences. The results are shown in Figures 4.4 and 4.5. Figure 4.4 shows the relative structural context preference of each RBP. RNAcontext's predicted structural preferences are consistent with co-crystal structures for Vts1p [111] (loop) and PTB [120] (ssRNA) and *in vitro* and *in vivo* binding data for HuR [23, 91, 108]. RNAcontext also predicts new structural preferences for SLM2, RBM4 and SF2/ASF. Of particular interest, is that RNAcontext predicts that SF2/ASF has a slight preference for RNA binding sites in bulges, internal loops, and/or multiloops (the M annotation). For FUSIP1, we report the motif model trained using only the weakly structured sequences even though the model trained on the full set (shown in Figure 4.6) had higher AU-PR. As mentioned in the legend of Figure 4.6, we could not rule out the possibility that this model reflected an artifact of our pool design despite the fact that the two models both suggest that FUSIP1 prefers its binding site to be 5' to an RNA stem.

Figure 4.5 compares the motif logo representations (generated by enoLOGOS software[121]) of RNAcontext's Φ parameters with previously reported motifs for those RBPs. To derive the energy parameters required by Enologos, we uniformly rescaled the elements of the Φ matrix so that $N^{seq}(s^*, \Theta)$ of the optimal binding site, s^*, would be 0.5 (as suggested by [60]). Underneath each of the logos for the RNAcontext motifs, we have displayed an estimate of the preferred structural context for each base. In order to identify this context, we found the top 20 best scoring k-mers in the test set under each motif model, averaged the annotation profiles for these 20 k-mers and deemed the annotation with the highest average frequency to be the preferred context for each position in the k-mer.

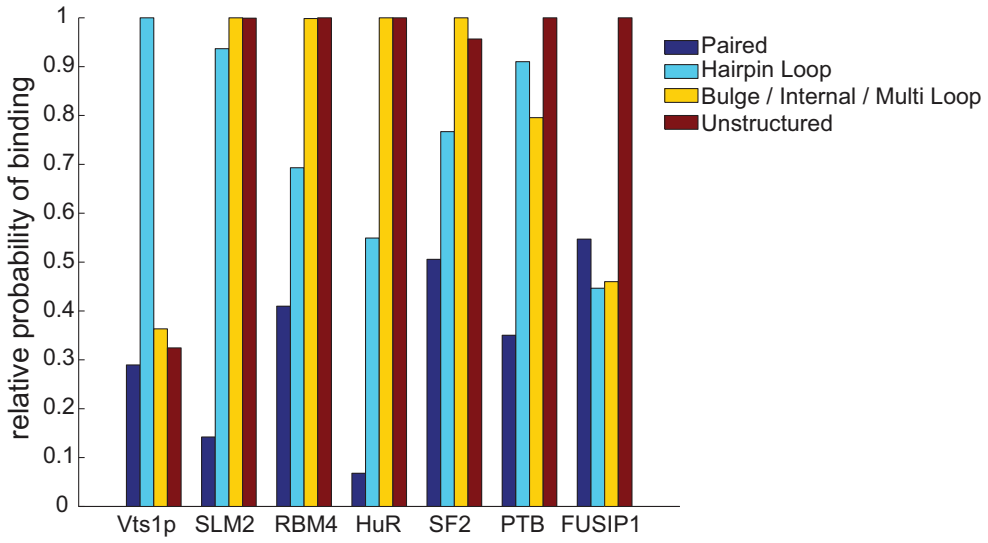


Figure 4.4: *Relative RNA structure preferences inferred by RNAcontext.* Y-axis indicates the ratio between the context scale factor $C(p^a)$ (see equation 4.4) for a structural context with probability one for the indicated annotation (a) for all bases (i.e. $p^a_{a,k} = 1, \forall k$) to the context scale factor $C(p^*)$ for the best possible structural context for the RBP (i.e. $p^*_{a*,k} = 1, \forall k$ where $a* = argmax_a \Gamma_a$). Displayed are ratios across parameters learned from the training set containing all the sequences (i.e. both Set A and Set B). For Vts1p, the most preferred context was predicted to be hairpin loop and this is consistent with the known binding preferences. SLM2, RBM4, and HuR have similar preferences, and predicted to bind regions that are not paired.

These estimates recover the fact that the Vts1p binding site (CNGG) occurs at the 5' end of the hairpin loop. Our RNAcontext motifs match previously reported binding sites and the motifs that we have previously derived from the RNAcompete data[23].

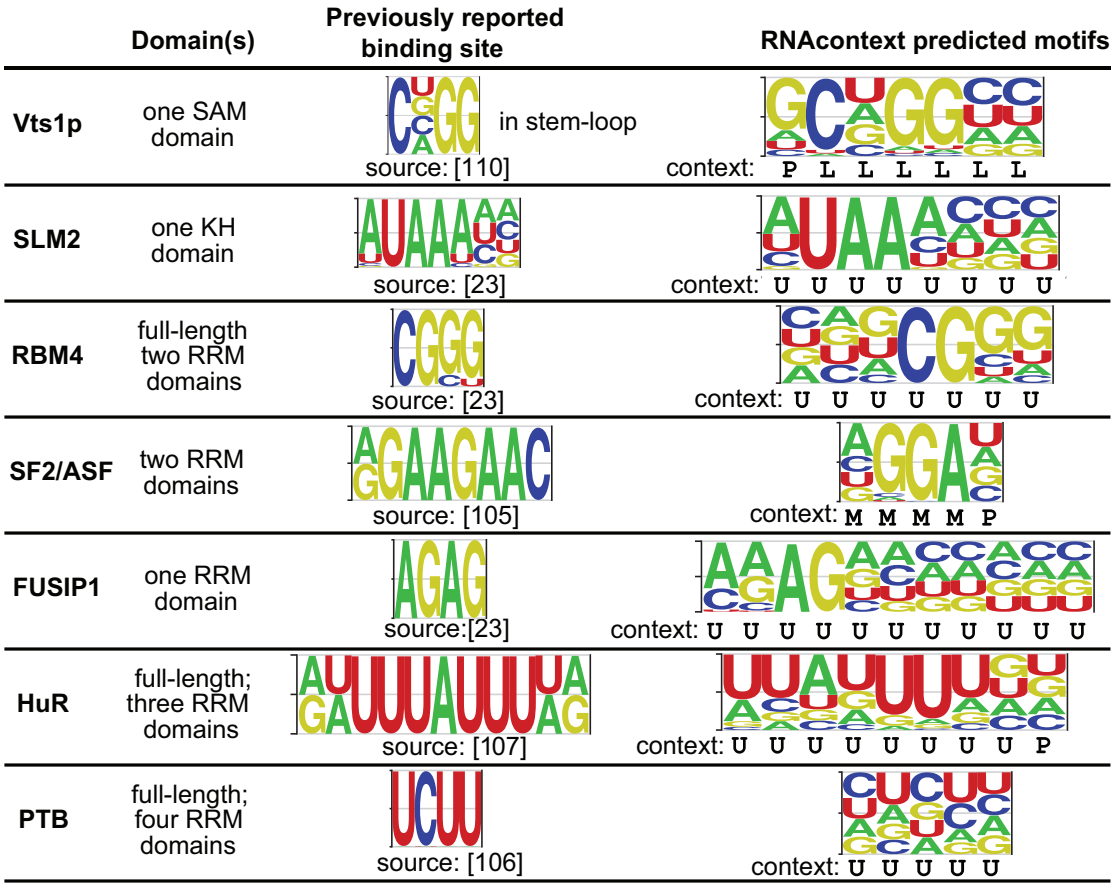


Figure 4.5: *Inferred sequence and structure binding preferences for seven RBPs.* RNA-binding domains of the proteins are displayed on the second column and previously reported binding sites are displayed on the third column for reference. RNAcontext predicted sequence parameters are shown as a PWM (fourth column). Also, an estimate of the preferred structural context for each base is displayed underneath each of the logos.

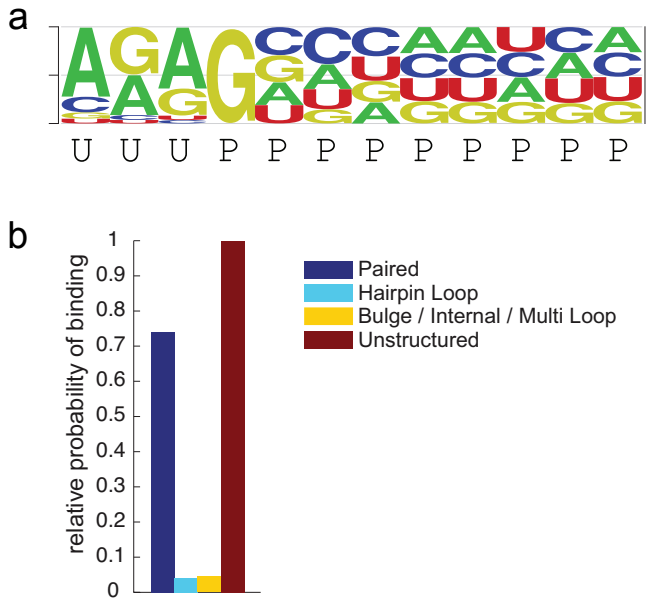


Figure 4.6: *Inferred sequence and structure binding preferences for FUSIP1 using all the sequences as the training set.* a) Predicted sequence parameters are shown using a sequence logo representation. An estimate of the preferred structural context for each base is displayed underneath the logo. b) The bar graph shows the relative RNA structure preferences of FUSIP1. Note that all sequences in the RNA pool, including the stem-loop sequences, have an unpaired 5'-AGA or 5'-AGG (the initiation sequence for T7 promoter) at their 5' end. In all stem-loop sequences, this initiation sequence is followed by a G because the bottom base pair of every stem-loop is G-C. Because AG(A/G)G is very similar to previously reported FUSIP1 binding sites [23, 110], we were concerned that this artifact of the pool design had an impact on the model of FUSIP1 binding preferences fit to the full training set. However, even in the model fit only to the weakly-structured sequences shown in Figure 4.4, there is a slight preference for the paired context compared to L and M.

4.2.8 *In vivo* confirmation of RNAcontext motif for SF2/ASF

In both Figure 4.4 and Figure 4.5, we observe a preference for the M structural context for the SF2/ASF motif. This preference has not been previously reported for SF2/ASF [122]. To confirm this unusual preference, we collected data on the *in vivo* targets of SF2/ASF from [123]. These targets were generated using the CLIP-Seq assay and consist of 296 short RNA fragments that cross-link to the protein in cultured cells which we call "bound"; and 314 transcript sequences not observed to cross-link which we call "unbound". These data supported our inferred structure preferences for SF2/ASF. In particular, by manual inspection, we discovered a number of cases of the RNAcontext motif within bulge and internal loops within the bound sequences. Also, using our model trained on the RNAcompete data, we were able to distinguish between bound and unbound sequences with higher accuracy using our model (AU-PR 0.915) compared with the version of our model with a single letter annotation alphabet (AU-PR 0.898) and MatrixREDUCE (AU-PR 0.898). Furthermore, when we train our RNAcontext model on the *in vivo* data, assigning bound sequences an affinity of 1 and unbound ones an affinity of -1, we recover the same structural preference for SF2/ASF (Figure 4.7).

4.3 Discussion

We have demonstrated that RNAcontext represents an advance over existing methods for modeling RBP binding preferences. Motifs learned by RNAcontext more accurately predicted a held out *in vitro* binding dataset for all of the nine RBPs tested. Seven of these differences were statistically significant. As expected, the size of an improvement depends on the relative representation of the preferred binding site in the preferred structural context (or contexts) in the RNAcontext dataset.

RNAcontext motif models reflect previously reported sequence and structure preferences for well-studied RBPs like HuR, Vts1p and PTB and predict new structure binding

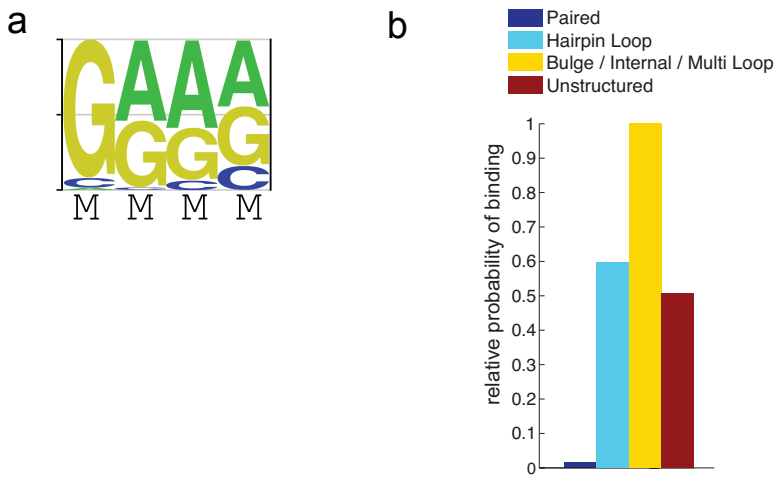


Figure 4.7: *Inferred sequence and structure binding preferences for SF2/ASF from in vivo data [123]* a) Predicted sequence parameters are shown use a sequence logo representation and an estimate of the preferred structural context for each base is displayed underneath the logo. b) The bar graph shows the relative RNA structure preferences of SF2/ASF.

preferences for SLM2, RBM4 and SF2/ASF. RNAcontext's predictions are supported by *in vivo* binding data for SF2/ASF: the RNAcontext *in vitro* motif model more accurately predicts *in vivo* binding of SF2/ASF, and RNAcontext motif models trained using the *in vivo* data recover the same structural context preference. We expect similar success with our other new predictions because, as we have previously established (in [23]), binding preferences inferred from RNAcompete data are consistent with *in vivo* binding preferences and more accurate prediction of RNAcompete-measured binding affinity translates into more accurate prediction of *in vivo* binding.

We have also provided evidence that the position weight matrix (PWM) motif representation is a better approximation for the RNA binding preferences of RBPs than it is for dsDNA binding preferences of TFs. In particular, in previous work [37], using a similar evaluation framework, we had found that that a "fully-specified 8-mer model" trained on protein-binding microarray (PBM) [32] data had greater predictive power for 7 of 10 TFs than a set of standard DNA motif-finding algorithms, including MatrixREDUCE, trained on the same data. These observations were consistent with many others (e.g.,

[41, 42, 124]) that PWMs were inaccurate approximations dsDNA binding affinities for the majority of TFs. In Figure 3, we show that the opposite holds for RNA-binding data: PWM models learned by MatrixREDUCE, which does not consider RNA structure, had greater predictive power than a fully-specified 7-mer model for a majority of RBPs. Although the sample size is small, this result may reflect the increased flexibility of RNA compared with dsDNA, which may permit more independent movement and recognition of individual bases. Our observations further suggests that modifications of the basic PWM model made for TFs that incorporate interactions between bases may not be as indispensible for modeling RBP binding preferences. Note that our conclusions here differ from our previous analyses on the same data [23]. We suspect that this difference is due our use, in the present study, of motif finding methods that take full advantage of the affinity data associated with each sequence. Indeed, MEMERIS, one of the algorithms we also used in [23] performed worse than the fully 7-mer model in Figure 4.3 for eight of the nine RBPs.

In summary, we have introduced a new motif model of RBP binding preferences and a corresponding algorithm for fitting this model to quantitative estimates of RBP binding affinity for short RNA sequences. Our RNAcontext model makes use of a new technique for representing RNA structure based on a structural context alphabet that we use to annotate individual bases of RNA sequence. This representation is particularly amenable to modeling RBP binding preferences. Although we provide a pipeline to annotate RNA sequences according to the PLUM alphabet, our motif finding code does not require a particular structural context annotation alphabet for bases or even a particular RNA structure prediction method. Hence, RNAcontext can easily be expanded to integrate more parsimonious annotations of structural context or improvements in RNA structure prediction methods.

Chapter 5

MaLaRKey: detailed in vivo binding models

In the previous chapter, we introduced RNAcontext, a motif model that can learn both the sequence and structure preferences of RBPs from large-scale binding data. RNAcontext recovers known binding preferences of several RBPs from RNAcompete data; however, it also has some limitations. To begin with, fitting RNAcontext to *in vivo* data sets is difficult for two main reasons. First, *in vivo* data sets usually contain long sequences (i.e. mRNAs) which are problematic because the noisy-or function that is used by RNAcontext (i.e. to combine k-mer scores) gets saturated when the number of k-mers is large. Also, predicting RNA secondary structure with Sfold becomes infeasible for long sequences. Second, most of the *in vivo* experiments only provide binary binding information (i.e. bound or unbound), and fitting a least squares loss function for this type of data might not be optimal. Another limitation of RNAcontext is the restricted representation of structural information. RNAcontext represents the structure of the binding site by summing profiles of single bases across the binding site. Thus, its structure parameters can only represent overall preferences to the structural context of the binding site. Considering additional structural features would allow a more detailed

representation of structural preferences. For instance, rather than summing profiles of single positions across the binding site, probabilities for the entire binding site to be in different structural contexts can be considered. Indeed, it has been shown that the accessibility of the entire binding site is a better predictor than single-base accessibility when predicting RBP binding [125].

In this chapter, we introduce a new motif finding method, MaLaRKey (Multilinear regression RNAContext), that extends RNAcontext by incorporating a much richer model of secondary structure preferences and by using a motif scoring method suitable for sequences of arbitrary length. MaLaRKey's motif models are fit to RNA binding data from large-scale assays by using multilinear regression to maximize the agreement between the RNA sequence affinity predicted by the motif model and that measured (or implied) by the binding assay. MaLaRKey uses a feature-based product model to represent RBP binding affinity for a given site. Although MaLaRKey can make use of arbitrary additional sequence and structural features, we have had success by adding just three: an overall preference for a secondary structure context of the binding site (i.e. RNAcontext's structure representation, e.g. does the RBP prefer its site to be in a hairpin loop?); predicted transitions between paired and unpaired bases in the binding site (e.g. does the RBP prefer the first base in its binding site to be paired and the rest in a hairpin loop?); and the tendency for the entire binding site to share the same secondary structure context.

5.1 Methods

5.1.1 Overview of input and output of MaLaRKey

The input to MaLaRKey is a set of sequences $S = \{S^1, S^2, ..., S^N\}$ and their associated binding affinities to the RBP of interest $R = \{r^1, r^2, ..., r^N\}$. When there is only information on whether or not a given sequence is bound, we assign $r^i = 1$ for bound sequences

and $r^i = 0$ for unbound sequences.

The output of MaLaRKey consists of RBP preferences for various sequence and structure features, and also the predicted affinities of the RBP for the given input sequences.

5.1.2 Structural annotation of RNA sequences

Previous studies of RNA secondary structure prediction methods show that considering the ensemble of all possible structures instead of the one with the minimum free energy improves the prediction accuracy [75]. Also, there is evidence that mRNAs fold locally instead of globally [82, 84, 126]. For these reasons, we decided to use RNAplfold [81] that considers the ensemble of all possible structures to calculate probabilities for a base or a set of bases to be unpaired with a maximal base pair span of L nucleotides. RNAplfold employs local folding by predicting structure in sliding windows of length W, and averaging base pair probabilities over all windows that contain the base pair. RNAplfold was modified [1] to output probabilities for not only unpaired context, but also for its subclasses (e.g. hairpin loop, internal loop). If the sequences are short (i.e. less than 80-nts), we assigned -W and -L to be equal to the length of the sequence. Otherwise, we ran RNAplfold with parameters -W 200 -L 150 for human transcripts and -W 80 -L 40 for yeast transcripts.

RNAplfold uses a backtracking procedure for the partition function-folding algorithm to calculate base pair probabilities or probabilities for multiple positions to be in unpaired context [79]. This procedure does not allow calculating probabilities for multiple positions to be in paired context, or probabilities of transitioning from a paired context to an unpaired context. In order to estimate these probabilities, we ran RNAshapes to enumerate all secondary structures that had predicted free energies within 70% of the minimum free energy (MFE) using the following call: RNAshapes -s -c 70.0 -r -M 30 -t 1 -o 2. We then parsed individual structures to scan for patterns of interest (e.g. PP, PU, PPU

[1] by Wei Jiao

where P stands for paired context and U stands for unpaired context). The probability of a structure pattern is then calculated as the sum of probabilities of structures (with energies within 70% of the MFE) containing that pattern. When the sequences are long, similar to RNAplfold, we ran RNAshapes in windows and averaged the probabilities across the windows that the base is in.

5.1.3 Motif model

Given a set of input sequences $S = \{S^1, S^2, ..., S^N\}$, we define $L = \{L^1, L^2, ..., L^N\}$ to be the set of feature values of the corresponding sequences. L^i contains an element for each K-mer (subsequence of length K) of the sequence S^i. Let L^{ij} be the set of feature values for the K-mer that starts at j-th position of S^i. L^{ij} consists of a set of vectors with each vector containing the values for a single feature. For example, if feature t is defined as "nucleotide at position 2 of the K-mer", L_t^{ij} is a vector containing four elements (i.e. L_{t1}^{ij}, L_{t2}^{ij}, L_{t3}^{ij},L_{t4}^{ij}) that correspond to the four nucleotides A,C,G,U. We require that the feature values are always non-negative for all K-mers.

MaLaRKey associates a non-negative parameter vector Θ^t with each feature t and uses these parameters to predict the score of a sequence, $f(S^i, L^i, \Theta)$:

$$f(S^i, L^i, \Theta) = \sum_{j=1}^{|S^i|-K+1} \prod_{t=1}^{T} c_t^{ij} \tag{5.1}$$

where T is the number of features and c_t^{ij} is the contribution of feature t for the K-mer that starts at j-th position of sequence i. The sum iterates over all K-mers, and a K-mer score is calculated as the product of feature contributions. Contribution of a feature is defined as follows:

$$c_t^{ij} = \sum_{n=1}^{n_t} \Theta_n^t L_{tn}^{ij} \tag{5.2}$$

where n_t is the number of states for feature t.

5.1.4 Feature design

Figure 5.1 summarizes the features that we consider in our motif model. The first set of features represents the sequence preferences with a Position-Specific Affinity Matrix (PSAM) that has been previously used in MatrixREDUCE model [36]. Elements of PSAM represent the change in the binding affinity when the corresponding position is mutated from the wild-type binding site. MaLaRKey represents the PSAM by fitting a feature for each position of the motif where each feature is associated with a vector Θ^t that consists of four parameters corresponding to the four possible bases. The second set of features is defined to represent the structure preferences. The "single nucleotide structural context feature" models each nucleotide independently and calculates the expected number of bases in each structural context by summing the annotation profiles across the positions of the binding site (i.e. equivalent to RNAcontext's structure representation). If the structures are annotated with paired, hairpin loop, bulge or internal loop, multiloop and external loop (i.e. unstructured) contexts, this feature contains five parameters corresponding to each of these contexts. Note that bulge and internal loop contexts are combined into a single context since RNAplfold does not distinguish them. The "full site structural context feature" calculates the probability of the entire site to be in specified structural contexts. Since the probabilities for multiple positions to be in paired context cannot be calculated with RNAplfold, the states for this feature consist of hairpin loop, bulge or internal loop, multiloop, external loop and dummy state where dummy state represents the remaining probability mass. Another structure feature that we consider models the 5' transition from paired to unpaired state (or from unpaired to paired state). Namely, the states of this feature correspond to the probability of the 5' transition from a paired to an unpaired state at a specific position within the binding site. For example, this feature can represent the preference for the first base of the binding site to be in

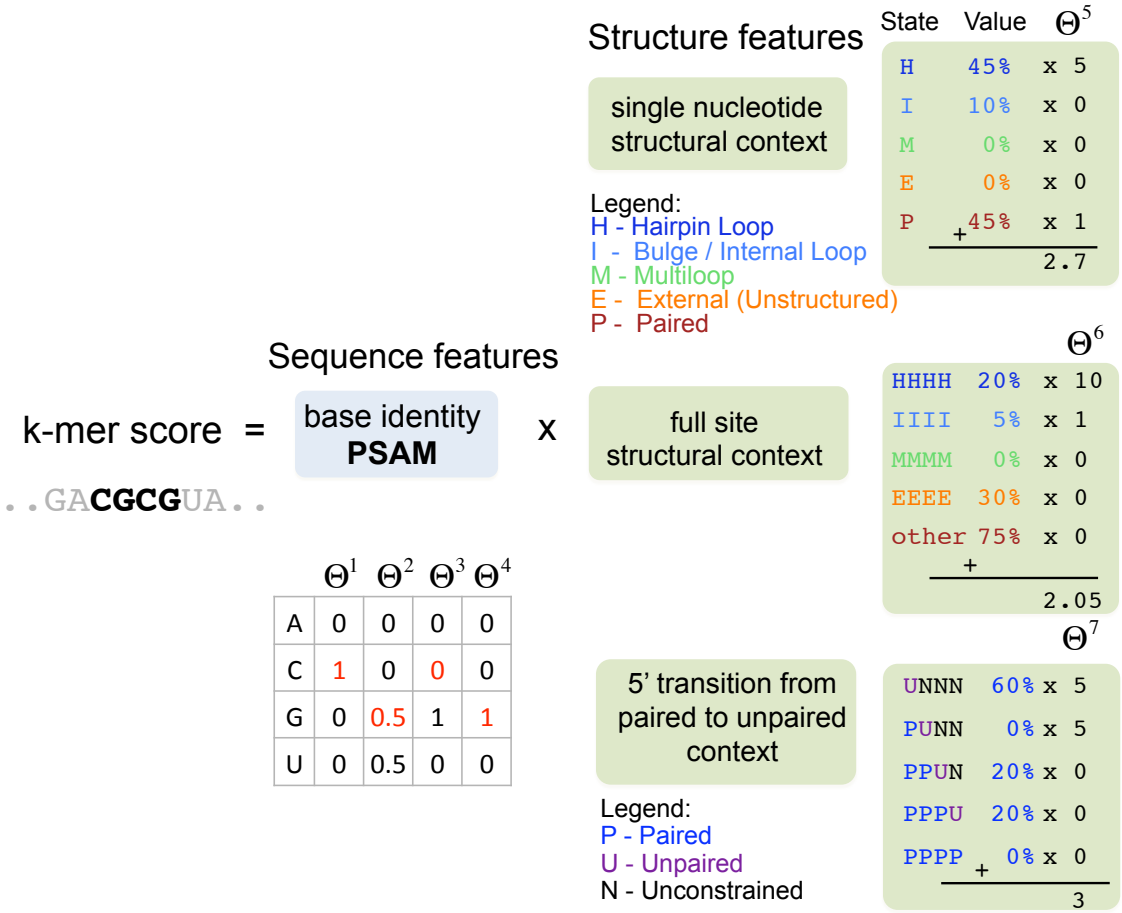


Figure 5.1: *Overview of MaLaRKey method: the input, parameters and motif model.* The input to MaLaRKey consists of a set of sequences together with their associated structure annotation profiles (estimated using RNAplfold and RNAshapes) and RNA-binding affinity estimates for the given RBP. Each feature t is associated with a parameter vector Θ^t. In this figure, random values are assigned to the entries of Θ^t for illustration purposes. The sequence features are represented with a Position-Specific Affinity Matrix (PSAM) [36]. Boxes on the right show the structure features that are considered: single nucleotide structural context feature, full site structural context feature, 5' transition from paired to unpaired state within a binding site. "N" indicates that the corresponding nucleotide is unconstrained, i.e., paired or unpaired. Note that 5' transition from unpaired to paired state is not shown in this figure.

paired context and the rest of the binding site to be in a hairpin loop context.

5.1.5 Parameter estimation

Parameters are fit by minimizing two different cost functions $E(\Theta)$ depending on the provided binding data. If quantitative binding affinity data is provided, we minimize

the sum of the squared differences, between the measured affinities r^i and our predicted affinities $\hat{r}^i$:

$$E(\Theta, \beta) = \sum_{i=1}^{N} (r_i - \hat{r}_i)^2 + \lambda(\sum_{t,n} {\Theta_n^t}^2) \tag{5.3}$$

We predict the affinity of S^i of the sequence score as $f(S^i, L^i, \Theta) + \beta$, where β is the bias term. We also include a L2 regularization term with λ as the regularization constant.

If only bound or unbound information is provided we assign $r^i = 1$ for bound sequences and $r^i = 0$ for unbound sequences, and minimize the cross entropy error between r^i and $\hat{r}^i$:

$$E(\Theta, \beta) = \sum_{i}^{N} [-r_i \log(\hat{r}_i) - (1 - r_i) \log(1 - \hat{r}_i)] + \lambda(\sum_{t,n} {\Theta_n^t}^2) \tag{5.4}$$

where $\hat{r}_i$ is defined as $\sigma(f(S^i, L^i, \Theta) + \beta)$. $\sigma(x) = (1 + \exp(-x))^{-1}$ corresponds to the logistic function that scales x to the range of [0,1].

We use L-BFGS-B package [118] to optimize our cost function where we utilize box constraints to enforce positivity constraints for Θ. Different initializations can give different results when running MaLaRKey. As such, it is important to start the search from a good initialization point. For this purpose, we use a pre-filtering procedure to determine the K-mers that are most predictive. That is, we input counts of all possible k-mers as features and affinity measurements as response variables to glmnet package [127] and fit a logistic regression model with lasso penalty. We use the top five k-mers with the highest coefficients as initialization points. For each of these k-mers, we initialize the sequence parameters by assigning 1 for the consensus base in each column and 0.1 for the non-consensus bases. Also, we determine the optimal motif length by calculating the average cross-validation area under ROC curve (AU-ROC) on glmnet predictions, and we choose the length that gives the minimum average cross-validation AU-ROC.

While learning the parameters, we first fit the sequence parameters using five different

initializations and choose the parameter setting that has the lowest error (on the training set). Then, to determine which structure feature to add, we fix the sequence parameters and only fit one structure feature at a time. In the default setting, we consider two alternative structure features: single nucleotide structural context feature and full site structural context feature. We choose the one that reduces the error most. Finally, we fit all the parameters (i.e. sequence features and the selected structural context feature) to determine the final settings of the parameters. If we have prior knowledge that the RBP binds to stem-loops, we also consider the 5' transition feature by adding it after we fit the first structural feature (i.e. single nucleotide or full site structural context).

5.2 Results

To evaluate the performance of MaLaRKey, we first compare it against RNAcontext on RNAcompete-v1 data. Here, we also show how the 5' transition feature can be used to construct a detailed binding model for Vts1p. Next, we test MaLaRKey's ability to reproduce various binding data that we compiled on mRNA targets of a diverse set of RBPs from yeast and human. We describe our results on binding data for 14 RBPs from RIP-chip data and for seven RBPs from CLIP data.

5.2.1 Comparison of RNAcontext and MaLaRKey

We ran MaLaRKey on binding data for nine RBPs from RNAcompete-v1 experiments (the details of this data is described in Chapter 3). We minimized the least squares error to train MaLaRKey. Except for Vts1p, we chose one of the two structural context features i.e., single nucleotide or full site structural context feature (as described in the previous section). Since we have prior knowledge that Vts1p binds to stem-loops, we checked whether we can model the transition from stem to hairpin loop within a binding site. As such, we considered two structural features for Vts1p: first was selected between

single nucleotide structural context feature and full site structural context feature, and the second was selected between the 5' transition from unpaired to paired context feature and 5' transition from paired to unpaired context feature.

We followed the same procedure described in the previous chapter (i.e. Chapter 4) for model selection. Namely, we fit two motif models for RNAcontext and MaLaRKey. One set of models was fit using only weakly structured sequences, and the other model was fit using the full data set containing both stem-loops and weakly structured sequences. Also, as before, we used the training set AU-PR to choose the optimal motif width, and to choose between the full training set and only the weakly structured sequences. We then scored the test sequences (i.e. full data set) with the selected motif model, and calculated AU-PR. We used RNAplfold (rather than Sfold) to calculate profiles for single nucleotide and full site structural context features, and RNAshapes to calculate the profiles for transition features. Because of this change, RNAcontext AU-PRs for some RBPs are different when compared to the results in the previous chapter (Table 4.3, first column). The first and second columns of Table 5.1 show the average AU-PRs across the two cross-validation sets (i.e. Set A and B) for RNAcontext and MaLaRKey. The AU-PRs of the two methods are the same for SLM2, SF2, HuR. MaLaRKey gives a better performance for FUSIP1 and Vts1p, and RNAcontext gives a better performance for RBM4, YB1, U1A and PTB. Since RNAcontext's motif scoring is more suitable for short sequences, it is not surprising that its predictions are more accurate than MaLaRKey for a number of RBPs. However, many of these differences are small suggesting that both algorithms can be applied to large binding data of short sequences.

Figure 5.2 shows MaLaRKey's predicted sequence and structure preferences for Vts1p when the entire data set (both Set A and Set B) is used for training. The sequence logo contains a G 5' to the core loop sequence CNGG (Figure 5.2-a). Since the predicted motif includes some flanking in addition to the core binding site, the single nucleotide structural context feature is selected. The bar graph on the bottom left (Figure 5.2-b)

Proteins	RNAcontext	MaLaRKey
RBM4	**0.91**	0.90
FUSIP1	0.34	**0.46**
Vts1p	0.65	**0.72**
YB1	**0.17**	0.16
SLM2	0.81	0.81
SF2	0.69	0.69
U1A	**0.42**	0.21
HuR	0.96	0.96
PTB	**0.72**	0.70

Table 5.1: *Comparison of predictive accuracy of RNAcontext and MaLaRKey on nine RBPs from RNAcompete-v1 data.* The values show the average AU-PR across two test sets. The first and second columns show the results for RNAcontext and MaLaRKey when motif lengths between 4 and 12-nts (inclusive) are considered, and the secondary structure profiles are predicted with RNAplfold. Bold values indicate the best performing method between the first and second columns.

shows that there is a strong preference for hairpin loop. Among the transition features, 5' transition to unpaired context is chosen, and the transition is predicted to occur after the first position (Figure 5.2-c). Taken together, learned sequence and structure parameters recover Vts1p's known binding preferences (i.e. preference to hairpin loops that contain the CNGG motif with closing base pair G-C), and allow constructing a more detailed representation of the binding site. We also checked whether the inclusion of 5' transition feature improves the prediction accuracy. The sequence model alone results in an average AU-PR of 0.56. The inclusion of the single nucleotide structural context feature increases the average AU-PR to 0.65. Finally, the addition of 5' transition feature further improves the average AU-PR to 0.72 confirming that the higher performance of MaLaRKey for Vts1p is due to the addition of 5' transition feature.

5.2.2 RIP-chip data

Hogan et. al. queried for the RNA targets of 40 RBPs in yeast using RIP-chip technique [6]. We used a subset of this data that had a large enough number of transcripts to fit a statistical model. The selection of the subset of the data and the definition of bound and

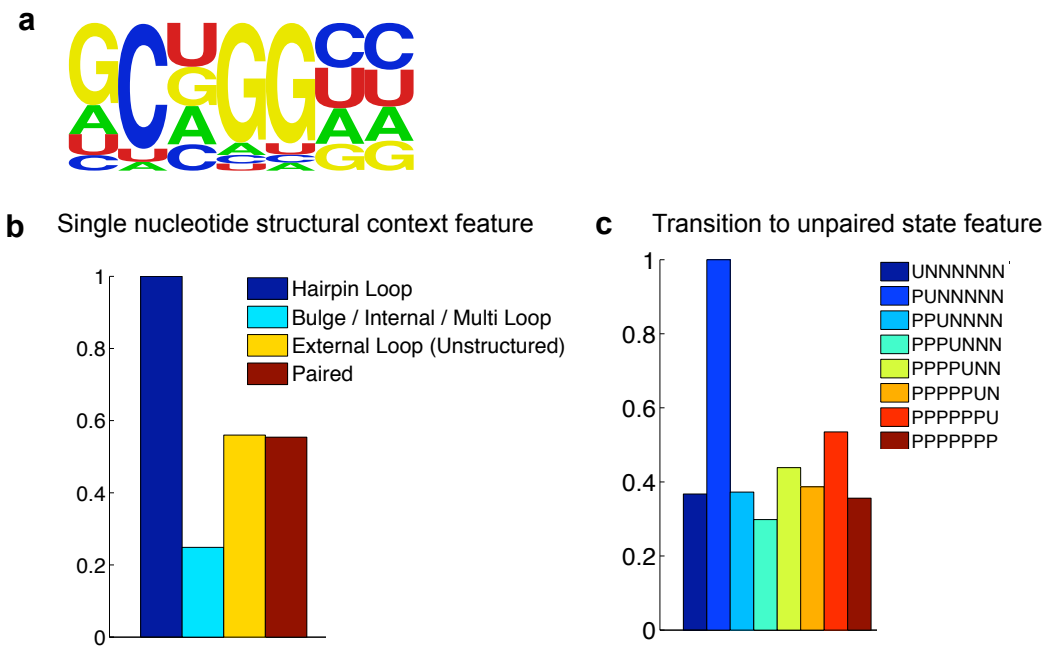


Figure 5.2: *Inferred sequence and structure preferences of Vts1p from RNAcompete-v1 data.* a) Sequence logo representation of inferred sequence preferences (generated by enoLOGOS software [121]). b-c) Relative structure preferences are shown with bar graphs where the parameters associated with the structure feature are scaled so that the maximum value is 1. Note that we combined bulge / internal / multi loop into a single context to compare against Chapter 4 results. b) single nucleotide structural feature. c) 5' transition to unpaired state feature.

unbound transcripts are based on a previous study [125]. Table 5.2 shows the number of bound and unbound transcripts for each of the 14 data sets that we used. Note that many of the data sets are unbalanced, thereby exacerbating the classification problem. Also, the length of the input sequences (i.e. full-length transcripts) is quite long (i.e. average length is 1851) compared to a typical motif-finding problem.

We fit MaLaRKey to binding data for 14 RBPs by minimizing the cross entropy error. To assess the confidence in inferred structure preferences, we ran MaLaRKey on 100 training sets (for each RBP) that are created by bootstrapping. We used stratified bootstrapping to ensure that each training set contains similar proportions of bound and unbound transcripts. We used the transcripts that are not selected for the training set as the test set, and evaluated the fitted motif models by predicting binding of these

RBPs	bound	unbound	total
Vts1p	121	1433	1554
Khd1	567	3108	3675
Gbp2	212	2180	2392
Nab2	620	962	1582
Ypl184	305	174	479
Scp160	1230	303	1533
Pub1	1378	1034	2412
Bfr1	985	790	1775
Puf3	212	2608	2820
Puf4	219	3889	4108
Puf5	178	949	1127
Yll032c	49	1872	1921
Nop56	292	105	397
Tdh3	125	452	577

Table 5.2: *The number of bound and unbound transcripts for 14 RBPs from RIP-chip data [6].*

transcripts.

Figure 5.3 contains the inferred sequence and structure binding preferences. Since displaying each of the 100 sequence models is difficult, we only show the sequence model that is learned from the entire data set (third column). MaLaRKey recovers previously reported sequence preferences for most of the RBPs [6, 125, 128]. For example, MaLaRKey correctly predicts that Vts1p binds to CNGG motifs, and Puf proteins (i.e. Puf3, Puf4, Puf5) prefer sequences that contain the UGUA core motif. The bar graph next to the sequence logo displays the median of the structure parameters that are learned from bootstrapped training sets. Note that the parameters associated with the structure feature are scaled so that the maximum value is 1. The structure feature that is selected in maximum number of training sets (i.e. out of 100) is shown on the left of the bar graph: *single* refers to the single nucleotide structural context feature and *full* corresponds to the full site structural context feature. MaLaRKey recovers the fact that Vts1p binds to hairpin loops with the CNGG motif in a hairpin loop, and Puf proteins (i.e. Puf3, Puf4, Puf5) recognize single-stranded regions [18, 111]. For four RBPs with unknown

structure preferences (i.e. Khd1, Ypl184, Scp160 and Bfr1), MaLaRKey predicts a strong preference to external regions.

We evaluated the fitted motif models by calculating area under ROC curve (AU-ROC) and area under precision-recall curve (AU-PR) on the test set predictions. The right-most four columns of Figure 5.3 show the average AU-ROC and AU-PR of the complete MaLaRKey model and a simpler version of it that only includes the sequence features. By comparing the performances of the two models, we can evaluate the added predictive value of structure features. Table 5.3 shows the p-values of two-sided Wilcoxon sign rank test on 100 AU-ROC or AU-PR pairs for each RBP. Including structure information improves the predictions significantly for 9 out of 14 RBPs (p-value < 0.005). The improvement due to structure features is prominent for RBPs such as Khd1, Vts1p, Ypl184 and Gbp2. Including structure information does not improve the prediction accuracy for RBPs Puf3, Puf5, Scp160, Nop56 and Bfr1. MaLaRKey can be easily modified such that the inferred structure preferences are not reported when their inclusion does not improve the prediction accuracy.

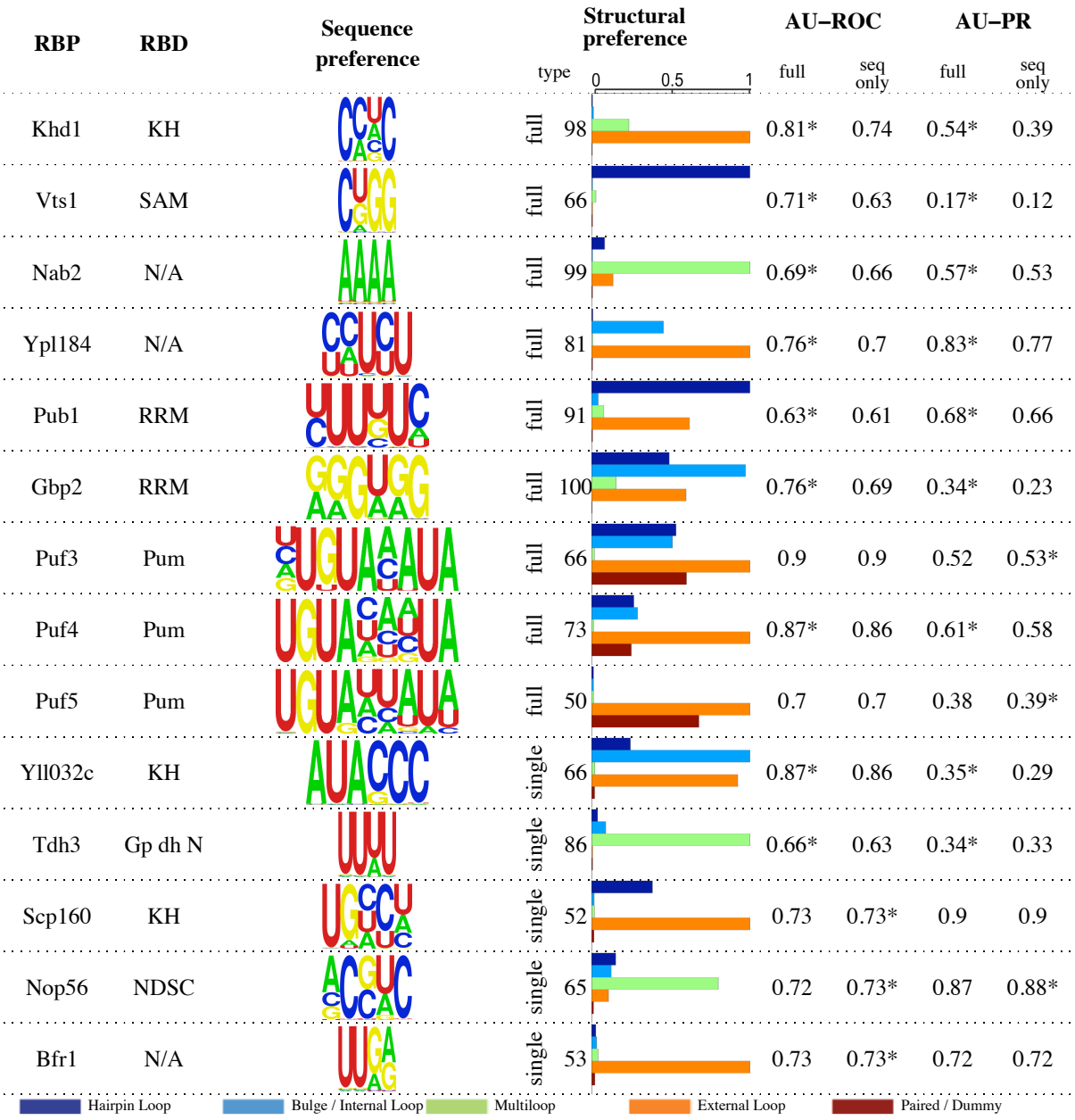


Figure 5.3: *Inferred sequence and structure preferences of yeast RBPs from RIP-chip data [6].* First and second columns display the names and the RNA-binding domains of the RBPs. Sequence preferences are shown with a motif logo representation (third column). In the fourth column, the type of the structural context feature that is selected in the majority of the bootstrap runs is reported. This column also shows the number of times this feature is selected among the 100 bootstrap runs. In the fifth column, relative structure preferences are shown with a bar graph where the parameters of the selected structure feature are scaled so that the maximum value is 1. The values show the median of the bootstrap results. Lastly, AU-ROC and AU-PR values are displayed for the complete MaLaRKey model and a simpler version of it containing only sequence features. * indicates that the difference is significant according to two-sided Wilcoxon sign rank test (p-val < 0.005).

Proteins	AU-ROC	p-value	AU-PR	p-value
Khd1	complete	3.9e-18	complete	3.9e-18
Vts1p	complete	8.1e-16	complete	1.3e-15
Nab2	complete	3.9e-18	complete	3.9e-18
Ypl184	complete	4.0e-18	complete	5.0e-18
Pub1	complete	7.3e-18	complete	3.9e-18
Gbp2	complete	3.9e-18	complete	3.9e-18
Puf3	-	1.6e-01	sequence-only	1.5e-06
Puf4	complete	2.1e-15	complete	2.3e-14
Puf5	-	4.1e-01	sequence-only	7.6e-07
Yll032c	complete	4.0e-12	complete	1.6e-13
Tdh3	complete	1.4e-14	complete	7.5e-10
Scp160	sequence-only	1.5e-03	-	4.2e-02
Nop56	sequence-only	3.1e-05	sequence-only	8.2e-04
Bfr1	sequence-only	4.5e-04	-	1.0e-01

Table 5.3: *P-values for Wilcoxon sign rank test on 100 AU-ROC and AU-PR pairs.* The performance of the "complete" model is significantly better than the "sequence-only" model for RBPs Khd1, Vts1p, Nab2, Ypl184, Pub1, Gbp2, Puf4, Yll032c and Tdh3. Including structure information either results in no change or gives significantly worse performance for the remaining set of RBPs (i.e. Puf3, Puf5, Scp160, Nop56, Bfr1). Though, the p-values for these RBPs are higher when compared to the p-values for RBPs for which including structure information improves the prediction accuracy. Entries with "-" shows that the hypothesis (i.e. the difference of the two distributions come from a distribution whose median is 0) cannot be rejected at the 5% significance level.

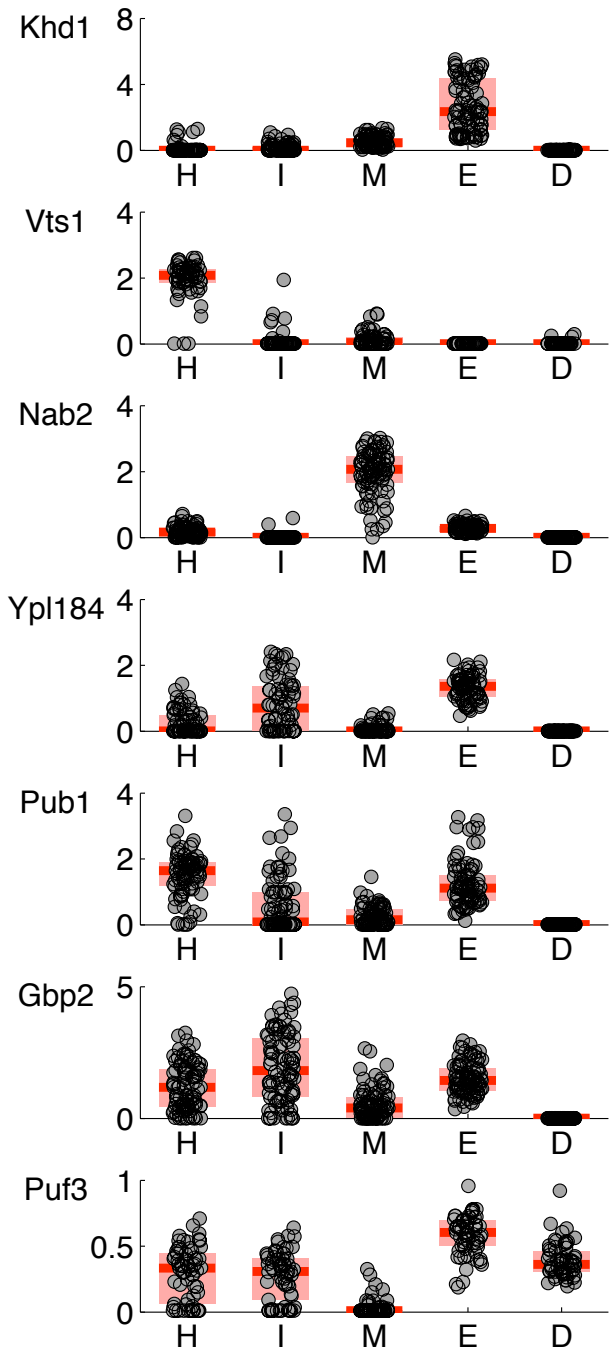
We next assessed the uncertainty in inferred structure preferences by plotting the distribution of structure parameters learned from bootstrapped training sets. Note that we only plot a subset of the structure parameters inferred from 100 bootstrapped training sets. We identify the structural feature (i.e. either single nucleotide or full site) that is selected in the majority of the runs and only plot the distribution of parameters of this feature type. Figure 5.4 and Figure 5.5 show these distributions with box plots where the red lines show the medians, and the bottom and top of the pink boxes correspond to the 25th percentile and 75th percentile, respectively. The tables next to the box plots compare the distributions of structure parameters using two-sided Wilcoxon sign rank test. For each RBP, we report eight comparisons i.e., four comparisons between the most preferred structural context (i.e. that has the highest median value) and remaining

structural contexts, and four comparisons between the paired / dummy structural context and remaining structural contexts. We observe that the magnitude of the structure parameters is larger for RBPs for which including structure information is helpful. For instance, the median values for the preferred structure context for Khd1 and Vts1p are above two. On the other hand, most of the values of the structure parameters are below one for RBPs Scp160 and Bfr1. Also, we see a clear difference between the preferred structural context and the other structural contexts for RBPs for which structure is important. For example, Khd1 and Puf4 prefers to bind external regions, and Vts1p prefers to bind hairpin loops. RBPs such as Pub1, Gbp2 and Ypl184 seem to have preferences for multiple contexts. For RBPs for which structure results in no gain of performance (e.g. Puf3, Nop56), the distributions of structural contexts are highly overlapping.

For the RBPs for which the inclusion of structure features does not improve the prediction accuracy (i.e. Scp160, Nop56, Bfr1, Puf3, Puf5), we investigated whether this is due to lack of signal in the data or MaLaRKey's inability to learn the correct structure preferences. We first determined the highest scoring k-mer (which will be called top k-mer hereafter) under the sequence model. Then, we found the occurrences of the top k-mer in the set of input transcripts, and compared the structure profiles of these occurrences between bound and unbound transcripts. Figure 5.6 shows the mean profile values and standard error for each structure context for RBPs Ypl184 (left) and Scp160 (right). Note that one of the structural contexts differs (i.e. Dummy vs Paired) between the two bar graphs since Ypl184 prefers the full site structural context feature, and Scp160 prefers the single nucleotide structural context feature. The bar graph for Ypl184 clearly shows that the top k-mers within bound sequences have a higher probability to be in external regions when compared to unbound transcripts. This difference is reflected in strong preference to external regions predicted for Ypl184. On the other hand, the profiles of the top k-mer occurrences between bound and unbound sequences are indistinguishable for Scp160. As expected, inclusion of structure information does not change the prediction performance

for Scp160.

We next checked whether fitting MaLaRKey with the alternative structure feature (the one that is not selected) changes the prediction performance. Namely, we fit MaLaRKey with the single nucleotide structural context feature on binding data for Vts1p, and with full site structural context feature on binding data for Tdh3, Nop56 and Yll032c. We found that the average AU-ROC decreases from 0.71 to 0.63 for Vts1p. The AU-ROCs remained the same for Nop56, Tdh3 and Yll032. Therefore, we decided to keep the two alternative structural context features; however, we note that the selection of single nucleotide structural context feature usually indicates a weak preference for structure. Although rare, the single nucleotide structural context feature can be useful for RBPs that prefer to bind paired regions because the alternative full site structural context feature cannot represent pairedness.



	H	I	M	D / E
E	8.3e-18	8.3e-18	8.6e-18	8.3e-18
D	8.3e-18	1.6e-08	8.3e-18	8.3e-18

	I	M	E	D / H
H	3.9e-12	3.9e-12	5.1e-12	2.2e-12
D	4.9e-02	5.8e-06	2.5e-01	2.2e-12

	H	I	E	D / M
M	1.1e-17	8.3e-18	9.6e-18	8.1e-14
D	2.6e-16	8.3e-18	5.7e-18	8.3e-18

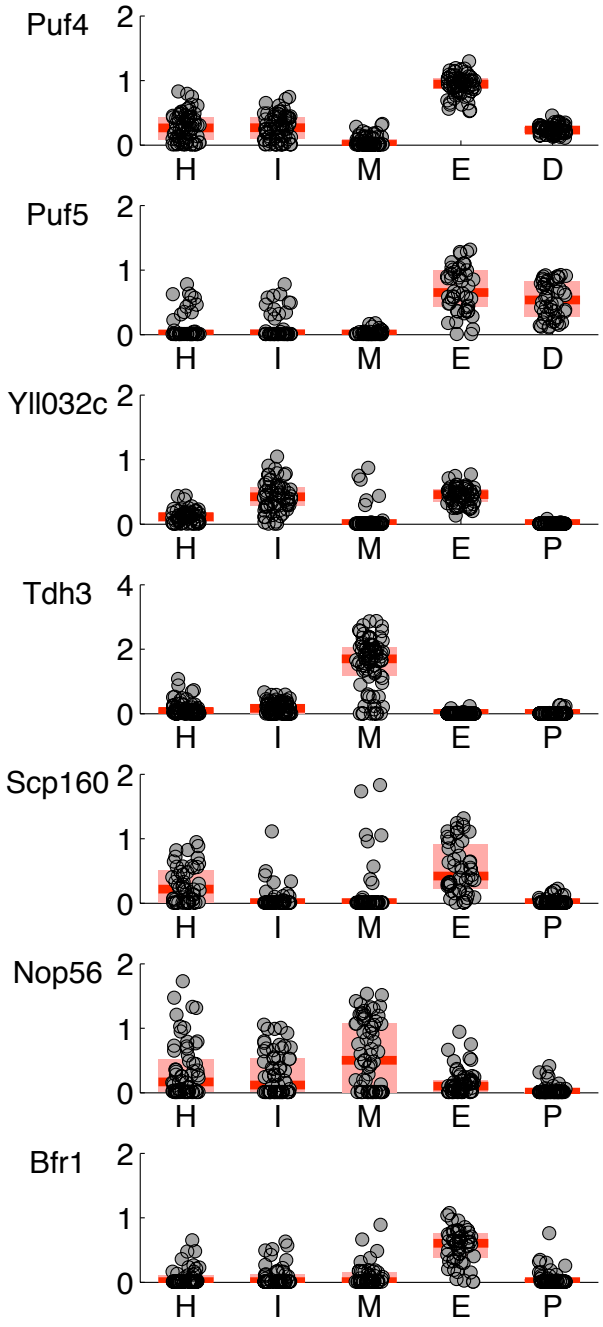
	H	I	M	D / E
E	1.3e-14	1.1e-04	5.4e-15	5.4e-15
D	1.1e-07	5.1e-11	4.0e-05	5.4e-15

	I	M	E	D / H
H	1.0e-04	7.8e-16	3.9e-04	5.5e-16
D	1.1e-09	7.6e-12	1.2e-16	5.5e-16

	H	M	E	D / I
I	1.1e-03	5.1e-14	3.4e-02	1.2e-16
D	5.6e-17	3.6e-14	3.9e-18	5.5e-16

	H	I	M	D / E
E	1.5e-08	1.6e-19	1.7e-12	7.5e-10
D	7.5e-02	1.4e-02	1.6e-12	7.5e-10

Figure 5.4: *Distribution of structure parameters inferred from bootstrapped training sets.* The box plots show the distributions of structure parameters for each RBP. Each box plot displays the parameters for structural contexts hairpin loop (H), internal loop (I), multiloop (M), external loop (E), paired (P) or dummy (D). Note that these plots are different from traditional box plots in that they display all the points rather than a subset of the points that do not reside in inter-quartile range. The tables next to the box plots display the p-values that are calculated by comparing the distributions of the structural contexts using two-sided Wilcoxon sign rank test. Highlighted entries indicate that the p-values are not significant i.e., ≥ 0.0005 considering 10 multiple tests.



	H	I	M	D / E
E	3.0e-13	2.5e-13	1.1e-13	1.1e-13
D	3.3e-01	1.9e-01	4.8e-13	1.1e-13

	H	I	M	D / E
E	2.3e-08	1.9e-08	1.6e-09	6.6e-02
D	8.7e-07	7.2e-07	8.0e-10	6.6e-02

	H	M	E	P / I
I	6.4e-10	8.7e-10	8.4e-01	5.2e-12
P	5.1e-11	2.4e-02	1.6e-12	5.2e-12

	H	I	E	P / M
M	3.8e-14	1.5e-14	1.7e-14	7.9e-15
P	8.6e-10	2.0e-09	4.2e-01	7.9e-15

	H	I	M	P / E
E	4.6e-03	5.2e-08	2.0e-06	3.4e-09
P	1.9e-07	3.7e-01	4.8e-01	3.4e-09

	H	I	E	P / M
M	1.3e-02	7.1e-04	2.2e-06	6.5e-09
P	9.5e-07	3.8e-06	1.1e-08	6.6e-09

	H	I	M	P / E
E	9.4e-10	1.4e-09	1.3e-09	3.1e-09
P	9.8e-10	2.9e-01	4.1e-01	3.1e-09

Figure 5.5: *Distribution of structure parameters inferred from bootstrapped training sets.* The box plots show the distributions of structure parameters for each RBP. Each box plot displays the parameters for structural contexts hairpin loop (H), internal loop (I), multiloop (M), external loop (E), paired (P) or dummy (D). Note that these plots are different from traditional box plots in that they display all the points rather than a subset of the points that do not reside in inter-quartile range. The tables next to the box plots display the p-values that are calculated by comparing the distributions of the structural contexts using two-sided Wilcoxon sign rank test. Highlighted entries indicate that the p-values are not significant i.e., ≥ 0.0005 considering 10 multiple tests.

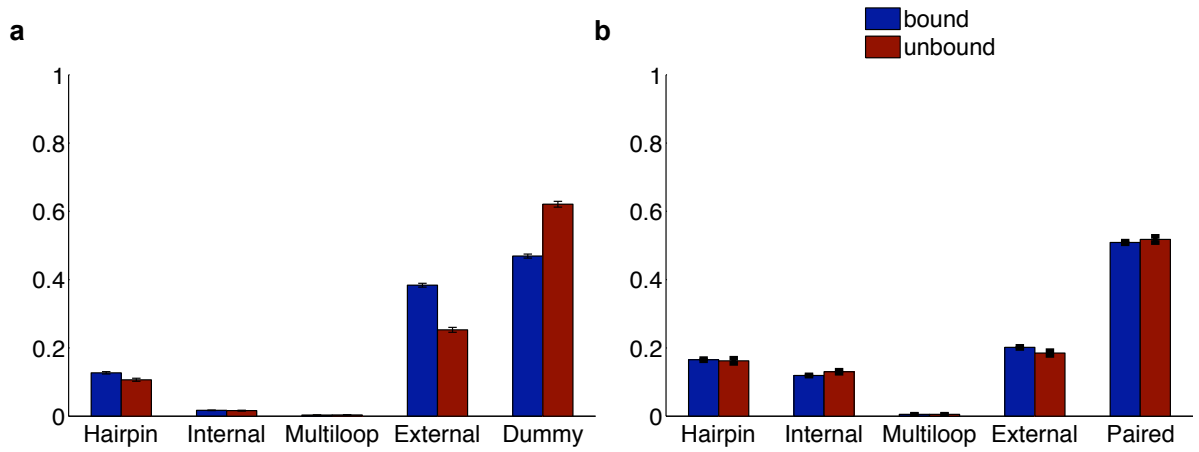


Figure 5.6: *Mean profile values for the top k-mer occurrences in bound and unbound sequences for RBPs a) Ypl184 b) Scp160.* The error bars show the standard error of the mean. Since full site structural context feature is selected for Ypl184, the profiles are calculated such that the whole k-mer is in the same structural context. On the other hand, single nucleotide structural context feature is selected for Scp160. As such, expected number of bases in each structural context was calculated by summing individual profiles across the positions of the k-mer. The profiles show these values divided by the length of the k-mer.

Next, we checked whether we could recover Vts1p's preferences that we inferred from RNAcompete-v1 data when we fit MaLaRKey to RIP-chip data for Vts1p. To do this, we followed the procedure that we previously used to fit 5' transition feature on RNAcompete-v1 data. Figure 5.7 shows the predicted sequence and structure preferences for Vts1p. Predicted preferences confirm that Vts1p prefers hairpin loops with the CNGG motif where the closing base pair is a G-C pair. We repeated our evaluation using bootstrapped training /test sets mentioned in the previous section to check whether the inclusion of transition feature improves the prediction accuracy. The motif model that only considers sequence features results in an average AU-ROC of 0.62. Average AU-ROC increases from 0.62 to 0.69 with the addition of single nucleotide structural context feature. The addition of transition feature further improves the average AU-ROC from 0.69 to 0.76 confirming its high predictive value.

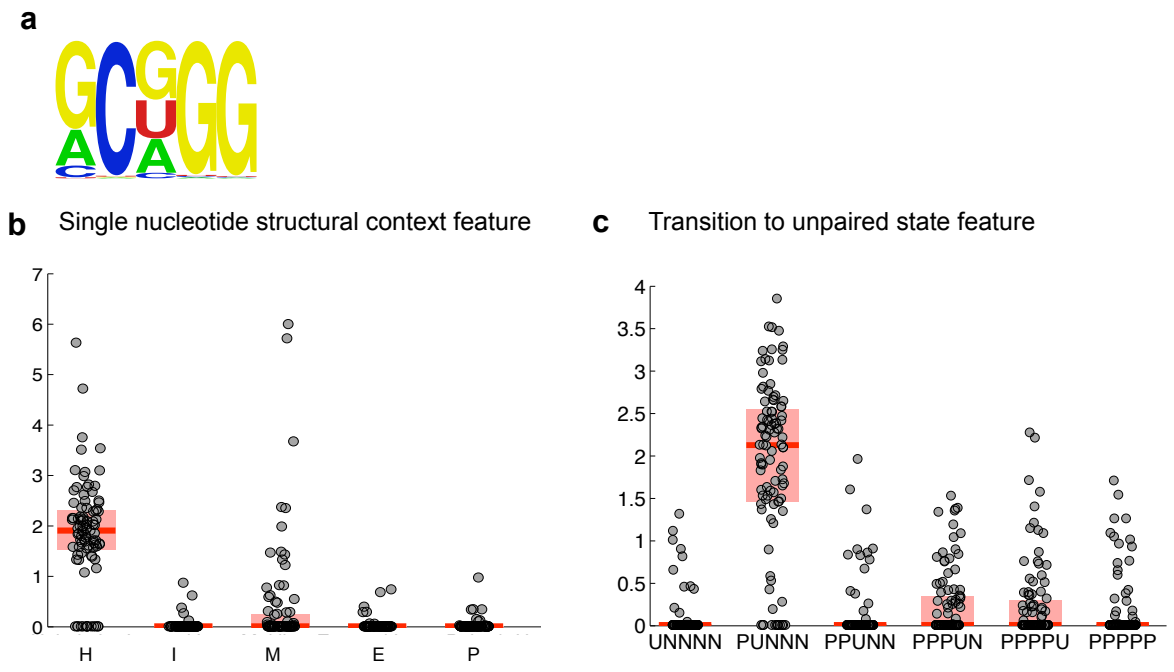


Figure 5.7: *Inferred sequence and structure preferences of Vts1p from RIP-chip data.* a) Sequence logo representation of inferred sequence preferences from the entire data set. b) Box plot shows the structure parameters learned from 93 bootstrapped training sets for which single nucleotide structural context feature is selected. The labels of the x axis corresponds to the structural contexts: H (hairpin loop), I (internal loop), M (multiloop), E (external loop) and P (paired). By comparing the distribution of the first column against the other columns, we can infer a strong preference for hairpin loop context. c) Box plot shows the distribution of parameters for the transition feature that are learned from 100 bootstrapped training sets (i.e. transition to unpaired state feature rather than transition to paired state feature is selected for all of the training sets). Strong preference to PUNNN implies that the transition from paired to unpaired state occurs after the first position.

5.2.3 CLIP data

CLIP-derived binding data of six human RBPs (i.e. IGFBP1-3, QKI, Pum2, TIA1, TIAL1 and TDP-43) were downloaded from doRINA database [129]. Also, CLIP data on the yeast RBP Khd1 were downloaded from the supplementary material of Wolf *et al.* [130]. Since CLIP experiments are followed by high-throughput sequencing to identify the cross-linked sites, the binding data consists of peak locations and their associated scores that are calculated based on the number of reads matching to the peak region. Sequences that correspond to peak locations were extracted from the human (hg18) and

yeast (Sigma1278b) genomes using BEDtools [131] utilities. Some flanking region (i.e. 80-nts for yeast and 200-nts for human) were also included for RNA secondary structure prediction. Sequences with scores in the top five percentile were defined as the bound sequences. If the number of bound sequences based on this selection is less than 1000, the percentile cutoff is reduced to include 1000 sequences. Random non-peak windows of matching length were selected from the same set of transcripts as the peaks to form the set of unbound sequences. These random regions were also constrained to be at least 300-nt away from the end of the peaks. This selection criterion is designed to produce a "negative set" of unbound sequences available for binding but reported not to be bound.

We assigned a score of 1 to bound sequences and 0 to unbound sequences and fit MaLaRKey by minimizing the cross entropy error. The results are shown in Figure 5.8. Our predicted sequence motifs for IGFBP1-3, QKI and Pum2 are highly similar to the motifs identified by the original study that produced the CLIP data [28]. For the RBPs TIA1 and TIAL1, our predicted sequence preferences confirm the finding that these proteins have similar preferences, and bind to U-rich sequences [132]. Sequence motifs for IGFBP1-3, QKI, and TIA1 are further confirmed by RNAcompete identified sequence preferences (See Figure 3-12). We predict that TDP-43 prefers UG-rich sequences that agrees with a previous study showing that TDP-43 binds to UG clusters [133]. Finally, sequence preferences for Khd1 are in accordance with those that we identified from RIP-chip data (Figure 5.3). Including structural features improve the prediction accuracy of IGFBP1-3 and Khd1 (p-val < 0.005 from two-sided Wilcoxon sign rank test). Table 5.4 shows a complete list of p-values of two-sided Wilcoxon sign rank test applied on a set of 100 AU-ROC and AU-PR pairs.

As we did for RIP-chip data, we plot the distributions of structure parameters in order to assess the uncertainty. Figure 5.9 shows these distributions for each of the seven RBPs. For Khd1, MaLaRKey infers a strong preference to external loop, which is in accordance with Khd1's preferences previously inferred from RIP-chip data. Interestingly, we see a

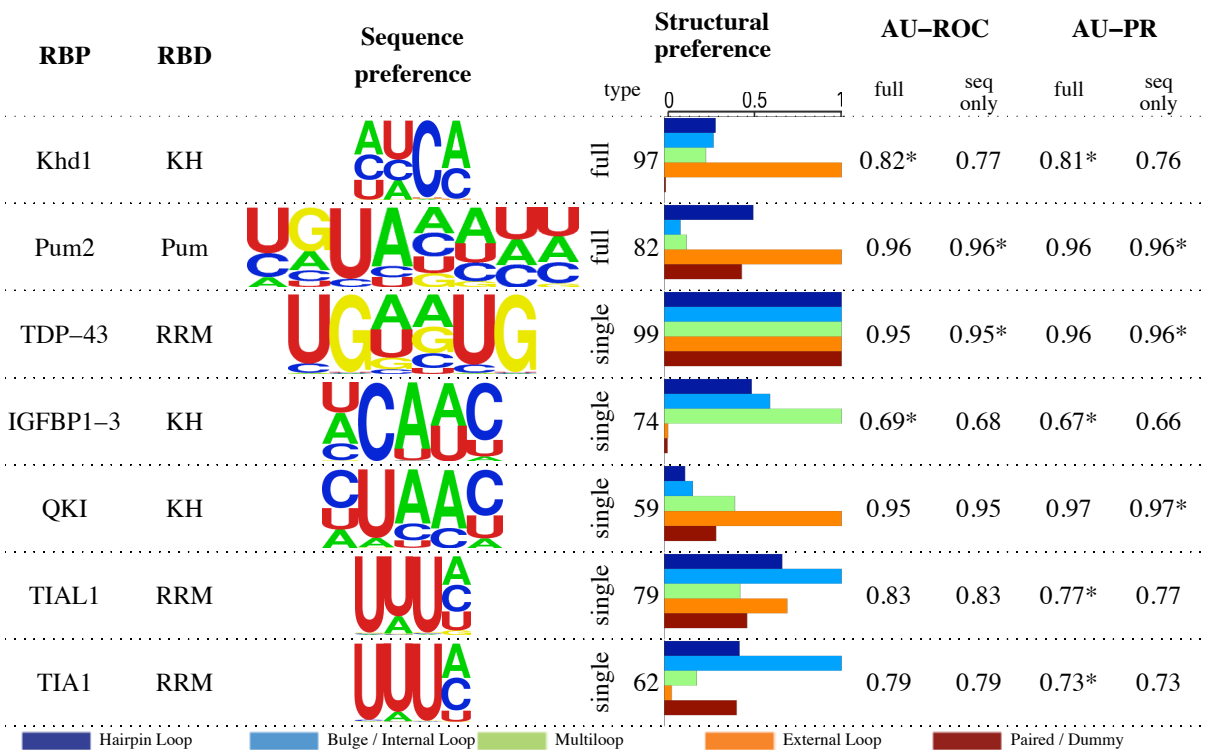


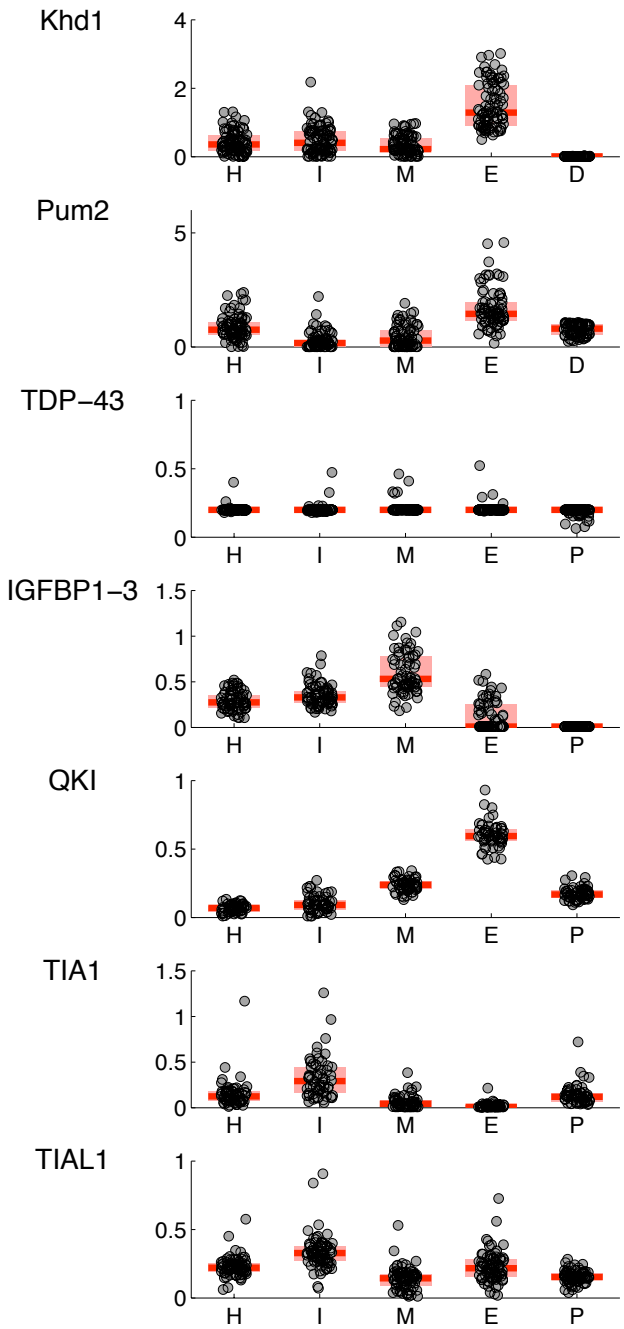
Figure 5.8: *Inferred sequence and structure preferences of seven RBPs from CLIP data.* First and second columns display the names and the RNA-binding domains of the RBPs. Third column shows the learned sequence preferences with a motif logo representation. In the fourth column, relative structure preferences are shown with a bar graph where the parameters of the selected structure feature are scaled so that the maximum value is 1. Lastly, AU-ROC and AU-PR values from the complete MaLaRKey model and a simpler model containing only sequence features are reported. * indicates that the difference is significant (p-val < 0.005, two-sided Wilcoxon sign rank test)

preference to external loop context for Pum2 and QKI, although including structure does not improve the performance for these RBPs.

Similar to our analysis on RIP-chip data, we investigated the profiles of the top k-mer occurrences between bound and unbound sequences. Figure 5.10 shows the mean profile values and standard errors for Khd1 and TIAL1. The profile for Khd1 shows a strong enrichment for external regions in bound sequences, whereas the difference between the profiles for bound and unbound sequences are small for TIAL1.

Proteins	**AU-ROC**	**p-value**	**AU-PR**	**p-value**
Khd1	complete	3.9e-18	complete	3.9e-18
Pum2	sequence-only	8.2e-06	sequence-only	3.1e-10
TDP-43	sequence-only	1.8e-04	sequence-only	6.8e-06
IGFBP1-3	complete	3.9e-18	complete	3.9e-18
QKI	-	1.1e-01	sequence-only	3.5e-03
TIA1	-	2.0e-01	complete	4.2e-11
TIAL1	-	6.3e-01	complete	1.6e-03

Table 5.4: *P-values of Wilcoxon sign rank test on 100 AU-ROC and AU-PR pairs.* The AU-ROCs and AU-PRs of the "complete" model is significantly higher than the "sequence-only" model for RBPs Khd1 and IGFBP1-3. For TIA1 and TIAL1, only AU-PRs of the "complete" model is significantly higher than the "sequence only" model. Including structure information either results in no change or gives significantly worse performance for the remaining set of RBPs (i.e. Pum2, TDP-43, QKI). However, we should note that the p-values for these RBPs are higher when compared to the p-values for RBPs for which including structure information improves the prediction accuracy. The entries with "-" indicates that the hypothesis (i.e. the difference of the two distributions come from a distribution whose median is 0) cannot be rejected at the 5% significance level.



	H	I	M	D / E
E	1.2e-17	1.0e-16	1.4e-17	1.2e-17
D	1.8e-17	2.6e-16	3.7e-16	1.2e-17

	H	I	M	D / E
E	1.1e-09	3.2e-14	1.7e-14	8.1e-14
D	9.1e-02	1.8e-11	7.0e-08	8.1e-14

	H	I	M	P / E
E	1.6e-04	1.9e-03	2.6e-04	1.2e-05
P	1.2e-05	1.2e-05	1.2e-05	1.2e-05

	H	I	E	P / M
M	8.7e-14	3.7e-13	4.2e-13	7.7e-14
P	1.2e-05	1.2e-05	1.2e-05	7.7e-14

	H	I	M	P / E
E	2.4e-11	2.4e-11	2.4e-11	2.4e-11
P	3.3e-11	2.9e-07	9.2e-04	2.4e-11

	H	M	E	P / I
I	7.6e-12	7.6e-12	7.6e-12	7.6e-12
P	6.6e-01	2.4e-08	7.6e-12	7.6e-12

	H	M	E	P / I
I	9.9e-14	1.2e-12	2.1e-09	1.1e-14
P	1.1e-12	9.9e-02	4.0e-07	1.1e-14

Figure 5.9: *Distribution of structure parameters inferred from bootstrapped training sets.*The box plots show the distributions of structure parameters for each of the seven RBPs. Each box plot displays the parameters for structural contexts hairpin loop (H), internal loop (I), multiloop (M), external loop (E), paired (P) or dummy (D). The tables next to the box plots display the p-values that are calculated by comparing the distributions of the structural contexts using two-sided Wilcoxon sign-rank test. Highlighted entries indicate that the p-values are not significant i.e., ≥ 0.0005 considering 10 multiple tests.

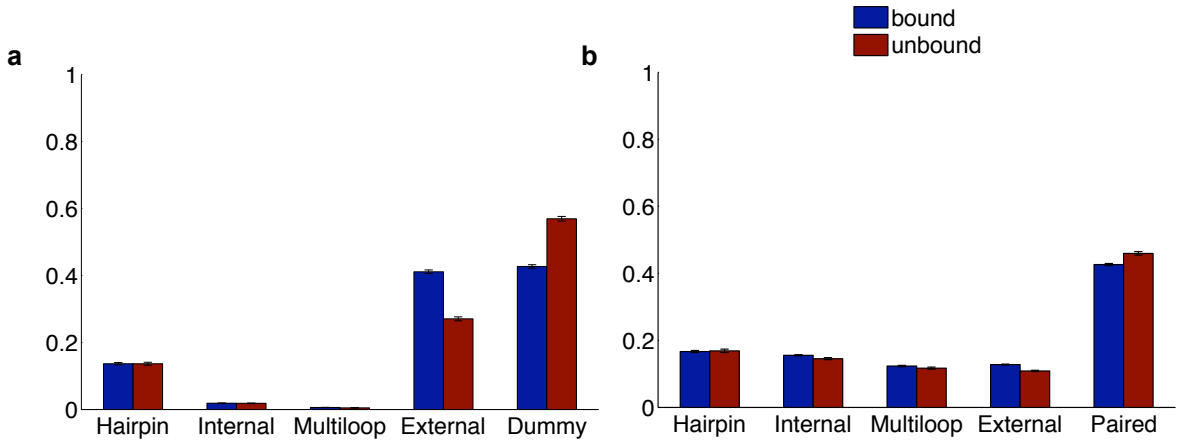


Figure 5.10: *Mean profile values for the top k-mer occurrences in bound and unbound sequences for RBPs a) Khd1 b) TIAL1.* The error bars show the standard error of the mean.

5.3 Discussion

In this chapter, we have introduced MaLaRKey, a new motif finding method that infers detailed binding preferences of RBPs from large-scale binding data. MaLaRKey is flexible enough to handle either binary-valued (i.e. bound and unbound) or real-valued binding data. Also, MaLaRKey can be fit to sequences of arbitrary length such as full-length transcripts or short probe sequences. In addition to sequence features, a rich set of secondary structure features are incorporated to build detailed models of RBP binding. Secondary structure features can be used to infer whether the RBP of interest prefers the entire binding site to be in the same structural context, or whether there is an overall preference for a structural context. For the latter case, additional structural features can be added to check whether there is a transition from unpaired to paired state, or vice versa.

To evaluate MaLaRKey, we first compared it against RNAcontext on the binding data of nine RBPs from RNAcompete-v1. The performances of RNAcontext and MaLaRKey are quite similar except for a few RBPs. Since RNAcontext's motif scoring (i.e. noisy-or function [60]) is much more suitable for short sequences, we do not expect MaLaRKey to outperform RNAcontext on this particular data set. However, we show that the additional features considered by MaLaRKey are useful. By fitting the sequence features, single nucleotide structural context feature and 5' transition to unpairedness feature, MaLaRKey recovers the preference of Vts1p for CNGG within a hairpin loop and with G as the 5' closing base of the loop.

We then evaluated MaLaRKey with 14 RIP-chip data sets that provide binding information on full-length yeast transcripts. Predicting binding of full-length transcripts is challenging both because of increased uncertainty of the location of the binding site(s) within the long sequence and also because of decreased accuracy of secondary structure prediction methods for long sequences. Despite these challenges, MaLaRKey is able to learn motif models that accurately predict binding for held-out transcripts. Sequence

preferences inferred by MaLaRKey agree with known preferences (when applicable) or with previous analysis of the same data. Incorporation of secondary structure features improves the prediction accuracy significantly for 9 out of 14 RBPs. For a set of RBPs for which the accuracy does not improve, we showed that this is due to lack of signal in data rather than MaLaRKey's inability to learn the correct structure preference. We also showed that the Vts1p's preferences inferred from RNAcompete-v1 data and RIP-chip data are consistent with each other.

Lastly, we ran MaLaRKey on CLIP data for seven RBPs. Learned sequence motifs for these RBPs are in agreement with previous analysis of these data, and predict binding of held-out data accurately. Including structural information improves the prediction accuracy significantly for two RBPs (i.e. IGFBP1-3 and Khd1). The lack of performance gain with the addition of structural features for the remaining five human RBPs could be due to one (or more) of the following reasons: (i) MaLaRKey is not able to learn the correct structural preferences from CLIP data; (ii) the RBP does not have a structural preference; and (iii) RNA secondary structure cannot be predicted accurately for human sequences. We argue that (i) cannot be the case since the binding models that are inferred from RIP-chip and CLIP data for Khd1 are highly similar, and the addition of structure improves the accuracy substantially. Since the five RBPs that are tested have unknown structure preferences (ii) could be the reason. The base content of the sequence motif is likely to also play a role in the ability to detect structural preferences. If the binding site is enriched for G's and C's, and the RBP prefers single-stranded regions, it is easy to detect this preference as GC-rich regions are likely to be paired. On the other hand, detecting a preference to single-stranded regions that are AU-rich can be difficult as only a small number of the occurrences are not in the preferred context (i.e. paired). Inaccuracies in predicting secondary structure may also play a role since the fact that the transcripts are occupied by many factors during secondary structure formation is not taken into account by computational prediction methods.

Although added predictive value of structural features is little or none for some RBPs, construction of detailed models of binding is still critical for identifying the true binding sites of an RBP. For example, a genome-wide scan of Vts1p binding sites would identify all occurrences of CNGG's when only its sequence preferences are known. On the other hand, once the fact that Vts1p binds to CNGGs within a hairpin loop is learned, the false positive matches where CNGG is not in a hairpin loop can be discarded. Furthermore, if Vts1p is known to prefer a G 5' to the hairpin loop, more confidence can be assigned to those sites that contain a G 5' to the hairpin loop.

Chapter 6

Conclusions and Future Work

Information about RBP-RNA interactions is fundamental for the understanding of post-transcriptional networks and its associated diseases. *In vitro* and *in vivo* studies, using affinity selection-based approaches, have successfully identified RBP targets. However, inferring RBP binding preferences from these data is challenging. The goal of this thesis is to develop models to identify RBP binding preferences from experimental data. We address two separate but related problems. First, we describe a new strategy to design a complex pool for *in vitro* assays to query RBP sequence and structure preferences. Second, we propose motif models to learn RBP sequence and structure preferences from experimental data.

In Chapter 3, we described a protocol to design a complex RNA pool that supports querying RBP binding preferences *in vitro*. A common strategy to design an RNA pool for *in vitro* assays is to randomly generate RNA sequences. Although simple, this approach fails to represent the maximum possible diversity of sequence elements, and ignores RNA secondary structure. We propose a new strategy for designing a pool that represents the entire space of possible k-mers as well as selected elements of RNA secondary structure. Moreover, our design fits into the limited space of the array, and minimizes microarray cross-hybridization. We utilized de Bruijn sequences to represent all possible k-mers

using minimum space, and we used existing RNA secondary structure prediction tools to design RNA sequences that fold into specific secondary structures. Our design has been implemented in the RNAcompete method, and applied to several RBPs across different organisms. To evaluate RNAcompete, we took advantage of the fact that our pool contains two distinct replicate designs. We observed strong correlations between relative binding affinities of 7-mers that are estimated from the two replicate designs. Furthermore, we confirmed that the identified motifs are consistent with established binding preferences when available. In conclusion, our design played a key role in the success of RNAcompete, an *in vitro* tool for comprehensive analysis of RNA binding specificities.

In Chapter 4, we developed RNAcontext, a discriminative probabilistic model to learn sequence and structure binding preferences of RBPs from large-scale binding data. Existing models to identify RBP binding preferences fail to take into account RNA secondary structure, or use limited representations of it. RNAcontext augments the standard PWM representation of RBP sequence binding preferences with a representation of its preferred structural context(s). RNAcontext represents the structural context of a base using a probability distribution over a set of non-overlapping annotated states (i.e. paired, hairpin loop etc.). The probability distribution over these states is computed using Sfold, which samples a large set of structures from the ensemble of all possible structures. The level of specificity of structural contexts can be easily adjusted by changing the annotation alphabet. We evaluated RNAcontext by comparing it against two existing motif models: MEMERIS and MatrixREDUCE. RNAcontext achieved higher average AU-PR values than MEMERIS and MatrixREDUCE on all of the nine RBPs analyzed. We also confirmed that the added predictive value for RNAcontext was due to the incorporation of structural context. Lastly, we showed that RNAcontext predicted motif models reflect previously reported sequence and structure preferences for well-studied RBPs like HuR, Vts1p and PTB. In summary, we demonstrated that RNAcontext represents an advance

over existing methods for modeling RBP binding preferences. However, RNAcontext's motif scoring scheme and RNA secondary structure prediction procedure is restricted to short input sequences, and its structure model is limited to query overall preferences of the binding site.

In Chapter 5, we described MaLaRKey, which extends RNAcontext in several ways. For example, MaLaRKey uses a motif scoring scheme suitable for input sequences of arbitrary length. To predict secondary structures of full-length transcripts, MaLaRKey uses RNAplfold, a local folding algorithm that considers the Boltzmann ensemble of all possible structures. MaLaRKey's structure model allows to infer whether the RBP prefers the entire binding site to be in the same structural context, or whether there is an overall preference for a structural context. For the latter case, additional structural features can be added to check whether there is a transition from unpaired to paired state, or vice versa. Another advantage of MaLaRKey is the use of a parameter fitting procedure that is flexible enough to handle either binary-valued (i.e. bound and unbound) or real-valued binding data. We evaluated MaLaRKey with both *in vitro* and *in vivo* data sets. MaLaRKey performs as good as RNAcontext for many of the RBPs from RNAcompete-v1 data. MaLaRKey is able to learn motif models that accurately predict binding for held-out transcripts of 14 RIP-chip datasets which provide binding information on full-length yeast transcripts. Incorporation of secondary structure features improves the prediction accuracy significantly for 8 out of 14 RBPs. We also ran MaLaRKey on CLIP data for seven RBPs. Learned sequence motifs for these RBPs are in agreement with previous analysis of these data, and predict binding of held-out data accurately. Including structural information improves the prediction accuracy significantly for two RBPs For the RBPs for which the accuracy does not improve, we showed that this is due to lack of signal in data rather than MaLaRKey's inability to learn the correct structure preference. In summary, MaLaRKey has the ability to successfully infer both sequence preferences and complex structure preferences of RBPs from input sequences of arbitrary length.

In conclusion, although RNA secondary structure is known to play an important role in RBP-RNA interactions, it is ignored in most experimental designs and subsequent analysis of the binding data. The methods developed in this thesis support the investigation of the effect of RNA secondary structure in RBP binding, and constructing detailed models of RBP preferences. An important motivation for characterizing RBP preferences is to be able to predict the role of disease-associated mutations. This could be made possible within the RNAcompete framework by designing and testing a pool containing the various variants of the wild-type binding site. Likewise, once the motif models are learned, RNAcontext and MaLaRKey can be used to scan new sequences containing variants of the wild-type binding site and predict whether these will be bound by the RBP of interest. As transcripts are occupied by many factors, rather than scanning for binding sites of individual RBPs independently, combined knowledge of RBP-RNA interactions needs to be taken into account. As such, an important goal of the research in this area is to increase the coverage of RBPs with characterized binding preferences.

6.1 Future directions

A number of challenges remain in modeling RNA-protein interactions.

An important fact ignored by both RNAcontext and MaLaRKey is that many RBPs have multiple RNA-binding domains that each binds RNA. For example, PTB has 4 RRMs domains that each bind short pyrimidine sequences with variable spacing between the binding sites of individual domains [120]. If each individual domain has the same binding specificities, summing the k-mer scores to obtain the score for the entire sequence (i.e. as in MaLaRKey model) can provide a crude approximation. However, when the binding specificities of individual domains are distinct and require variable spacing in between, more complex models are needed. A similar problem in DNA motif finding has been addressed by the use of hidden Markov models (HMM) or conditional random fields

(CRF), and similar approaches can be used for RNA motif models [134, 135].

The feature set used by MaLaRKey can be extended to allow for more detailed representations of sequence and structure preferences. The sequence model of MaLaRKey assumes that the individual positions within the site make independent contributions to binding. However, there exist examples of RBP binding preferences where two or more positions are dependent on each other. MaLaRKey could be easily extended to include additional features that span multiple positions, and represent the dependencies in between. Similarly, the set of structural features considered by MaLaRKey can be expanded to represent more complex structure elements. Our procedure for calculating the probabilities of structure patterns for transition features could be extended to include arbitrary structure patterns such as PPIIPP where P and I correspond to paired and internal loop contexts, respectively. In general, including additional features would be useful to construct more detailed binding models, but may also lead to overfitting. This problem can be solved by switching to a Bayesian setting where we can automatically determine the relevant features by introducing priors for parameters and performing Bayesian inference.

The performance of the methods introduced in this thesis heavily depends on the accuracy of existing RNA secondary structure prediction tools. We used a small subset of the available tools (i.e. RNAshapes, Sfold, RNAplfold) in this thesis. An immediate future step would be to try other secondary structure prediction tools such as probabilistic models, and those that incorporate probing data or that consider pseudoknots to assess whether their predictions result in better models of RBP binding. Also, existing secondary structure prediction tools can be improved by considering co-transcriptional folding (i.e. 5' of the transcript is available to fold first).

Finally, an important major goal would be to combine existing knowledge on miRNA binding sites and RBP binding sites to explain several post-transcriptional phenomena such as RNA localization and stability. Such a model needs to incorporate competi-

tive interactions among RBPs and miRNAs as several studies demonstrate the interplay between these two classes of regulators[136].

Bibliography

[1] SP Gygi, Y Rochon, BR Franca, and R Aebersold. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, 19(3):1720–1730, 1999.

[2] MP Washburn, A Koller, G Oshiro, RR Ulaszek, D Plouffe, C Deciu, E Winzeler, and JR Yates. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA*, 100(6):31073112, 2003.

[3] CW Smith and J Valcarcel. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci*, 8:381–388, 2000.

[4] JD Keene and SA Tenenbaum. Eukaryotic mRNPs may represent posttranscriptional operons. *Mol Cell*, 9:1161–1167, 2002.

[5] JD Keene. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet*, 8:533–543, 2007.

[6] DJ Hogan, DP Riordan, AP Gerber, D Herschlag, and PO Brown. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol*, 6(10):e255, 2008.

[7] N Mukharjee, PJ Larger, MB Friedersforf, MA Thompson, and JD Keene. Coordinated posttranscriptional mRNA population dynamics during T-cell activation. *Mol Syst Biol*, 5(288):288–302, 2009.

[8] Y Barash, JA Calarco, W Gao, Q Pan, X Wang, O Shai, BJ Blencowe, and BJ Frey. Deciphering the splicing code. *Nature*, 465:53–59, 2010.

[9] KE Lukong, K Chang, EW Khandjian, and S Richard. RNA-binding proteins in human genetic disease. *Trends Genet*, 24(8):416–425, 2008.

[10] BA Oostra and R Willemsen. FMR1: a gene with three faces. *Biochim Biophys Acta*, 1790:467–477, 2009.

[11] TA Cooper, L Wan, and G Dreyfuss. RNA and disease. *Cell*, 136:777–793, 2009.

[12] L Cartegni and AR Krainer. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet*, 30(4):377–384, 2002.

[13] G Dreyfuss, VN Kim, and N Kataoka. Messenger-RNA-binding proteins and the messages that they carry. *Nat Rev Mol Cell Biol*, 3:195–205, 2002.

[14] N Mittal, N Roy, MM Babua, and SC Jangaa. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proc Nat Sci Acad USA*, 106(48):20300–20305, 2009.

[15] S Kishore, S Luber, and M Zavolan. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Brief Funct Genomic*, 9(5), December 2010.

[16] DE Draper. Themes in RNA -protein recognition. *J Mol Biol*, 293(2):255–70, 1999.

[17] KM Weeks and DM Crothers. RNA recognition by Tat-derived peptides: interaction in the major groove . *Cell*, 66(3):577–588, 1991.

[18] D Zhu, CR Stumpf, JM Krahn, M Wickens, and TMT Hall. A 5' cytosine binding pocket in Puf3p specifies regulation of mitochondrial mRNAs. *Proc Nat Acad Sci USA*, 106(48):20192–20197, 2009.

[19] PE Johnson and LW Donaldson. RNA recognition by the Vts1p SAM domain. *Nat Struct Mol Biol*, 13:177–178, 2006.

[20] A Ramos, S Grunert, J Adams, DR Micklem, MR Proctor, S Freund, M Bycroft, DS Johnston, and G Varani. RNA recognition by a Staufen double-stranded RNA-binding domain. *Embo J*, 19(5):997–1009, 2000.

[21] JW Szostak AD Ellington. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346:818–822, 1990.

[22] A Jolma, T Kivioja, J Toivonen, L Cheng, G Wei, M Enge, M Taipale, JM Vaquerizas, J Yan, MJ Sillanpaa, M Bonke, K Palin, S Talukder, TR Hughes, NM Luscombe, E Ukkonen, and J Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, 20(6):861–873, 2010.

[23] D Ray, H Kazan, ET Chan, LP Castillo, S Chaudhry, S Talukder, BJ Blencowe, Q Morris, and TR Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*, 27(7):667–670, 2009.

[24] JD Keene, JM Komisarow, and MB Friedersdorf. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature Prot*, 1(1):302–307, 06 2006.

[25] J Zhao, TK Ohsumi, JT Kung, Y Ogawa, DJ Grau, K Sarma, JJ Song, RE Kingston, M Borowsky, and JT Lee. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol cell*, 40(6):939–953, 12 2010.

[26] S Mili and JA Steitz. Evidence for reassociation of RNA-binding proteins after cell lysis : implications for the interpretation of immunoprecipitation analyses. *RNA*, 10(11):1692–1694, 2004.

[27] J Ule, K Jensen, A Mele, and RB Darnell. CLIP:A method for identifying protein-RNA interaction sites in living cells. *Methods*, 37:376–386, 2005.

[28] M Hafner, M Landthaler, L Burger, M Khorshid, J Hausser, P Berninger, A Rothballer, M Ascano, AC Jungkamp, M Munschauer, A Ulrich, GS Wardle, S Dewell, M Zavolan, and T Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 04 2010.

[29] J Konig, Z Kathi, R Gregor, T Curk, M Kayikci, B Zupan, DJ Turner, NM Luscombe, and J Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17:909–915, 2010.

[30] LA Newberg, WA Thompson, S Conlan, TM Smith, LA McCue, and CE Lawrence. A phylogenetic gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics*, 23(14):1718–1727, 2007.

[31] CE Lawrence LE Carvalho. Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc Natl Acad Sci U S A*, 105(9):3209–3214, 2008.

[32] MF Berger and ML Bulyk. Protein binding microarrays (PBMs) for the rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol*, 338:245–260, 2006.

[33] SJ Maerkl and SR Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.

[34] EM Conlon, XS Liu, JD Lieb, and JS Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Genetics*, 100(6):3339–3344, 2003.

[35] HJ Bussemaker, H Li, and ED Siggia. Regulatory element detection using correlation with expression. *Nature Genet*, 27:167–171, 2001.

[36] BC Foat, AV Morozov, and HJ Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22(14):e141–e149, 2006.

[37] X Chen, TR Hughes, and Q Morris. RankMotif++: a motif search algorithm that accounts for relative ranks of k-mers in binding transcription factors. *Bionformatics*, 23:72–79, 2006.

[38] A Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, 16:962–972, 2006.

[39] E Eden, D Lipson, S Yagev, and Z Yakhini. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol*, 3:e39, 2007.

[40] TL Bailey, N Williams, C Misleh, and WW Li. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34:W369–W373, 2006.

[41] P Benos, ML Bulyk, and GD Stormo. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, 30(20):4442–4451, 2002.

[42] ML Bulyk, PLF Johnson, and GM Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–1261, 2002.

[43] Y Barash, G Elidan, N Friedman, and T Kaplan. Modeling dependencies in protein-DNA binding sites. *In Proc RECOMB*, pages 28–37, 2003.

[44] FP Roth OD King. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res*, 31(19):e116, 2003.

[45] E Segal E Sharon, S Lubliner. A feature-based approach to modeling protein-DNA interactions. *PLoS Comp Biol*, 4(8):E1000154, 2008.

[46] WN Grundy, TL Bailey, CP Elkan, and ME Baker. Meta-MEME: Motif-based hidden Markov models of biological sequences. *Comput App Biosciences*, 3(4):397–406, 1997.

[47] SR Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[48] X Liu, DL Brutlag, and JS Liu. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, pages 127–138, 2001.

[49] M Djordjevic, AM Sengupta, and BI Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13(11):2381–2390, 2003.

[50] OG Berg and PH von Hippel. Selection of DNA binding sites by regulatory proteins.statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193:723–750, 1987.

[51] GD Stormo. DNA binding sites: Representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[52] BC Foat, SS Houshmandi, WM Olivas, and HJ Bussemaker. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci U S A*, 102:17675–17680, 2005.

[53] JS Liu XS Liu, DL Brutlag. An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat Biotechnol*, 20(8):835–839, 2002.

[54] AP Dempster, NM Laird, and DB Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *J Royal Stat S Ser B*, 39(1):1–38, 1977.

[55] MH Tang, A Krogh, and O Winther. BayesMD: Flexible biological modeling for motif discovery. *J Comput Biol*, 15(10):1347–1363, 2008.

[56] L Narlikar, R Gordan, and AJ Hartemink. Nucleosome occupancy information improves de novo motif discovery. *In Proc RECOMB*, pages 107–121, 2007.

[57] M Hiller, R Pudimat, A Busch, and R Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res*, 34(17):e117, 2006.

[58] HG Roider, A Kanhere, T Manke, and M Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, 2007.

[59] E Segal, Y Barash, I Simon, N Friedman, and D Koller. From promoter sequence to expression: a probabilistic framework. *In Proc RECOMB*, pages 263–272, 2002.

[60] JA Granek and ND Clarke. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biology*, 6:R87, 2005.

[61] E Segal, T Raveh-Sadka, M Schroeder, U Unnerstall, and U Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451:535–540, 2008.

[62] M Kertesz, Y Wan, E Mazor, JL Rinn, RC Nutter, HY Chang, and E Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 7311:103–107, 2010.

[63] JG Underwood, AV Uzilov, S Katzman, CS Onodera, JE Mainzer, DH Mathews, TM Lowe, SR Salama, and D Haussler. FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Meth*, 7(12):995–1001, 12 2010.

[64] JB Lucks, SA Mortimer, C Trapnell, S Luo, S Aviran, GP Schroth, L Pachter, JA Doudna, and AP Arkin. Multiplexed RNA structure characterization with

selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci USA*, 108(27):11063–11068, 2011.

[65] I Tinoco and C Bustamante. How RNA folds. *Mol Biol*, 293:271–281, 1999.

[66] PP Gardner and R Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5:140–158, 2004.

[67] I Hofacker, M Fekete, and PStadler. Secondary structure prediction for aligned RNA sequences. *J of Mol Biol*, 319(5):1059–1066, 2002.

[68] B Knudsen and J Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, 2003.

[69] D Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J App Math*, 45:810–825, 1985.

[70] J Gorodkin, SL Stricklin, and GD Stormo. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res*, 29(10):2135–2144, 2001.

[71] DH Mathews and DH Turner. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *J of Mol Biol*, 317(2):191–203, 2002.

[72] IL Hofacker, S Bernhart, and P Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20:2222–2227, 2004.

[73] CB Do, C Foo, and S Batzoglou. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, 24(13):i68–i76, 2008.

[74] DH Mathews, J Sabina, M Zucker, and H Turner. Expanded sequence dependence on thermodynamix parameters provides robust prediction of RNA secondary structure. *Jour of Mol Biol*, 288(5):911–940, 1999.

[75] DH Mathews and DH Turner. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol*, 31(13):270–278, 2006.

[76] P Steffen, B Vo, M Rehmsmeier, J Reeder, and R Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

[77] Y Ding and CE Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*, 31:7280–7301, 2003.

[78] M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9:133–148, 1981.

[79] JS McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[80] IL Hofacker, W Fontana, PF Stadler, S Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Chemie*, 125:167–188, 1994.

[81] SH Bernhart, IL Hofacker, and PF Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics)*, 22(5):614–615, March 2006.

[82] SJ Lange, D Maticzka, M Mohl, JN Gagnon, CM Brown, and R Backofen. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res*, 2012.

[83] C Flamm, W Fontana, IL Hofacker, and P Schuster. RNA folding at elementary step resolution. *RNA*, 6(3):325–338, 2000.

[84] EM Mahen, PY Watson, JW Cottrell, and MJ Fedor. mRNA secondary structures fold sequentially but exchange rapidly in vivo. *PLoS Biol*, 8(2):e1000307, January 2010.

[85] CB Do, DA Woods, and S Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based methods. *Bioinformatics*, 22:e90–e98, 2006.

[86] S Zakov, Y Goldberg, M Elhadad, and M Ziv-Ukelson. Rich parameterization improves RNA structure prediction. *In Proc RECOMB*, 6577:546–562, 2011.

[87] M Andronescu, A Condon, HH Hoos, DH Mathews, and KP Murphy. Computational approaches for RNA energy parameter estimation. *RNA*, 16:2304–2318, 2010.

[88] DH Mathews, MD Disney, JL Childs, SJ Schroeder, M Zuker, and DH Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA*, 101:7287–7292, 2004.

[89] KE Deigan, TW Li, DH Mathews, and KM Weeks. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci USA*, 106:97–102, 2009.

[90] S Washietl, IL Hofacker, PF Stadler, and M Kellis. RNA folding with soft constraints: Reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res*, 2012.

[91] J Hackermuller, NC Meisner, M Auer, M Jaritz, and PF Stadler. The effect of RNA secondary structure on RNA-ligand binding and the modifier RNA mechanism: a quantitative model. *Gene*, 345(1):3–12, 2005.

[92] NC Meisner, J Hackermuller, V Uhl, A Aszodi, M Jaritz, and M Auer. mRNA openers and closers: Modulating AU-rich element controlled mRNA stability by a molecular switch in mRNA secondary structure. *Chembiochem*, 5:1432–1447, 2004.

[93] SR Eddy and R Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Res*, 22:2079–2088, 1994.

[94] Z Yao, Z Weinberg, and WL Ruzzo. CMfinder: A covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–52, 2006.

[95] Z Yao, Z Weinberg, and WL Ruzzo. CMfinder 1.0 Manual: http://wingless.cs.washington.edu/CMfinder/manual.htm.

[96] M Rabani, M Kersetz, and E Segal. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci U S A*, 105(39):14885–14890, 2008.

[97] MT Miller, JJ Higgin, and TM Hall. Basis of altered RNA-binding specificity by PUF proteins revealed by crystal structures of yeast Puf4p. *Nat Struct Mol Biol*, 15(4):397–402, 2008.

[98] R Stefl, M Xu, L Skrisovska, RB Emeson, and FHT Allain. Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs . *Structure*, 14:345–355, 2008.

[99] SR Eddy. User's guide for COVE - covariance models of RNA sequence families. *MRC Lab of Mol Biol, Cambridge, UK*, [Guide tex in cove-2.4.4.tar.Z]., 1993.

[100] BC Foat and GD Stormo. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. *Mol Syst Biol*, 5(268):268, January 2009.

[101] AA Philippakis, MF Berger AM Qureshi, and ML Bulyk. Design of compact, universal DNA microarrays for protein binding microarray experiments. *J Comput Biol*, 15(7):655–665, 2008.

[102] JL Gross and J Yellen. *Handbook of Graph Theory*. CRC Press, New York, 2004.

[103] A Golomb. *Shift Register Sequences*. Aegean Park Press, Laguna Hills, CA, 1967.

[104] WA Huber, HS Heydebreck, A Poustka, and M Vingron. Variance stabilization applied to microarray data calibration and the quantification of differential expression. *Bioinformatics*, 18:96–104, 2002.

[105] DE Tsai, DS Harper, and JD Keene. U1-snRNP-A protein selects a ten nucleotide consensus sequence from a degenerate RNA pool presented in various structural contexts. *Nucleic Acids Res*, 19:4931–4936, 1991.

[106] R Tacke and JL Manley. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA-binding specificities. *EMBO J*, 14:3540–3551, 1995.

[107] I Perez, JG McAfee, and JG Patton. Multiple RRMs contribute to RNA binding specificity and affinity for polypyrimidine tract binding protein. *Biochemistry*, 36:11881–11890, 1997.

[108] FB Gao, CC Carson, T Levine, and JD Keene. Selection of a subset of mRNAs from combinatorial 3' untranslated region libraries using neuronal RNA-binding protein Hel-N1. *Proc Natl Acad Sci U S A*, 91:11207–11211, 1994.

[109] T Aviv, Z Lin, S Lau, LM Rendl, F Sicheri, and CA Smibert. The RNA-binding SAM domain Smaug defines a new family of post-transcriptional regulators. *Nat Struct Mol Biol*, 10:614–621, 2003.

[110] C Shin and JL Manley. The SR protein SRp38 represses splicing in M phase cells. *Cell*, 111:407–417, 2002.

[111] T Aviv, Z Lin, G Ben-Ari, CA Smibert, and F Sicheri. Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1. *Nat Struct Mol Biol*, 13:168–176, 2006.

[112] S Griffiths-Jones, S Moxon, M Marshall, A Khanna, SR Eddy, and A Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33 (Database issue):D121–D124; doi:10.1093/nar/gki081, 2005.

[113] I López de Silanes, M Zhan, A Lal, X Yang, and M Gorospe. Identification of a target RNA motif for RNA-binding protein HuR. *Proc Nat Acad Sci USA*, 101(9):2987–92, 2004.

[114] AP Gerber, D Herschlag, and PO Brown. Extensive association of functionally and cytotopically related mRNAs with PuF family RNA-binding proteins in yeast. *PLoS Biol*, 2(3):e79, 2004.

[115] S Sinha. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, 22(14):e454–e463, 2006.

[116] MJ Law, AJ Rice, P Lin, and IA Laird-Offringa. The role of RNA structure in the interaction of U1A protein with U1 hairpin II RNA . *RNA*, 12:1168–1178, 2006.

[117] L Wickham, T Duchaine, M Luo, IR Nabi, and L DesGroseillers. Mammalian Staufen is a double-stranded-RNA- and tubulin-binding protein which localizes to the rough endoplasmic reticulum. *Mol Cell Biol*, 19:2220–2230, 1999.

[118] RH Byrd, P Lu, and J Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM J Sci and Stat Comp*, 16:1190–1208, 1995.

[119] TM Bailey and M Gribskov. Combining evidence using p-values: applications to sequence homology searches. *Bioinformatics*, 14(1):48–54, 1998.

[120] FC Oberstrass, SD Auweter, M Erat, Y Hargous, A Henning, P Wenter, L Reymond, B Amir-Ahmady, S Pitsch, DL Black, and FHT Allain. Structure of PTB bound to RNA: Specific binding and implications for splicing regulation. *Science*, 309:2054–2057, 2005.

[121] CT Workman, Y Yin, DL Corcoran, T Ideker, GD Stormo, and PV Benos. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res*, 33:W389–W392, 2005.

[122] R Tacke and JL Manley. The human splicing factors ASF/SF2 and SC35 possess different, functionally significant RNA binding specificities. *EMBO J*, 14:3540, 1995.

[123] JR Sanford, X Wang, M Mort, N VanDuyn, DN Cooper, and SD Mooney. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res*, 19:381–394, 2008.

[124] G Badis, MF Berger, AA Philippakis, S Talukder, AR Gehrke, SA Jaeger, ET Chan, G Metzler, A Vedenko, X Chen, H Kuznetsov, C Wang, D Coburn, DE Newburger, Q Morris TR Hughes, and ML Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324:1720–1723, 2009.

[125] X Li, G Quon, H Lipshitz, and Q Morris. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, doi:10.1261/rna.2017210, 2010.

[126] T Pan and T Sosnick. RNA folding during transcription. *Ann Rev of Biophys and Biomol Struct*, 35(1):161–175, 2006.

[127] R Tibshirani J Friedman, T Hastie. Regularization paths for generalized linear models via coordinate descent. *J Stats Soft*, 33(1):1–22, 2010.

[128] PO Brown DP Riordan, D Herschlag. Identification of RNA recognition elements in the Saccharomyces cerevisiae transcriptome. *Nucleic Acids Res*, pages 1–9, 2010.

[129] G Anders, SD Mackowiak, M Jend, J Maaskola, A Kuntzagk, N Rajewsky,

M Landthaler, and C Dieterich. doRINA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res*, 40(D1):180–186, 2011.

[130] JJ Wolf, RD Dowell, S Mahony, M Rabani, DK Gifford, and GR Fink. Feed-forward regulation of a cell fate determinant by an RNA-binding protein generates asymmetry in yeast. *Genetics*, 185(2):513–522, 2010.

[131] IM Hall AR Quinlan. BEDtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[132] Z Wang, M Kayikci, M Briese, K Zarnack, and NM Luscombe. iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol*, 8(10):e1000530, 2010.

[133] JR Tollervey, T Curk, B Rogelj, M Briese, M Cereda, M Kayikci, J Konig, T Hortobagyi, AL Nishimura, V Zupunski, R Patani, S Chandran, G Rot, B Zupan, CE Shaw, and J Ule. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat Neurosci*, 14(4):452–458, 04 2011.

[134] MC Frith, SFW Saunders, B Kobe, and TL Bailey. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol*, 4(5):e1000071, 2008.

[135] W Fu, P Ray, and EP Xing. DISCOVER: A feature-based discriminative method for motif search in complex genomes. *Bioinformatics*, 25(12):i321–i329, 2009.

[136] M Van Kouwenhove, M Kedde, and R Agami. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nature Reviews Cancer*, 11(9):644–656, 2011.