Identification of functional RNA structures in sequence data

Author: Shermin Pei

BOSTON COLLEGE

Identification of functional RNA structures in sequence data

by

Shermin Pei

A Dissertation
submitted to the Faculty of
the department of Biology
in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

in the
BOSTON COLLEGE
Morrissey College of Arts and Sciences
Department of Biology
Meyer Lab

December 2016

Identification of functional RNA structures in sequence data

Shermin Pei

Abstract

Identification of functional RNA structures in sequence data

by Shermin Pei

Advisor: Dr. Michelle Meyer

Structured RNAs have many biological functions ranging from catalysis of chemical reactions to gene regulation. Many of these homologous structured RNAs display most of their conservation at the secondary or tertiary structure level. As a result, strategies for natural structured RNA discovery rely heavily on identification of sequences sharing a common stable secondary structure. However, correctly identifying the functional elements of the structure continues to be challenging. In addition to studying natural RNAs, we improve our ability to distinguish functional elements by studying sequences derived from *in vitro* selection experiments to select structured RNAs that bind specific proteins. In this thesis, we seek to improve methods for distinguishing functional RNA structures from arbitrarily predicted structures in sequencing data. To do so, we developed novel algorithms that prioritize the structural properties of the RNA that are under selection. In order to identify natural structured ncRNAs, we bring concepts from evolutionary biology to bear on the *de novo* RNA discovery process. Since there is selective pressure to maintain the structure, we apply molecular evolution concepts such as neutrality to identify functional RNA structures. We hypothesize that alignments corresponding to structured RNAs should consist of neutral sequences. During the course of this work, we developed a novel measure of neutrality, the structure ensemble neutrality (SEN), which calculates neutrality by averaging the magnitude of structure retained over all single point mutations to a given sequence. In order to analyze *in vitro* selection data for RNA-protein binding motifs, we developed a novel framework that identifies enriched substructures in the sequence pool. Our method accounts for both sequence and structure components by abstracting the overall secondary structure into smaller substructures composed of a single base-pair stack. Unlike many current tools, our algorithm is designed to deal with the large data sets coming from high-throughput sequencing. In conclusion, our algorithms have similar performance to existing programs. However, unlike previous methods, our algorithms are designed to leverage the evolutionary selective pressures in order to emphasize functional structure conservation.

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Michelle Meyer, for her constant support, encouragement, and mentorship throughout graduate school. She has helped me grow as a scientist and a writer over the years. I could not have asked for a better advisor.

Second, I'd like to thank members of my thesis committee (Dr. Peter Clote, Dr. Charles Hoffman, Dr. Welkin Johnson, and Dr. Brian Tjaden) for their insights and suggestions throughout my graduate school career.

Additionally, I would like to thank my family and friends for their support and encouragement. In particular, thank you to members of the Meyer lab: Arianne Babina for her tough love and support, Betty Slinger all the great discussions we've had over beer, and Jon Anthony for teaching me to code and the joy of Clojure.

Contents

List of Figures

List of Tables

Abbreviations

Bp-distance	Normalized base-pair distance
CFG	Context free grammar
CM	Covariant model
DBP	DNA-binding protein
db	Dot-bracket
HTS	High-throughput sequencing
JSD	Jensen-Shannon divergence
KLD	Kullback-Leibler divergence
MFE	Minimum free energy
miRNA	Micro RNA
mpwi	Mean pairwise identity
MSA	Multiple sequence alignment
NCM	Nucleotide cyclic motif
ncRNA	Non-coding RNA
nt	Nucleotide
PCC	Pearson's correlation coefficient
RBP	RNA-binding protein
RFam	RNA families database
ROC	Receiver operating characteristic
rRNA	ribosomal RNA
SCFG	Stochastic context free grammar
SELEX	Systematic evolution of ligands by exponential enrichment
SEN	Structure ensemble neutrality

sRNA	small RNA
SVM	Support vector machine
tRNA	transfer RNA

Chapter 1

Introduction

RNA has long been the forgotten middle child of the central dogma of biology — DNA makes RNA makes protein. RNA for a time was relegated to only messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), which are all exclusively involved in the protein biosynthesis process. However, in the past two decades, there has been a resurgence in the interest in RNA as a gene regulatory mechanism. Non-coding RNAs (ncRNA) do not encode a protein product. Instead, they have roles in both bacterial and eukaryotic gene regulation, localization, translational efficiency and metabolite sensing [1–5]. Changes to RNA structure have been implicated as causes for human genetic diseases [6]. In terms of regulatory mechanisms, the ncRNAs can be broadly separated into two categories: *cis* and *trans*-acting. *Cis*-acting RNAs work by forming a secondary structure that transriptionally or translationally regulates nearby genes, usually on the same transcript. For example, in *Bacillus subtillis*, the guanine riboswitch will repress expression of the xpt-pbuX operon in the presence of guanine via the formation of a rho-independent intrinsic terminator stem [7]. On the other hand, *trans*-acting RNAs such as small RNAs (sRNA) [8] and micro RNA (miRNA) [9] are encoded on separate transcripts and use sequence specificity to regulate gene expression.

In bacteria, *cis*-acting structured ncRNAs regulate gene expression via ligand binding. While gene regulation is essential, not all bacteria use the same regulatory mechanisms. These *cis*-regulatory structures have widely variable distributions across the

bacterial phylogenetic tree. Interestingly, some such RNA structures are widely distributed throughout the phylogenetic tree to many bacterial phyla but more often they are narrowly distributed and specific to a single phyla [10]. Particularly interesting is that highly conserved genes that are shared by many phyla can be regulated by distinct RNA structures in different species. The narrow distribution and distinct structure make identification and discovery a significant problem.

1.1 RNA structure discovery

Unlike protein sequences, which are readily identified in genomic sequences, RNAs with homologous functions may be difficult to identify in genomic sequences due to a lack of well defined start and stop signals and poor primary sequence identity [11, 12]. Rather, the biological function of structured RNAs often depends on a well-defined three-dimensional shape that is largely determined by interactions between discrete and stable secondary structure elements [13–15]. These structural constraints lead to covarying mutations, a conservation pattern characterized by the maintenance of base-pairing interactions involved in RNA secondary structure [16, 17]. These features are exploited to identify homologous sequences of previously characterized structured RNAs and to discover new putative RNAs [18]. However, this process is often further complicated by the potential for multiple biologically functional conformations [19], and cases where only a portion of a larger RNA structure is required for biological function. For example, RNAse P is a ribozyme involved in the maturation of small noncoding RNAs whose phylogentically conserved core is functional in isolation, although with significantly decreased activity [20, 21]. Despite these challenges, several computational tools have been developed both for RNA homology searching and *de novo* structured RNA identification [18, 22].

Predicting RNA secondary structure is a fundamental component to building multiple sequence alignments of RNA, RNA homology searches, RNA *de novo* discovery, and is the basis for RNA evolutionary models. Folding algorithms fall into two camps: maximization of the structural stability using the Turner thermodynamic energy model based on

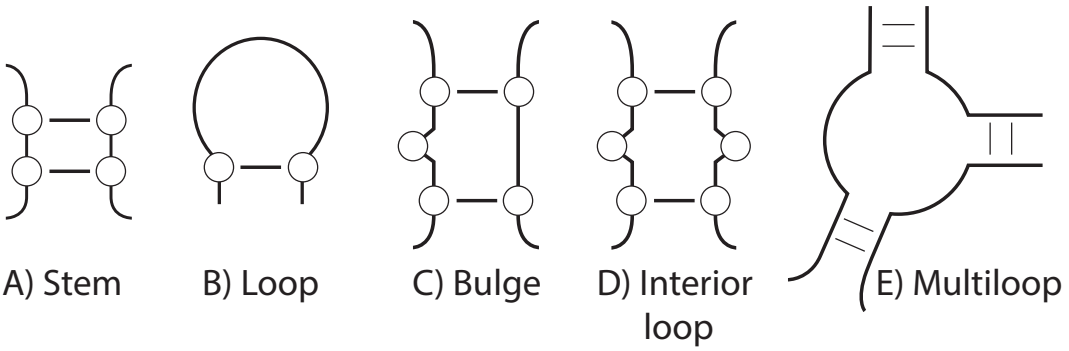

FIGURE 1.1: **RNA secondary structure elements.** RNA secondary structure is classified into five main categories. **A)** Stem is formed from a set of base-paired nucleotides. **B)** Unpaired nucleotides enclosed by a base-pair form a loop. **C)** Bulges refer to unpaired nucleotides in a stem without a corresponding base in the opposite strand. **D)** Interior loops are unpaired nucleotides in a stem with corresponding bases in the opposite strand. **E)** Multiloops are structures composed of multiple stems enclosed by a closing base-pair that is not part of any interior stems.

measurements [23–25], and probabilistic models using parameters derived from statistical learning of known RNAs [26]. The structure predicted for a sequence and subsequent alignment of sequences will vary depending on the type of folding algorithm used. In order to provide some understanding of the structure discovery process, the rest of this section is structured to provide background of the basic ideas in *de novo* identification including: RNA folding, construction of RNA structure alignments, identification of RNA structures using machine learning techniques, and probabilistic models. For the remainder of this thesis, I will refer to "secondary structure" simply as "structure."

1.1.1 Computational prediction

Sequence dictates structure; therefore, the sequence and structure are commonly represented together as a pair. Secondary structure prediction is crucial to RNA discovery because conservation of the same or similar structure in multiple organisms suggests functional importance. An RNA sequence is composed of standard bases $b = \{A, C, G, U\}$. Just like in DNA, RNA nucleotides also base pair with each other following Watson-Crick rules with the addition of the G-U wobble pair. The base-pairing interaction allows RNA to fold into secondary structure, which is composed of elements such as stems, loops, bulges, interior loops and multiloops (Figure 1.1). Stems are formed

when adjacent base-pairs are stacked on top of each other, which can be interrupted by unpaired bases in bulges and interior loops. Loops are at least three unpaired nucleotides closed by a base-pair. Multiloops are formed when multiple distinct stems are enclosed by a closing stem. Let a sequence S with structure T of length L be represented as a set $S = \{s_1, s_2, ..., s_L | s_i \in b\}$ and $T = \{t_1, t_2, ..., t_L | t_i \in \{(, .,)\}\}$. The base-pairs are a set of all (i, j) pairs such that s_i base-pairs with s_j. Here, the structure T is represented in dot-bracket notation, which denotes unpaired positions as "." and paired positions as "(" and ")", where the opening and closing parenthesis are base-paired to each other. For example, given a sequence "GGGGAAAACCCC", the G's would pair with C's and the resulting structure in dot-bracket notation is "((((....))))". Other alternative representations include arcs, trees, and coarse grained. Each of the various representations provides an intuitive visualization for various aspects of the structure: arcs for suboptimal structures, trees for nested structure, and coarse grained for abstracting structures.

To provide some basic insight into how RNA folding algorithms work, I introduce the Nussinov-Jacobson algorithm, which is one of the earliest structure prediction algorithms [27, 28]. Despite the sophistication of modern folding algorithms, the recursions in Nussinov's folding algorithm are still applicable because the computation is fundamentally an optimization problem (equation 1.1). The primary assumption of the Nussinov algorithm is that the most stable structure has the most base-pairs; therefore, Nussinov wanted to identify the structure that maximizes the number of valid base-pairs. The number of base-pairs on the (i, j) interval can be counted using a weight function that assigns 1 to any valid base-pair and 0 otherwise (Equation 1.2). In the base case, we only have to consider the interval containing the smallest physically possible loop such that $j - i > 3$. The recursive function works by reducing the problem to only consider the possible structures within the interval (i, j). This recursion is conceptually important because finding the best structure on the interval (1,L) is algorithmically identical to finding the best structure on the interval $(i, i + 4)$, the smallest possible loop. Within each interval, 4 possible cases are structurally possible: i is unpaired, j is unpaired, i and j base-pair to each other, or the interval contains two stems, referred to as a bifurcation (Figure 1.2).

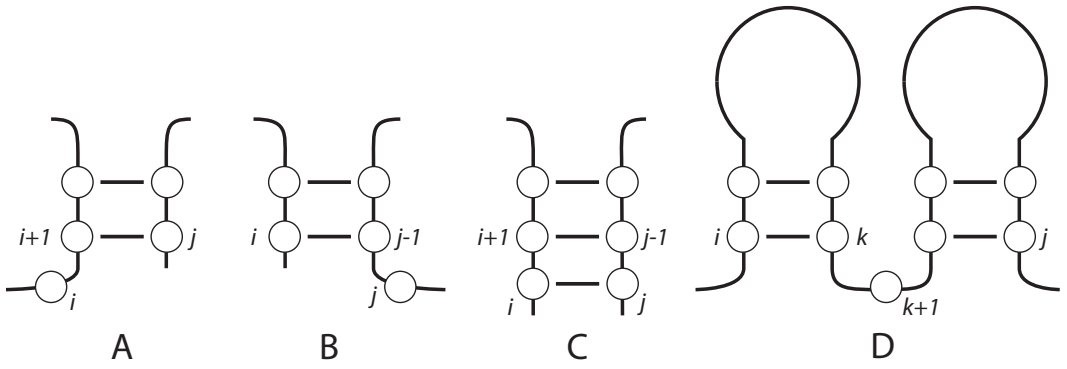

FIGURE 1.2: **Nussinov recursion cases** There are four possible cases to consider in the Nussinov algorithm. **A)** i is unpaired. **B)** j is unpaired. **C)** (i, j) base-pair to each other. **D)** Bifurcation.

$$\gamma(i,j) = max \begin{cases} \gamma(i+1,j) & \text{i is unpaired} \\ \gamma(i,j-1) & \text{j is unpaired} \\ \gamma(i+1,j-1) + w(i,j) & \text{i and j are paired} \\ max_{i<k<j}[\gamma(i,k) + \gamma(k+1,j)] & \text{bifurcation} \end{cases} \tag{1.1}$$

where the matrix is initialized such that $\gamma(i,i) = \gamma(i,i-1) = 0$ and

$$w(i,j) = \begin{cases} 1 & \text{If } (i,j) \text{ base-pair and } j - i > 3 \\ 0 & \text{Otherwise} \end{cases} \tag{1.2}$$

Maximizing base-pairs for the interval $(1, L)$ results in a structure with the most base-pairs. The recursion has implicit base-pairing rules such that given a base-pair (i, j), the following must be true: $i < j$, and there may be no crossing structures such that given two base-pairs $\{(i,j),(k,l)|i < k < j < l\}$. These rules in particular make the folding problem computationally tractable and allow the use of dynamic programming to optimally compute valid structures and find the one with the highest score. However, instead of counting all base-pairs as equal, accuracy may be improved by changing the weight function to add 3 points for a G-C (3 hydrogen bonds), and 2 points for A-U and G-U pairs (2 hydrogen bonds).

Modern structure prediction algorithms improved accuracy by minimizing the Gibb's free

energy (ΔG), rather than counting base-pairs [29]. The structure with the lowest free energy is referred to as the minimum free energy (MFE) structure. The Nussinov algorithm uses a dynamic programming method to optimally solve RNA secondary structure. Zuker and Stiegler improved prediction accuracy by employing a similar dynamic programming method to minimize free energy [30]. The current thermodynamic parameters are derived by the Turner nearest-neighbor energy model, which contains experimentally derived free energies and models the base-pair as a stacking interaction [24]. Minimizing free energy maximizes thermodynamic stability. Since 1986, a major improvement to the model was in determining free energy parameters corresponding to the loop entropies. Additional stability contributions encompassed by the model include adjacent base-pairs, loops, various types of interior loops, multiloops, dangles, and coaxial stacking [25].

While the nearest-neighbor energy model greatly improved prediction accuracy compared to base-pair maximization, there are some limitations to the algorithm. First, programs implementing these algorithms (RNAfold [31], Mfold/UnaFold [32, 33], RNAstructure [34]) cannot predict pseudoknotted structures because base-pairs must be nested and cannot cross [30]. This constraint on crossing structures exists because it turns the structure prediction into a much more complicated problem with a higher computational complexity [35]. These programs return the MFE and corresponding structure, which is not always the biologically relevant structure [36]. Additionally, the accuracy of predicted structure is limited by the accuracy of measured parameters, which in turn is limited by the experimental conditions of the RNAs used to measure thermodynamic parameters [37]. For short RNAs ($<$ 700 nt), the current thermodynamic model correctly predicts approximately 73% of base-pairs [25]. Currently, in order to help conform predicted structure to experimentally derived structures, the thermodynamic energy parameters have been optimized using known RNA structures [38, 39]. However, there is also diminishing returns for each new parameter added [40].

1.1.2 The structure ensemble

Partition function

RNA structures are not static, and they are likely to fluctuate through many different low energy conformations both *in vitro* and *in vivo* [36]. The MFE structure is simply one of many different possible structures, which is not always representative of all possible structure conformations. Therefore, predicting alternative structures provides more insight into the RNA folding landscape. Not every structure is equally likely to appear. The probability of the RNA taking a specific structure can be described using the Boltzmann distribution, which is commonly used in statistical mechanics to describe the probability of a system being in a particular state. Thus by representing each structure as a state, the probability of each structure is proportional to its thermodynamic stability, represented as

$$P_s = \frac{e^{\frac{-\Delta G_s}{RT}}}{Z} \tag{1.3}$$

The denominator (Z) is known as the partition function and represents all possible structures in the structure ensemble

$$Z = \sum_{j \in S} e^{\frac{-\Delta G_j}{RT}} \tag{1.4}$$

In both equations (1.3) and (1.4), R is the ideal gas constant and T is the temperature in Kelvin. However, determining the partition function though enumeration becomes impossible for even sequences of moderate length as the number of possible structures grows exponentially with length [23]. In order to calculate the partition function, McCaskill made the key observation that structure energies are additive. This observation implies that they are multiplicative in the partition function [41]. Thus, this algorithm is an optimal calculation of the partition function in $O(n^3)$ time.

The partition function allows us to calculate structure properties while simultaneously considering many different suboptimal structures. Due to the sheer number of structures, the individual probability of any particular structure can be vanishingly small. The MFE structure is only one of many structure configurations. Instead of only representing the

MFE structure, the structure configuration can be represented as base-pair probabilities, or the percent of structures containing a particular base-pair [42]. Using base-pair probabilities improves on structure prediction accuracy [37, 43–45]. The ensemble is efficiently calculated during McCaskill's recursions, which have been implemented in the ViennaRNA folding package [31].

Suboptimal structures

The RNA structure ensemble represents all possible structures that a given sequence can fold into. Within the ensemble, there are multiple stable structures, referred to as suboptimal structures. Often, the biologically relevant structure is a suboptimal structure near the MFE [46]. Since the relevant structure is likely to be a stable structure within ensemble, instead of considering the full ensemble, we are only concerned with conformational changes to the low energy structures. Using the Boltzmann distribution, these low energy structures can be sampled from the secondary structure ensemble proportional to the structure stability.

Centroid structures

Instead of considering the ensemble as a set of structures, the ensemble can be represented as a single average structure. The centroid structure is defined as the structure that minimizes the base-pair distance between it and all other ensemble structures. This structure represents an average structure thus removing unlikely base-pairs only found in the MFE structure. Programs such as Sfold [36] and centroidfold [47] estimate the centroid structure using the Boltzmann weighted ensemble. Centroid structures may be more accurate than MFE structures because they contain more of the base-pairs predicted by comparative genomic approaches [36, 43].

Within a structure ensemble, there can be alternative stable structures that are composed of different base-pairs but have similar 3D shape. Thus, the centroid structure may not adequately represent all structure clusters. RNAshapes merges adjacent base-pairs into stems thus generalizing the individual base-pairs into an overall shape represented

as stems [48]. Additionally, RNAshapes samples representative structures from each structure cluster within the ensemble, which can improve structure prediction accuracy. However, the cluster containing the true structure is rarely known a priori.

1.1.3 Predicting consensus structures from RNA Alignments

Modern RNA structure prediction, homology searches, and *de novo* ncRNA discovery center around building a multiple sequence alignment (MSA) of similar sequences. Sequence alignment has been a fundamental technique used in comparative genomics and is the gold standard for discovery in all genomic, proteomic, and ncRNA studies. The growth of comparative genomics has been fueled by the exponential increase in genomic data resulting in much more evidence to support findings. In the case of structure prediction, aligning multiple sequences with common evolutionary origin, improves structure prediction accuracy because of the multiple examples of shared structure [18].

The MSA is generated by systematically aligning sequences in such a way that the maximum number of matching residues are put into the same column. However, building the globally optimal alignment is a computationally complex problem that is NP-complete [49]. Therefore, alignment algorithms must balance speed and accuracy. When only considering sequence, the scoring function only needs to reflect the magnitude of selection for a given residue change. Modern algorithms use heuristics for alignment that calculate locally optimal solutions. Progressive alignment techniques such as ClustalW [50] work by aligning sequences starting with the most similar sequences. Then the algorithm constantly updates the entire alignment every time a new sequences is added. Alternatively, instead of relying on scoring matrices and heuristics, alignments can be built using probabilistic models such as hidden Markov models (HMM), specifically the profile-HMM. Profile-HMMs are probabilistic models trained on specific sequence profiles and are used to generate the most likely set of matches and gaps in the alignment based on both the sequence position and the probability of transitioning between match and gap states. The advantage of the HMM over heuristic methods is that it produces a model in addition to the alignment, thus any new sequences can be aligned without having to redo the multiple sequence alignment process. Additionally, because of the

consideration of position, the model automatically puts insertion states together leading to a simple biological interpretation (e.g. highly variable = loop region).

A purely sequence based alignment approach has limited success in aligning RNA sequences. RNA sequences can be more difficult to align than DNA or protein sequences due to the degeneracy of the RNA sequence relative to structure. Since the structure, not sequence, is conserved, the selective pressure to maintain the structure becomes more obvious when multiple co-varying mutations occur in stems. These mutations occur such that the base-pairings are maintained but the original nucleotides change. For example, let a sequence S contain a A-T base pair at (i, j). If an A -> G mutation followed by a T -> C mutation occurred, then, in essence, there was an A-T -> G-C base-pair mutation. The structure remains the same while the new sequence differs from original sequence S at two positions. Therefore, in addition to sequence, RNA alignment algorithms must also account for the thermodynamic stability and structure maintaining co-varying mutations.

Depending on the degree of sequence conservation, there are multiple approaches to determining a RNA consensus structure from an alignment: align-first, fold-first, or fold and align simultaneously [51]. In the align-first approach, a sequence aligner, such as ClustalW, aligns a set of sequences to make an MSA. The advantage of starting with a sequence alignment is that it can be generated quickly and techniques for sequence alignment are well understood. Because of the considerable effort invested into sequence alignment algorithms, a large number of sequences can be aligned quickly. High identity sequences are likely to be correctly aligned thus yielding higher quality alignments. Once the sequences are aligned, the MSA consensus structure can be calculated using RNAalifold [52]. RNAalifold calculates an optimal consensus structure by combining the Turner energy model with additional parameters that adjust for evolutionary conservation and sequence identity. The downside of the align-first approach is that it will perform significantly worse with $< 50\%$-60% sequence identity [51].

Since structured RNAs have shared structure, the fold-first approach seeks to align the secondary structure stems to each other. This approach is primarily used in motif detection. First, each sequence is folded independently. Next, the sequences and structures

are aligned using the common pairing elements. Programs that can be used for motif detection include: RNAshapes [48, 53], and RNAsampler [54]. Just as with sequence alignment, the structure is aligned using different algorithms, depending on the program. However, the consensus structure is generated by combining conserved stems that are compatible with all sequences in the alignment. The advantage of fold first is that aligning the common secondary structure elements better tolerates low sequence identity and identifies common shapes.

Simultaneous alignment of both sequence and structure leverages all available information and can produce an optimal alignment. The optimal solution is guaranteed by the Sankoff algorithm [55]. However, the algorithm is very computationally slow, with a runtime of $O(L^{3N})$, where L is the sequence length and N is the number of sequences being aligned. In the best case scenario, aligning two sequences will take L^6 time, which is prohibitively expensive as each doubling of length increases the run time by a factor of 2^6. Because of the algorithmic complexity, almost no programs directly implement the Sankoff algorithm. To compensate for the algorithmic complexity, programs such as Dynalign [56, 57], Foldalign [58], PMcomp [59], Stemloc [60], mxscarna [61], and LocARNA [62] incorporate heuristics and they are limited to aligning a small number of sequences. Programs built upon LocARNA (LocARNA-P and SPARSE [63, 64]) can now align many more sequences than their predecessors.

1.1.4 Discovery and post-processing

De novo non-coding RNA (ncRNA) discovery in genomic sequence is largely accomplished with computational tools that identify a stable thermodynamic structure that is maintained across many species [65–68]. While thermodynamic stability alone is not sufficient to distinguish functional structured RNAs from random genomic sequence [69], the rapid growth of sequence databases has allowed the use of comparative genomics to determine whether such putative stable structures are conserved, and to identify the characteristic co-varying mutation pattern of structure conservation within predicted pairing elements [18]. Potential candidate alignments are post-processed to filter likely candidates. The process starts by choosing a sequence, or a small number of similar sequences

that fold into a stable structure. A major challenge is that these sequences may or may not contain a functional RNA structure. Then, machine learning algorithms screen the sequences for conserved sequence and structure, and finally producing an alignment.

The main challenge to identifying structured RNAs from genomic data is the difficulty of differentiating false positives from true positives. The genomic sequences often have a shared lineage. These RNAs are difficult to separate because there is a need for both multiple homologous RNAs, and shared stable structure. However, other than structured RNAs, many homologous sequences will be identified from genomic data with a shared lineage. Filtering on stable structure alone is not enough because many of these sequences are likely to be predicted to fold into stable structures [69]. The number of false positive hits is exacerbated when searching larger sequence databases.

In order to reduce the false positive rate, many RNA discovery programs have implemented a machine learning model to act as a post-processing filter. These models are the result of machine learning algorithms that have been trained to differentiate structured from unstructured RNAs using positive and negative training examples. The machine learning algorithms learn to separate positive and negative training data. Once the model is trained, it can act as a classifier or predictor for any new data, such as an RNA alignment. The training data must be carefully constructured such that the algorithm does not over-fit the data, which would cause the resulting model to be biased and only perform well on the training data.

Training machine learning models requires both positive and negative data. The positive examples are true structured RNAs, such as those found on the RNA families database (RFam) [70, 71]. Negative training examples are artificially generated. These negative data sets are built to mimic true positives by having some but not all qualities. There is a lot of emphasis placed on negative training examples because of the impact it has on the final model. For example, base-stacking interactions are major parameters in the Turner energy model. If random sequence does not account for base composition, then the model will differentiate training data purely on sequence composition instead of structure [72–74].

1.1.5 Probabilistic models

The structure prediction performance of stochastic context free grammars (SCFG) and thermodynamic energy models are comparable [26, 75]. SCFGs are stochastic context-free grammars (CFG) that can be trained to fold RNAs by statistically learning common structure features. A CFG is a set of production rules that can generate a structure from the inside outwards. For example, all possible dot-bracket structure can be generated using the grammar

$$S \rightarrow \bullet | S \bullet | (S) | S(S) \tag{1.5}$$

where if i and j base pair, then $j-i \geq 1$. An example derivation of the structure (..).((..)) is

$$S \rightarrow S(S) \rightarrow S\bullet(S) \rightarrow (S)\bullet(S) \rightarrow (S\bullet)\bullet(S) \rightarrow (\bullet\bullet)\bullet(S) \rightarrow (\bullet\bullet)\bullet((S)) \rightarrow (\bullet\bullet)\bullet((S\bullet)) \rightarrow (\bullet\bullet)\bullet((\bullet\bullet)) \tag{1.6}$$

A CFG can be made into an SCFG by assigning probabilities to all productions from a given production rule such that the probabilities sum to 1. By aligning a trained SCFG to a given sequence, the optimal alignment represents the most likely structure. An advantage of using SCFGs is that the grammar produces structures that account for the nested long-range base-pairing interactions that have been conserved through evolution. The probabilities are learned, SCFGs must be trained using high quality alignments. A limitation of using SCFGs is that they cannot fold RNA structures that do not resemble the training set, whereas algorithms using the thermodynamic energy model can fold any given input sequence.

SCFGs can be incorporated into increasingly complex models that account for phylogenetic distance. By employing phylo-SCFGs, Pfold [76, 77] showed the necessity of considering phylogenetic information in structure prediction. Combining the thermodynamic energy model with Pfold, PETfold [78] further reinforced the idea that phylogenetic information improves prediction accuracy. By training the phylo-SCFG from Pfold on different input alignments, EvoFold [79] leveraged an 8-way genome wide alignment to identify conserved ncRNAs in humans.

A major application of probabilistic models is in RNA homology searches [26]. Profile HMMs have been successfully used to identify and align homologous proteins in protein databases [80]. HMMs are not suitable for RNA homology searches because they only consider local interactions and cannot deal with the long range interactions of RNA secondary structure. Instead, RNA structure profiles are modeled using covariance model, which is the SCFG analogue to the profile-HMM. Infernal [81] is the current state-of-the-art program for RNA homology searches. As part of the input, Infernal requires a CM (also produced by Infernal) that has been previously trained on an RNA alignment. The advantage of using a CM is that it simultaneously aligns sequences and predicts structure. Additionally, the probabilistic model assigns a likelihood score to each sequence, which allows ranking hits. Using a CM trained on tRNAs, tRNAscan-SE has been used to identify tRNAs with high fidelity from genomic data with a low false discovery rate [82].

1.2 Distance measures

Throughout this work, we will compare sequence, structures and probability distributions against each other. Distance measures are useful for determining the number of differences between sequences or structures as well as classifying RNA structures from "other." The most common type of data we work with are sequences, which are represented as strings. In order to compare strings, we use Levenshtein (or string edit) distance or general string alignment. Levenshtein distance returns the minimal number of changes to make the two strings identical. The Levenshtein distance is normalized to the length of the longer sequence so that strings of different lengths can be compared. This distance metric does not return an alignment, but it is fast, which makes it especially useful for generating sequence clusters from high-throughput sequencing data. String alignment has a similar in run time to string edit distance [83]. The main difference is that string edit distance uses the same penalty for gap open and gap extension. This difference can result in discontinuous gaps, which is less preferable in sequence alignment.

Given that structure is crucial to function, many calculations in RNA evolution models heavily rely on RNA folding and distance measure algorithms. Summarized in a review by Cowperthwaite and Meyers, RNA evolutionary models use changes to secondary structure are a proxy for the change in fitness [84]; therefore, these models assume there is a direct correlation between changes to the secondary structure and fitness. Because there are different structure representations, there are different algorithms for calculating structure distance. The most common distance metric is the base-pair distance (equation 1.7). The base-pair distance between two structures T and T' is calculated as the number of base-pairs found in one structure but not the other, represented as

$$\text{base-pair distance}(T, T') = |T \cup T'| - |T \cap T'| \tag{1.7}$$

where the structures T and T' are the set of base-pairs found in T or T', respectively. A drawback of comparing structures with base-pair distance is that it can only be used to compare structures with the same length sequence. This limitation derives from the fact that base-pairs are represented as a set of ordered pairs (i.e. an (i,j) pair in a structure of length 10 is not the same as an (i,j) pair in a structure of length 100).

Representing unordered data as a probability distribution is useful in identifying the difference between divergent sequences using base frequency. By representing base frequency as a probability distribution, the distance between two sequences can be calculated using the Kullback-Leibler divergence (KLD). Given two probability distributions (P and Q) where P is the true distribution and Q is the observed distribution, the KLD calculates the relative entropy between P and Q. Intuitively, the relative entropy between P and Q is a measure of "surprise" or information gain from using P beyond the information of Q. The KLD is calculated as

$$KLD(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{1.8}$$

An application of KLD, and other information theory metrics, is to elucidate common short sequence motifs in sequences. These short sequence motifs can be represented as both probability distributions and pairwise weight matrices [85]. The disadvantage of using KLD is that the divergence is not symmetric and there is no upper bound on

the relative entropy, which makes it difficult to determine the distance between multiple sequences. Therefore, we primarily use Jensen-Shannon divergence (JSD), a modification of KLD, to calculate the difference between two probability distributions. JSD is calculated using KLD

$$JSD(P||Q) = 0.5(KLD(P||M)) + 0.5(KLD(Q||M)) \tag{1.9}$$

where M is an average of the two distributions calculated as

$$M = 0.5(P + Q) \tag{1.10}$$

This distance measure is useful because it is symmetric and bound between 0 and 1. A symmetric distance measure refers to the fact that the distance between two distributions P and Q is the same regardless of the order (i.e. JSD(P||Q) = JSD(Q||P)). Since JSD is a distance measure, 0 indicates the two distributions are the same and 1 indicates completely different. Being bound between 0 and 1 is useful in machine learning because features calculated using JSD are readily normalized.

1.3 RNA evolution models

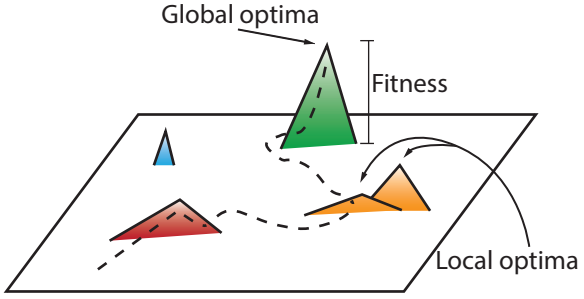

FIGURE 1.3: **The RNA fitness landscape.** An extension of the neutral network, every sequence on the network folds into the same structure. Each point on the fitness landscape is a sequence (genotype) and any two sequences ($a = a_1, ..., a_n$ and $b = b_1, ..., b_n$) are connected by an edge 1) if they differ by a single nucleotide or 2) they differ at two nucleotides such that $a_i \neq b_i$ and $a_j \neq b_j$ for an (i, j) base pair in the MFE structure. The height of the peak is proportional to the fitness.

Non-coding RNA discovery borrows some concepts from the evolutionary biology field. As previously discussed, RNA folding has been essential to the RNA discovery field because it focuses on sequences with stable structure. Some RNA folding algorithms such as evoFold [79], pFold [76, 77], and PETfold [78] rely on phylogenetic relationships between sequences to identify the structure. The key observations derived from RNA function studies are that there is a direct link between sequence and structure, and that functionality depends on structure. Therefore, RNA has a direct genotype (sequence) to phenotype (structure) relationship. Numerous computational models have been performed *in silico* using a finite population of RNA molecules replicating with a set mutation rate [11, 86–89].

The relationship between genotype and phenotype is often represented as a fitness landscape, which is an abstract framework for showing the relative fitness advantages and disadvantages for a set of genotypes [90]. Typically drawn in 3-dimensions, the xy-plane represents the genotypes as points. For a given genotype, the height corresponds to the relative fitness (Figure 1.3). In this model, genotypes represent sequences and higher fitness genotypes are selected for by evolution, thus over time, the population moves towards higher fitness. These evolutionary steps are the equivalent of moving across the fitness landscape, which is naturally represented as a network graph known as a mutational landscape. On this graph, each node is a sequence and an edge connects two sequences ($a = a_1, ..., a_n$ and $b = b_1, ..., b_n$) 1) if they differ by a single nucleotide or 2) they differ at two nucleotides such that $a_i \neq b_i$ and $a_j \neq b_j$ for an (i, j) base pair in the MFE structure. The height of the peak is proportional to the fitness.

1.4 Thesis organization

In this dissertation, I present my work in two parts. In my first project, we utilize sequence neutrality as a novel feature in an RNA structure classifier. In order to do so, we developed a novel method for calculating neutrality known as sampled ensemble neutrality (SEN). SEN improves on existing measures by heavily weighting changes to the existing structure averaged over the Boltzmann low energy RNA structure ensemble. We show this measure, compared to existing measures, has a larger dynamic range, and

increased separation between known structured RNAs and decoy sequence. Furthermore, RNA classifiers using neutrality alone or in combination with existing features have high performance. My second project focuses on understanding how evolution derives distinct RNA structures with analogous function. To do so, we analyze high-throughput sequencing data from an *in vitro* selection experiment designed to evolve a random RNA sequence pool to bind a target protein and identify common motifs that could be responsible for RNA-protein interactions. Analysis of the selection data show high sequence and structure diversity. During the course of the analysis, we develop a novel algorithm to elucidate sequence specific substructures that are enriched in the sequence pool over time. These two projects combine to show that structure conservation is a pervasive theme throughout RNA evolution and structure conservation can be exploited to identify structured RNAs and possibly ligand binding sites within an RNA structure.

Chapter 2

Sampled ensemble neutrality as a feature to classify potential structured RNAs[1]

2.1 Background

Machine learning techniques, specifically support vector machines (SVMs) [66, 68], leverage both the thermodynamic stability of structured RNA, and the presence of covarying mutations as an indicator of conserved structure, to distinguish alignments corresponding to putative biologically functional structured RNAs from alignments of sequences conserved for other reasons and non-conserved thermodynamically stable structures. There are six quantifiable features commonly used by *de novo* ncRNA prediction approaches including: the thermodynamic stability of the structures formed by individual sequences, as measured by the mean of the Z-score of the minimum free energy (MFE) structure of sequences in a putative alignment [72, 74]. The ability of the alignment sequences to fold into the common predicted consensus structure is measured by the structure conservation index [66]. The extent to which sequences are diverse and contain covarying mutations is measured by the mutual information [16], entropy of base-pairing regions [91], and the

[1] Adapted from Pei, Shermin, Jon S. Anthony, and Michelle M. Meyer. "Sampled ensemble neutrality as a feature to classify potential structured RNAs." BMC Genomics 16.1 (2015): 1.

mean pairwise sequence identity between alignment sequences. Finally, because more sequences lead to higher prediction accuracy, the number of sequences in the alignment is a common feature.

There exists a facile computational link between RNA sequence and secondary structure due to the considerable efforts toward RNA secondary structure prediction. As a result, simulation of RNA evolution using structure as a proxy for fitness has been used to explore a variety of evolutionary ideas [92–94]. These studies have shown that sequences with the same structure can be represented on a neutral network [11, 95]. *In silico* experiments reveal that some structures are mutationally robust because they have large networks of highly connected sequences [86] allowing them to maintain structure while tolerating many different mutations. Using *in silico* methods, mutational robustness has been demonstrated for naturally occurring RNAs such as pre-miRNAs [96] and virus genome elements [97], though RNAs without structure (e.g. sRNAs) do not seem to display this feature [98, 99]. However, these results heavily depend on the structure and inverse folding algorithm used [100].

Mutational robustness, therefore, should be a feature that can distinguish between random putative structures formed by genomic sequence, and biologically relevant ncRNA structures. Robustness is measured using neutrality, which is calculated as the mean secondary structure similarity (i.e. normalized base-pair similarity) between a sequence and those that differ by exactly one point mutation (1-mutant neighbors) [96]. There are a variety of existing computational methods [101] and programs designed to evaluate RNA robustness (e.g. RNAmute, RDMAS, RSRE, RNAmutants, SNPfold, RNAsnp, RemuRNA, and Rchange) [6, 12, 102–107]. All of these approaches focus on a single input sequence and the ability of its neighboring mutants to maintain a "wild-type" structure. RNAmute, RSRE, and RDMAS evaluate the normalized base-pair similarity between an MFE starting structure and the low energy suboptimal structures generated for mutant sequences using the Vienna RNA package [102–104]. However, using the MFE structure as the sole reference limits the accuracy of predicted structure-disrupting mutations [108]. RNAmutants samples mutant sequences and structures according to their probability in the structural ensemble to identify sequences that severely disrupt structure, but fundamentally determines the structural disruption based on the MFE structure of

the mutant [105]. To improve the accuracy of structure comparisons, SNPfold compares the structure ensemble of an RNA sequence with that of its mutants using Pearson's correlation coefficient (PCC) [6], and RNAsnp uses this measure in combination with the base-pair distance to evaluate structural similarity and disruption [106]. RemuRNA measures the effect of a mutation on the entire RNA secondary structure distribution using relative entropy rather than sampling from the structural ensemble [12]. Alternatively, Rchange takes a different approach and reports the expected change in mean ensemble free energy and thermodynamic entropy of structures [107].

In this work, we propose utilizing sequence neutrality as an SVM feature to classify potential structured RNAs. To do so, we introduce a new measure of neutrality, the structural ensemble neutrality (SEN). Similar to previous efforts to assess RNA robustness, this measure considers the thermodynamic ensemble of structures for 1-mutant neighbors and their difference from a given reference structure. However, rather than utilize the MFE structure of our initial sequence as the reference structure, we utilize a structure that is derived from a multiple sequence alignment (MSA) of homologous RNAs to more accurately reflect the biologically relevant structure [109]. In addition, to account for the over-prediction of secondary structure elements relative to tertiary structure interactions necessary for function, our similarity measure prioritizes maintenance of the existing structure rather than considering all base-pair changes (both newly formed and broken base-pairs) as equal. We demonstrate that this measure of neutrality successfully distinguishes alignments of known bacterial structured regulatory RNAs from several different types of decoy data including both shuffled alignments and alignments constructed from intergenic or protein-coding sequence. We extend this finding to evaluate neutrality as a feature for classification of putative ncRNA alignments using an SVM. This analysis shows that neutrality can correctly classify ncRNA alignments nearly as well as the combination of existing features implying that the calculation of neutrality encompasses many of these existing features. Finally, we also show that many RNAs involved in bacterial regulation are mutationally robust using the structural ensemble neutrality.

2.2 Methods

Sequence neutrality

Before calculating neutrality, some common variables must be defined. Let a given input sequence S, without gaps and of length L, fold into a structure T. The set of sequences that differ from S by one point mutation are the 1-mutant neighbors

$$1mut(S) = \{\text{1-mutant neighbors}\} \tag{2.1}$$

Additionally, the size of the set $1mut(S)$ is $|1mut(S)| = 3L$. A single 1-mutant neighbor of S is represented by S' such that $S' \in 1mut(S)$. Let the structure ensemble of S' be

$$e(S') = \{\text{structure ensemble of } S'\} \tag{2.2}$$

The set of all $e(S')$ created from $1mut(S)$ is defined

$$\Gamma_S = \{e(S')|S' \in 1mut(S)\} \tag{2.3}$$

We represent the set containing the MFE structure of the structure ensemble $e(S')$ as

$$MFE(e(S')) = \{\text{the MFE structure of } e(S')\} \tag{2.4}$$

which has size = 1.

$$T'_{Nsamp} = \{sample(N, e(S'))\} \tag{2.5}$$

is created using RNAsubopt which samples N structures with replacement from $e(S')$ according to their probability of occurrence [31, 46, 110]. Let the secondary structure be represented as an $L \times L$ adjacency matrix M where an entry

$$M_{i,j} = \begin{cases} 1, & \text{if position } i \text{ and } j \text{ base pair} \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

The base-pair probability matrix for all base-pairs i, j in T'_{Nsamp} is determined by calculating

$$BPROB(T'_{Nsamp}) = \frac{1}{|T'_{Nsamp}|} \sum_{T' \in T'_{Nsamp}} M_{T'} \tag{2.7}$$

where $M_{T'}$ is the adjacency matrix M for a sampled structure in T'_{Nsamp}. Alternatively, the base-pair probabilities can be explicitly calculated using 'RNAfold -p' in the Vienna RNA folding suite and parsing the resulting postscript file. However, we find this process to be somewhat slower in aggregate. The centroid structure represents the structure that has the minimal distance to all other structures in the Boltzmann low energy ensemble [36]. We approximate this centroid structure by identifying base-pairs that occur in more than half of the sampled structures and represent the structure in an array $CENT$ where each element in the array is 1 if $BPROB(T'_{Nsamp}) > 0.5$ and 0 otherwise

$$CENT_{i,j} = \begin{cases} 1, & \text{if } BPROB(T'_{Nsamp})_{i,j} > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{2.8}$$

For some similarity calculations, the secondary structure must be converted to an array representation. Given structure array V, let V_i be 1, if the ith position of the structure is a base-pairing character and 0, otherwise.

Neutrality calculations fundamentally rely on two factors: the accuracy of the two structures being compared (T and $MFE(e(S'))$ or $CENT(T'_{Nsamp})$), and the similarity calculation used to measure the similarity between these two structures. In this work, we develop a novel measurement of neutrality, the structural ensemble neutrality (SEN) and compare it with several existing neutrality measures. These include neutrality as determined by the programs RNAmute and RemuRNA. To allow direct comparison of different similarity measures, we implemented the normalized base-pair similarity (bp-similarity), and the Pearson's correlation coefficient (PCC). RNAmute takes a sequence S and reports neutrality. RemuRNA takes an input sequence (S) and calculates the Kullback-Leibler divergence (KLD) between $e(S)$ and each $e(1mut(S))$. In our assessment, we take the mean KLD over all 1-mutant neighbors.

We implement normalized base-pair similarity as

$$\frac{1}{|1mut(S)|}\sum_{S'\in 1mut(S)} 1-\frac{bpdist(T,MFE(e(S')))}{L} \tag{2.9}$$

where $bpdist(T, MFE(e(S')))$ is the base-pair distance between the given structure T and the MFE structure of S' $(MFE(e(S')))$ [96]. Structure similarity using PCC is calculated by

$$\frac{1}{|1mut(S)|}\sum_{S'\in 1mut(S)} \frac{1+PCC(V(T),V(cent(T'_{Nsamp})))}{2} \tag{2.10}$$

where $PCC(V(T), V(CENT))$ is the Pearson's correlation coefficient between the structure vector $V(T)$ and the centroid structure vector $V(CENT)$ created from 1000 sampled structures of $e(S')$ [108].

Our novel neutrality measure, the structural ensemble neutrality (SEN), leverages two factors to increase the biological relevance of neutrality measurements. First, we focus on maintenance of the core RNA structure (i.e. minimal structure for biological function). Rather than consider all base-pair changes deleterious, only base-pairs in the original structure T disrupted in T' are counted by our measurement. Second, we utilize a structure derived from comparative genomics as the reference structure T rather than the $MFE(e(S))$. This choice reflects understanding in the field that consensus structures defined from phylogenetic studies are much more likely to be accurate [111]. Structural ensemble neutrality is calculated by

$$\frac{1}{|1mut(S)|}\sum_{S'\in 1mut(S)} \frac{1}{|T'_{Nsamp}|}\sum_{T'\in T'_{Nsamp}} \frac{|T\cap T'|}{|T|} \tag{2.11}$$

T' is a suboptimal structure of S', $|T|$ is the number of base-pairs in T and $|T\cap T'|$ is the number of base-pairs shared by both structures; therefore, $\frac{|T\cap T'|}{|T|}$, a modification of Jacard distance, is the fraction of base-pairs in T retained in T'. To simplify equation 2.11, the similarity measure comparing T to T'_{Nsamp} is the mean fraction of bases retained

$$sim(T, T'_{Nsamp}) = \frac{1}{|T'_{Nsamp}|}\sum_{T'\in T'_{Nsamp}} \frac{|T\cap T'|}{|T|} \tag{2.12}$$

Here, $|T'_{Nsamp}| = 1000$ because sampling 10000 structures does not significantly improve the results and sampling 100 structures causes inconsistent results due to small sample size. To understand the margin of error from 1000 sampled structures, we randomly selected five alignments (RF01692, RF01826, RF00515, RF01767, RF01693) to calculate the standard deviation of SEN calculated neutrality. At 95 % confidence, we expect the margin of error to be 3.1% at worst. The median margin of error was 1.8%. Using a sample size of 1000, Substituting equation 2.12 into equation 2.11 results in

$$SEN = \frac{1}{|1mut(S)|} \sum_{S' \in 1mut(S)} sim(T, T'_{Nsamp}) \tag{2.13}$$

where

$$|A \cap B| = \sum_{0 \leq i < j \leq L} I[(i,j) \in A] I[(i,j) \in B] \tag{2.14}$$

and

$$I[(i,j) \in A] = \begin{cases} 1 & \text{if } (i,j) \text{ are paired in structure A} \\ 0 & \text{otherwise} \end{cases} \tag{2.15}$$

Alignment neutrality calculation

To streamline our process, we created a pipeline to calculate the neutrality of sequences in an MSA that can accommodate all neutrality measures in a uniform manner. This pipeline consists of a 3-step workflow and produces a alignment consensus structure T such that there are no non-canonical base-pairs or open and close base-pairs that are too close ($j - i <\leq 3$). Starting with a structure alignment, 1) S and T are created by selecting a sequence and simultaneously degapping both the sequence and structure. In addition, structure positions with non-canonical base-pairings (not Watson-Crick or G-U wobble) are considered single stranded. 2) From S, we calculate $1mut(S)$ (Equation 2.1) and Γ_S (Equation 2.3). 3) Neutrality is calculated by utilizing the similarity between the elements of Γ_S and T, which are calculated using a specified similarity function: normalized base-pair distance (bp-similarity) (Equation 2.9), Pearson's correlation coefficient (PCC) (Equation 2.10), or sampled ensemble neutrality (SEN) (Equation 2.13).

Test data

The test data sets were constructed using 35 seed alignments of regulatory structured RNAs found in bacteria (Table 2.2) from the RNA Families database (Rfam) [112]. Regulatory RNAs in bacteria were chosen due to the large size and diversity of alignments available, as well as the structural data that verify many of the predicted structures. Several data sets were constructed by varying how the positive and negative alignments were generated. Positive alignments were generated by either utilizing all sequences in the Rfam seed alignment (all), or a randomly chosen subset of 3-6 sequences (subset). Structural information for these alignments was either derived directly from the RFam seed alignment (given) or calculated using RNAalifold (predicted) [52] (Table 2.1). For each positive data set, a corresponding set of negative training alignments were created using one of three methods: using the program SISSIz to shuffle the alignment columns while preserving dinucleotide content of the positive alignments (shuffled) [113], gathering 5'-flanking, or 3'-flanking, genomic sequence for each entry in the alignment (5' and 3' respectively). To control for sequence versus structure alignment, the 5' and 3'-flanking sequences are aligned using ClustalW or Mxscarna [114]. All negative alignment consensus structures are calculated using RNAalifold [52].

TABLE 2.1: Summary of data set sources

Data set	Sequence	Structure	Negatives
1	subset	predicted	shuffled
2	all	given	3',5',shuffled
3	subset	predicted	3',5'

TABLE 2.2: List of *cis*-regulatory RNAs in Dataset2 and the number of sequences in each alignment

Rfam ID	Name	Number of sequences	Aligned sequence length (nt)
RF00050	FMN riboswitch (RFN element)	112	221
RF00057	RyhB RNA	26	71
RF00059	TPP riboswitch (THI element)	86	318
RF00114	Ribosomal S15 leader	62	140
RF00167	Purine riboswitch	106	113
RF00168	Lysine riboswitch	36	274
RF00234	glmS glucosamine-6-phosphate activated ribozyme	16	236
RF00379	ydaO/yuaA leader	93	335
RF00380	ykoK leader	78	221
RF00442	ykkC-yxkD leader	80	181
RF00504	Glycine riboswitch	40	215
RF00506	Threonine operon leader	17	142
RF00514	Histidine operon leader	22	167
RF00515	PyrR binding site	48	240
RF00522	PreQ1 riboswitch	31	70
RF00555	Ribosomal protein L13 leader	26	107
RF00557	Ribosomal protein L10 leader	84	234
RF00558	Ribosomal protein L20 leader	36	155
RF00559	Ribosomal protein L21 leader	36	151
RF01051	GEMM cis-regulatory element	104	136
RF01055	Moco (molybdenum cofactor) riboswitch	123	255
RF01057	S-adenosyl-L-homocysteine riboswitch	37	172
RF01070	SucA RNA motif	26	102
RF01385	isrA Hfq binding RNA	8	130
RF01402	STnc150 Hfq binding RNA	9	283
RF01482	AdoCbl riboswitch	5	161
RF01510	M. florum riboswitch	2	64
RF01692	Bacteroidete tryptophan peptide leader RNA	13	168
RF01693	Bacteroidales-1 RNA	7	210
RF01694	Bacteroides-1 RNA	8	96
RF01727	SAM/SAH riboswitch	12	55
RF01767	SMK box translational riboswitch	11	148
RF01769	Enterobacteria greA leader	19	129
RF01793	ffh sRNA	36	62
RF01826	SAM-V riboswitch	3	69
Total		1458	

Impact of alignment quality on SEN

In order to assess the impact of alignment quality on SEN values, we determined the difference between SEN values obtained using an entire Rfam seed alignment (full alignments, positive Dataset2) or subsets of this alignment (subalignments, positive Dataset3). The delta SEN (SEN of full alignment - SEN of subalignment) is an estimate for the distance from the "true" SEN value obtained when using a subset of sequences that may result in a lower quality alignment and structure. To gauge how the delta SEN corresponds to differences between the structure predicted from the subalignment and the given structure from the Rfam alignment we examined the delta SEN as a function of two measures of structural difference: the bp-similarity, and the ratio of the number of base pairs in the full alignment compared to the subalignment.

Positional neutrality

Let S'_i be the set of three possible point mutations of S at a given position i.

$$S'_i = \{S' \in 1mut(S) | S' \text{contains point mutation at } i\} \tag{2.16}$$

Positional neutrality is calculated by averaging equation 2.12 over S'_i

$$SEN(T, T'_{Nsamp}, i) = \frac{1}{|S'_i|} \sum_{S'_i} sim(T, T'_{Nsamp}) \tag{2.17}$$

Mutational robustness

For a sequence S to be considered mutationally robust, $neutrality(S)$ must be greater than the mean background neutrality (i.e. inverse folded sequences). Mutational robustness of S is calculated by comparing its neutrality to the mean neutrality of 100 inverse folded sequences (Equation 2.18).

$$neutrality(S) > \frac{1}{100} \sum_{i=1}^{100} neutrality(inv)_i \tag{2.18}$$

For each sequence tested for robustness, RNAinverse [115] was used to generated 10 inverse folded sequences and each of those are used to seed 10 random walks resulting in a total of 100 inverse folded sequences for each S. Input sequences were omitted if no inverse folded sequence could be made from its structure.

RNAinverse [115] was used to generate an initial null set of sequences for comparison. As an alternative, we also used RNAifold [116, 117] to generate inverse folded sequences. However, the alignment consensus structure is not necessarily the MFE structure, which often causes RNAifold to fail and return no sequences. Because of this failure-mode, we did not force the inverse folded sequences to have an MFE structure identical to the target structure when using RNAinverse. To maintain similar base composition [96], sequences that approximate solutions of inverse folding were constrained by Jensen-Shannon divergence (JSD) < 0.01 such that $JSD(S||S_{inverse}) < 0.01$. This process yielded an initial set of background sequences.

To ensure that background sequences generated by RNAinverse [115] are unbiased with respect to neutrality [118], the inverse folded sequences were used as a seed for a random walk along neutral sequences [98]. These neutral sequences are defined as sequences that fold into the target structure. As done by Rodrigo *et. al*, 4L steps are attempted and a step will be accepted only if the structure does not change. Any mutation that occurs in a base-pair will also get a compensatory mutation to restore base pairing. If the random mutation results in the base being changed to a G, then the compensatory mutation will be randomly chosen, with equal probability, between a C and U.

Support vector machine

To implement a binary classifier support vector machine (SVM), the LibSVM package [119] in R was used. The SVM uses the calculated features to classify an input sequence as either "structured RNA" or "other." The features used are a standard 6-feature set, including the Z-score of the MFE, structure conservation index, mean entropy of stems, mean mutual information of stems, mean pairwise identity and number of sequences [66, 68], and neutrality, which is calculated using the measures described above. Performance of the SVM is evaluated by using 10x cross-fold validation on a data set and compared by

calculating the area under the curve (AUC) in receiver operating characteristic (ROC) curve analysis.

Statistical analysis

All statistical tests were done in the R project for statistical computing. To test the significance of the separation of neutrality between structured and unstructured sequence, we used the Wilcoxon rank sum test, which is a non-parametric test and does not assume normally distributed data. Individual measures of neutrality were considered independently in this analysis.

To test correlation of neutrality using different measures, we first standardized the data by calculating the mean neutrality of RNA families. If a sequence is predicted to not fold into a structure then its neutrality cannot be calculated thus it is omitted. Then the correlation was determined using the Spearman's rank correlation coefficient.

Logistic regression was carried out using a generalized linear model where neutrality was used to predict the structure disruption, represented as 0 (no disruption) or 1 (disruption).

2.3 Results and Discussion

Reference structure and similarity measure impact calculated neutrality

A set of structured RNA alignments derived from Rfam seed alignments (Dataset2, Table 2.1, Table 2.2) was used to validate SEN as a measure of neutrality by comparing its performance to other measures that are the basis of most programs designed to capture RNA structural robustness: bp-similarity and PCC. First, bp-similarity performance was evaluated using both the original method which only takes an input sequence, implemented in RNAmute, and a modified version we implemented, which requires a given sequence and structure. By comparing these bp-similarity implementations, we examine the effect of the input structure on neutrality and establish a baseline performance to

compare SEN with existing methods. In addition, RNAmute can use two different structure representations to provide either a fine grained view (dot-bracket (db) notation) or coarse grained view (Shapiro representation) of structure to calculate base-pair distance. The neutrality RNAmute calculated using the db notation shows a small separation between structured (median = 0.8454) and unstructured sequences (medians = 5'-Clustal = 0.7807, 5'-Mxscarna = 0.7855, 3'-Clustal = 0.8069, 3'-Mxscarna = 0.8069, Shuffled = 0.7731) (Figure 2.1A). Using the Shapiro structure as an alternative representation to calculate neutrality shifted the neutrality lower (structured median = 0.7777, unstructured medians = 0.6553, 0.6850, 0.6925, 0.6925, 0.6615), but the results remain highly correlated (ρ = 0.9306) (Table 2.3) with the db structure notation results (Figure 2.1B) indicating similar performance. However, using our modified version of bp-similarity that imports the structure from the alignment does incrementally improve separation of structured RNAs and negative data (0.7654 vs. 0.6293, 0.7229, 0.6692, 0.6692, 0.6618) compared to RNAmute (Figure 2.1C) demonstrating that using the consensus structure from the alignment improves the accuracy of the structure. The correlation between using the MFE structure and a given structure (ρ = 0.565) indicates that using the given structure may improve the neutrality calculation but does not completely deviate from existing methods.

To assess alternative similarity measures to bp-similarity, we also compared the performance of SEN and PCC over Dataset2. Using PCC to calculate neutrality shows a better separation between structured (median = 0.2631) and unstructured sequences (medians = 0.4431, 0.4143, 0.4445, 0.4351, 0.5465) than bp-similarity (Figure 2.1D). While there is moderate inverse correlation to RNAmute (ρ = -0.608), using PCC shows that natural sequences have lower neutrality than unstructured sequence. This inverse correlation suggests PCC calculated neutrality is consistently opposite to existing measures. SEN performance creates the largest degree of separation between structured (median = 0.5991) and unstructured sequences (medians = 0.04368, 0.2625, 0.0791, 0.0789, 0.0215) (Figure 2.1E) as well as consistent performance to established methods (ρ = 0.608).

We also assessed RemuRNA, a program that compares the structural ensemble of an RNA sequence and its mutants. RemuRNA returns the KLD between the "wild-type" structure ensemble compared to the mutant-neighbor ensemble, therefore a low value

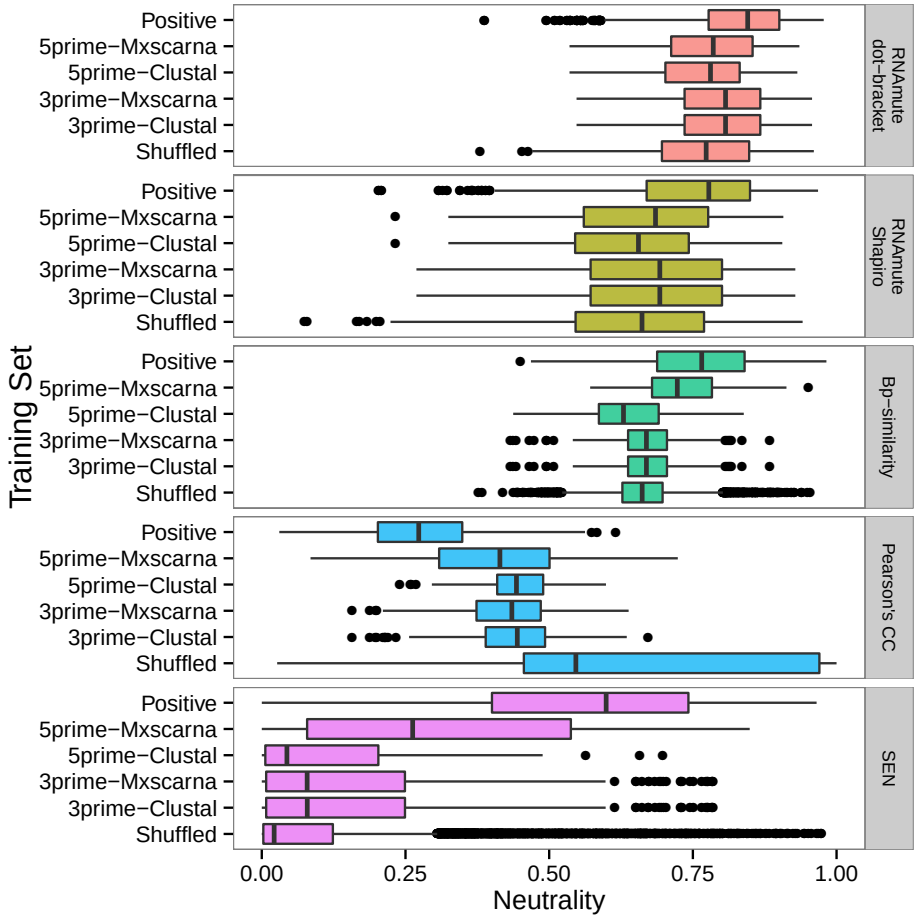

FIGURE 2.1: **SEN calculated neutrality has larger separation between structured and unstructured sequence.** Distribution of neutrality values from Dataset2 compare the performance of various distance functions **(A)** RNAmute dot-bracket representation, **(B)** RNAmute Shapiro representation, **(C)** bp-similarity, **(D)** Pearson's Correlation Coefficient (PCC), and **(E)** Sampled Ensemble Neutrality (SEN). The 3' and 5' flanking region used for negatives are referred to as 3prime and 5prime, respectively. The SEN on the positive test set has a larger separation between the negatives, compared to other measures. All similarity measures, except PCC, show unstructured sequence to be low on their respective scales. PCC calculated neutrality show structured sequences to be less neutral than unstructured sequence. Lastly, the SEN uses a large dynamic range of values compared to bp-similarity, which will increase its sensitivity between highly similar structures.

indicates that the mutant secondary structure distribution is not significantly different. Using RemuRNA, there is no significant difference between the positive sequences in Dataset2 (structured median = 2.3269) and most decoy sequences (unstructured medians = 2.244, 2.246, 2.271, 2.271). Shuffled sequences do show a significant loss of neutrality compared to other data (unstructured median = 2.785) (Table 2.4, Figure 2.2).

TABLE 2.3: Spearman's correlation between distance measures

			RNAmute	
	PCC	SEN	db	Shapiro
bp-similarity	-0.221	0.256	0.565	0.501
PCC		-0.614	-0.608	-0.595
SEN			0.608	0.651
RNAmute-db				0.930

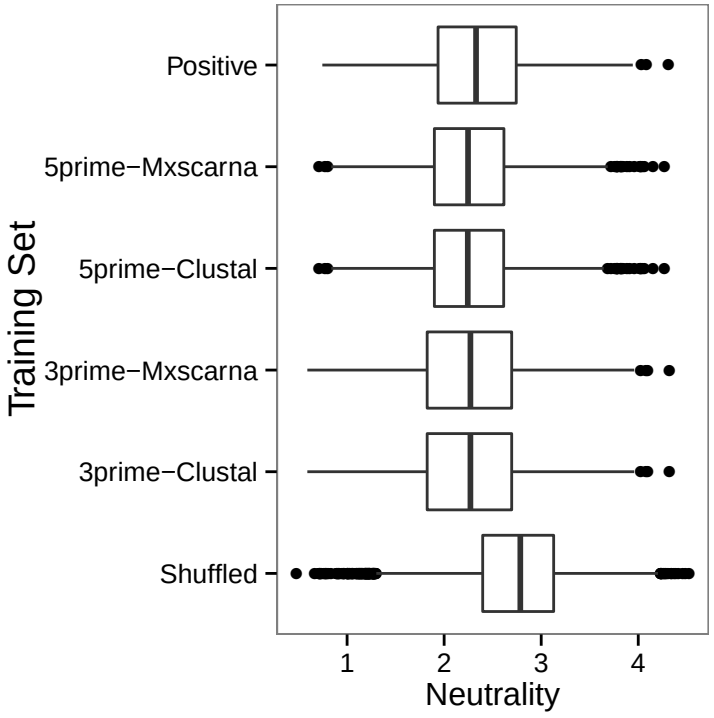

FIGURE 2.2: **RemuRNA performance against Dataset2.** RemuRNA distinguishes shuffled data but not other decoy sequences from positive sequences in Dataset2. Distribution of neutrality values from Dataset2 were compared using RemuRNA. The 3' and 5' flanking region used for negatives are referred to as 3prime and 5prime, respectively. There is no difference in the positive test set compared to the 5' and 3' negatives. There is a difference between the positive and shuffled alignments.

All the neutrality measures except RemuRNA we examined are able to distinguish between structured RNAs and negative sequence datasets with statistical significance (Table 2.4). The neutrality of negative sequences is near the bottom of the range for each measure. In addition, shuffled sequences are particularly easy to distinguish from structured RNAs using the PCC and the SEN compared with negative data derived from flanking genomic sequence. This, combined with the fact that RemuRNA is only able

TABLE 2.4: Wilcoxon rank sum determined P-values show significant difference between the neutrality of sequences

		3' Flanking		5' Flanking	
Similarity measure	Shuffle	ClustalW	Mxscarna	ClustalW	Mxscarna
SEN	0	2.06e-92	2.03e-92	4.99e-33	3.49e-21
Pearson's CC	0	1.20e-108	1.92e-84	5.07e-28	1.45e-23
bp-similarity	1.14e-285	7.40e-44	7.40e-44	6.12e-19	3.75e-05
RNAmute: dot-bracket structure	2.72e-107	2.15e-10	2.15e-10	8.19e-08	1.34e-09
RNAmute: Shapiro structure	4.51e-121	8.07e-16	8.07e-16	2.43e-09	3.45e-10
RemuRNA	4.68e-132	9.99e-01	9.99e-01	9.99e-01	9.99e-01

to distinguish shuffled sequences from structured RNAs, suggests that column shuffled alignments may not be the most effective way to generate negative data meant to mimic natural sequences. Aligning 5' and 3' flanking negative data based purely on sequence (ClustalW), or using more sophisticated algorithms that consider potential structure (Mxscarna), typically does not change the results. However, the 5'-flanking negative dataset aligned using Mxscarna (5'-Mxscarna) does show significantly higher neutrality as calculated by SEN. This higher neutrality is caused by a poorly conserved predicted structure where each structure is composed of a small number of predicted base pairs. This reduction in the number of base pairs in the reference structure (24.2 versus 10.9 mean base pairs per alignment for positive and 5'-Mxscarna, respectively) artificially increases SEN calculated neutrality as the potential number of base pairs that may be broken and considered deleterious is small. Despite this potential drawback in the SEN calculation, by combining an alignment based reference structure and relaxing the distance measure to consider only core structure, SEN calculated neutrality better distinguishes structured RNAs from decoy sequences than existing approaches. In addition, SEN utilizes a wider dynamic range that may allow it to have higher sensitivity. These properties are especially important for measurements that may be used as features in machine learning approaches.

Impact of alignment quality on SEN

In order to assess the effect of reduced alignment quality on SEN, we compared the difference between SEN values determined using an entire Rfam seed alignment (full alignment, Dataset2), and a subset of these sequences (subalignments, Dataset3). We observe a relatively small difference (delta) on most SEN values between the full and subalignment of the same ncRNA (Figure 2.3A). One common result of a lower quality alignment is altered predicted structure. To determine whether altered structure contributed to a large delta SEN, we examined the delta SEN as a function of base-pair distance between the predicted structure for the subalignment and the given structure of full alignment and found no strong correlation (Figure 2.3B). Since the structures for a given pair of full and sub alignments can vary in length, base-pair distance may be an imperfect comparison. Therefore, we also examined the delta SEN as a function of the ratio of the number of base pairs in the full alignment compared to the subalignment. (Figure 2.3C). From this comparison we observe that there are a small number of subalignments that are highly impacted by using subsets of the aligned sequences. Often, these are alignments that have limited biologically relevant structure in the Rfam seed alignment, and thus may be especially prone to overprediction of structure in the subalignment. Specifically the STnc150 Hfq binding RNA (RF01402) Rfam full alignment structure has many fewer base pairs than those predicted for the subalignments.

Overall we find that SEN is robust to changes to the alignment. Most SEN values derived from lower quality alignments are within 0.2 of the full alignment (Figure 2.3A). The SEN calculation does not depend on perfect accuracy of the consensus structure and tolerates minor changes to the number of base pairs present. This result suggests that even alignments of relatively few sequences can be used to calculate neutrality using SEN without a large decrease in accuracy.

Neutrality as an SVM feature

Given that most of the neutrality measures we examined exhibited a statistically significant difference between the structured and unstructured sequence, neutrality should be

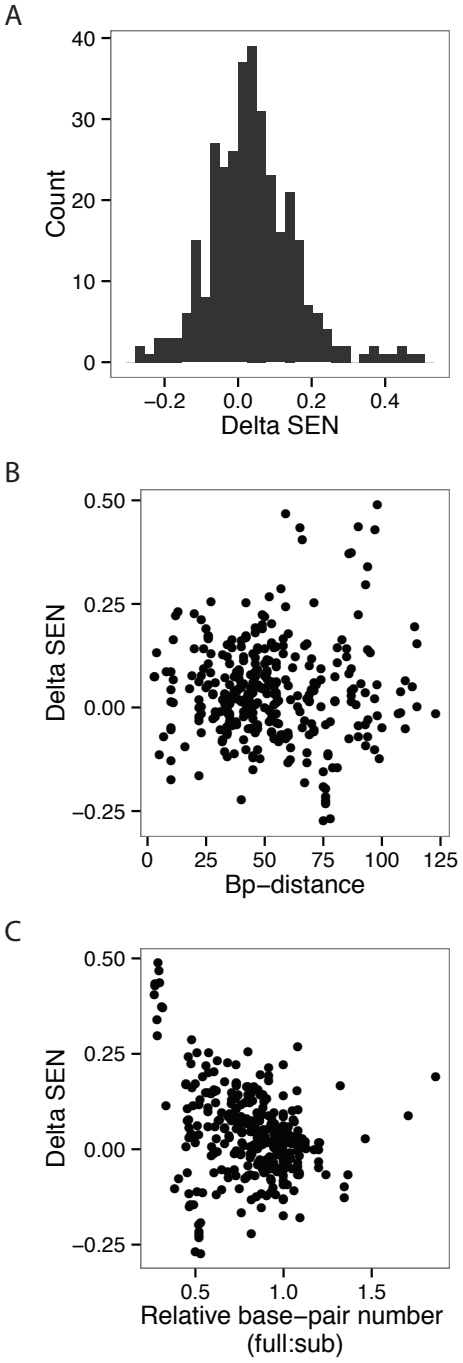

FIGURE 2.3: **Lower alignment quality has small impact on SEN** The effect of alignment quality on SEN. Lower quality alignments simulated by subalignments derived from Dataset3. The delta refers to (*delta* = full alignment SEN - subalignment SEN). **A)** Poorer quality alignments have a modest effect on SEN. **B)** No correlation is observed between the delta SEN and the base-pair distance between the structures derived from the full and subalignments. **C)** Large changes in relative number of base-pairs (full/subalignment) do affect SEN values.

a highly discriminative feature in an SVM binary classifier. Because of the large separation between structured and unstructured sequence, the classification performance of SEN and PCC was predicted to be comparable to each other and higher than bp-similarity. To test neutrality as a feature, we use neutrality as both an independent classifier and as part of an existing feature set for comparison with existing 6-feature SVMs [68]. First, as independent classifiers, neutrality calculated by both the SEN (Dataset2 AUC = 0.87, Dataset3 AUC = 0.903) and PCC (Dataset2 AUC = 0.864, Dataset3 AUC = 0.900) demonstrate a similar ability to correctly classify structured and unstructured sequence in all training examples regardless of sequence or structure origin (Table 2.5). Both of these methods significantly outperform bp-similarity (Dataset2 AUC = 0.735, Dataset3 AUC = 0.766). This is likely because SEN and PCC are less stringent forms of comparison than bp-similarity which equally weighs all base-pair changes, additions and disruptions.

TABLE 2.5: SVM performance using neutrality as a feature

Data set	Feature(s)	Area under curve (AUC)
Dataset1	6-feature set	0.918
	6-feature set + SEN	0.925
	3-feature set	0.927
	SEN	0.925
Dataset2	SEN	0.870
	PCC	0.864
	bp-similarity	0.735
Dataset3	SEN	0.903
	PCC	0.900
	bp-similarity	0.766

Natural RNA structures do not necessarily require all base-pairs to form a biologically relevant tertiary structure. It is common to see RNA alignments containing many homologs that have pairing elements of variable length, or with mismatches within pairing elements. From biology we know that these differences in structure do not necessarily affect function. Thus, because PCC only considers effects on the overall structure, and SEN only considers changes to the core structure they more accurately reflect requirements for biological function. Consistent with our previous analysis of delta SEN, SVM performance with Dataset2 (full alignments) and Dataset3 (subalignments) is comparable.

Next, to determine whether neutrality could be used as an additional feature to improve classification of putative ncRNA alignments, we added the SEN to the 6-feature set SVM revealing a marginal improvement with SEN (Dataset1 AUC = 0.925) verse without (Dataset1 AUC = 0.918). Interestingly the SEN used in isolation as a feature has equivalent performance (Dataset1 AUC = 0.925). Using the top 3 discriminative features (Z-score of MFE structure, mean mutual information of stems, and neutrality) also had comparable performance (Dataset1 AUC = 0.927) to using SEN alone.

Overall, neutrality as an independent classifier was able to separate structured and unstructured sequences. This finding is based on the similar classification performance when using either SEN or the currently used 6-feature set (Table 2.5). In fact, using the most discriminating features (Z-score of the MFE structure, mean mutual information of stems and SEN) offers comparable performance indicating the remaining features are redundant. The comparable performance of neutrality with existing feature sets is likely because current methods capture aspects of neutrality: structural maintenance despite sequence mutation and thermodynamic stability. The Z-score of MFE structure measures the thermodynamic stability which is also quantified in neutrality when comparing the alignment structure to 1-mutant neighbors ensemble of structures. The structure maintenance through covarying mutation is measured using the mean mutual information of stems which neutrality encompasses as the effect of single mutations on the structure.

Using SEN to detect structure disruption

One objective of many neutrality measures is to predict which bases are most disruptive to structure [6, 12, 106]. To evaluate whether SEN can be used to predict such bases, we sampled multiple sequences from our training set and interrogated the effect of position specific mutations on the calculated neutrality. By plotting positional neutrality for the purine riboswitch (RF00167), we observed that not every position has the same impact to the sequence neutrality. However, the neutrality predicted by SEN has consistent performance across multiple sequences drawn from the same alignment. In agreement with previous observations [107, 120], mutations to bases in structured regions (Figure 2.4) are more likely to be disruptive. In general, there are more disruptive mutations

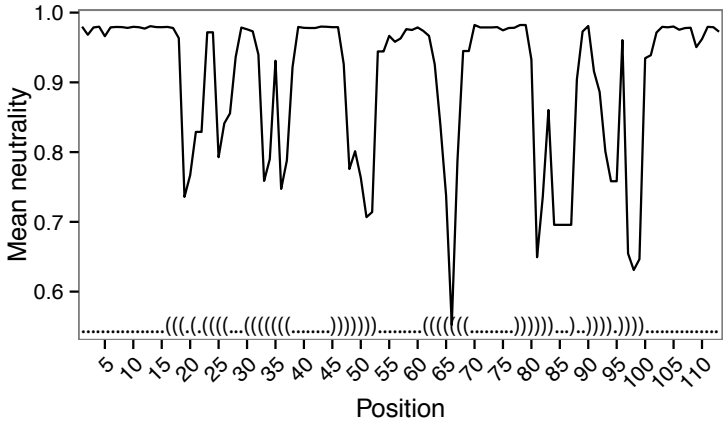

FIGURE 2.4: **Structure disruption generally occurs in stem regions.** Profile view of the purine riboswitch (RF00167) showing the mean neutrality at each position of all mutant neighbors at that position. The structure has been overlaid onto the graph. Mutations in the stems show larger structure disruption whereas mutations which occur in the single stranded regions do not significantly affect structure.

occur at the edges of stems. Mutations in the middle of stems appear to create either bulges or internal loops which have a small effect on the neutrality. Mutations in the loop regions also had little effect on the structure. We observed that there is a mutation at position 66 that strongly disrupts the structure. Given the strong impact, this mutation likely disrupts the formation of the stem.

To assess the accuracy of predicted structure disrupting mutations, we compared our predictions to experimental data obtained on the purine riboswitch using 2D SHAPE (Selective 2'-hydroxyl acylation analyzed by primer extension) [121]. Like evaluating neutrality using 1-mutant neighbors, 2D SHAPE interrogates the changes in RNA structure when making single mutations to an RNA sequence. To compare our predictions to the 2D SHAPE data, the reported change in base reactivity was converted to the expected structure disruption coefficient (eSDC) where $eSDC = (1 - PCC) * \sqrt{L}$ [108]. The top 50% of eSDC values are considered to be "structure disrupting." Logistic regression using SEN to predict structure disruption indicates that predicting which bases disrupt structure continues to be very difficult (AUC = 0.55) (Figure 2.5).

Current methods rely on RNA folding algorithms to predict which nucleotides can potentially be structure disrupting. Incorporating the structure ensemble does improve prediction accuracy [108] but such methods fundamentally still have poor performance.

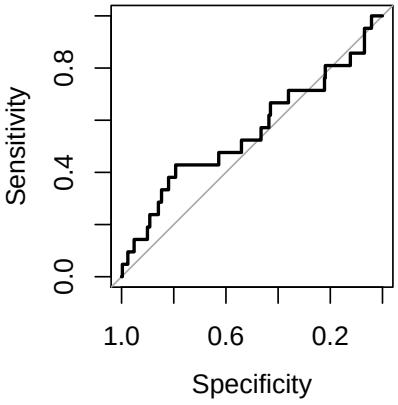

FIGURE 2.5: **Ability to predict which base mutations disrupt structure.** The receiver operator characteristic (ROC) curve shows the performance of SEN to correctly call structure disrupting mutations compared to random guessing (diagonal line). The line reveals SEN performs better than random.

The similar predictions of both SEN and current methods to detect structure disruption is likely due to the use of the same thermodynamic model for RNA folding that cannot fully encompass three-dimensional interactions, which results in similar prediction accuracy. However, the inability of SEN to make accurate predictions could also be due the limited data on structure disrupting bases derived from 2D SHAPE. Because a vast majority of positions have small impacts on structure, it is very difficult to establish the eSDC threshold at which the structure is disrupted. Furthermore, if the eSDC threshold is too high, then there is very little data available to build regression or machine learning models.

SEN detects mutational robustness

Finally, we use SEN to calculate the mutational robustness of positive sequences in our data sets. Robustness is defined as the ability of a sequence to maintain its structure despite perturbations to the sequence. The sequence is considered mutationally robust when its neutrality is greater than the mean background neutrality. Using SEN as a similarity measure detects 74.9% of the sequences in Dataset2 as being mutationally robust (Table 2.6). In comparison, using PCC (41.2%) or bp-similarity (40.5%) detected fewer robust sequences. The background neutrality calculated by PCC and bp-similarity

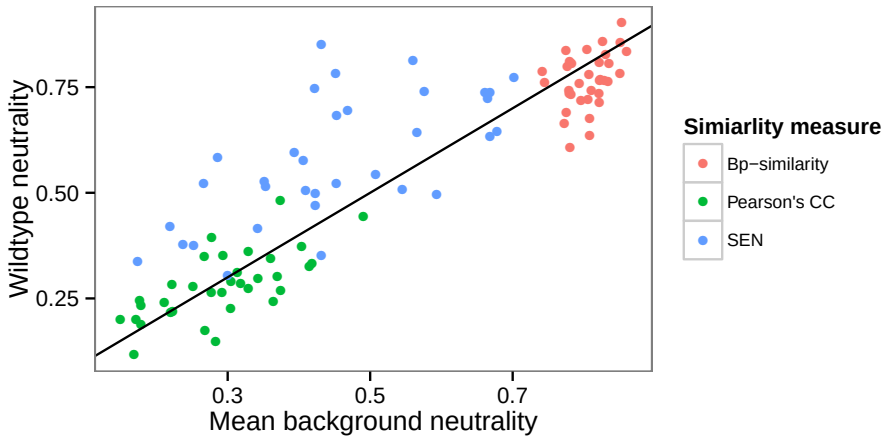

FIGURE 2.6: **Mean alignment neutrality organized by similarity measure.** The line represents wildtype sequence neutrality equal to mean background neutrality. If the wildtype sequence neutrality is higher than the mean background neutrality, the sequence is considered robust. To reduce the number of points, only the mean sequence neutrality for an alignment is compared against the average of the mean background neutrality. Plotting individual sequence neutrality reveals a similar trend (Figure 2.7). The SEN better detects mutational robustness of these sequences compared to PCC or bp-similarity.

is relatively high compared to the SEN background neutrality and likely contributes to the ability of distance measures to detect mutational robustness (Figure 2.6, Figure 2.7).

Using PCC or SEN calculated neutrality have equivalent performance as an SVM feature. However, using PCC neutrality shows that structured sequences are less neutral than unstructured sequence. This lower neutrality suggests that many sequences are not mutationally robust. The PCC calculation involves converting the structure into a binary vector; therefore, the base pairing information is removed and only the base-pairing status remains. By removing this information, the PCC potentially has difficulty differentiating similar distributions of 0's and 1's which could represent different structures. Bp-similarity had difficulty detecting mutational robustness in the data, likely due to the high stringency of the neutrality measure. Thus, existing commonly used measures of neutrality, normalized base-pair distance and PCC have potentially decreased accuracy for opposite reasons. The ability of SEN to detect mutational robustness in ncRNA regulators can likely be attributed to the hybrid nature of the calculation which still considers individual base pairs but is only concerned with maintaining the core structure and not with additional base pairs added by in 1-mutant neighbor.

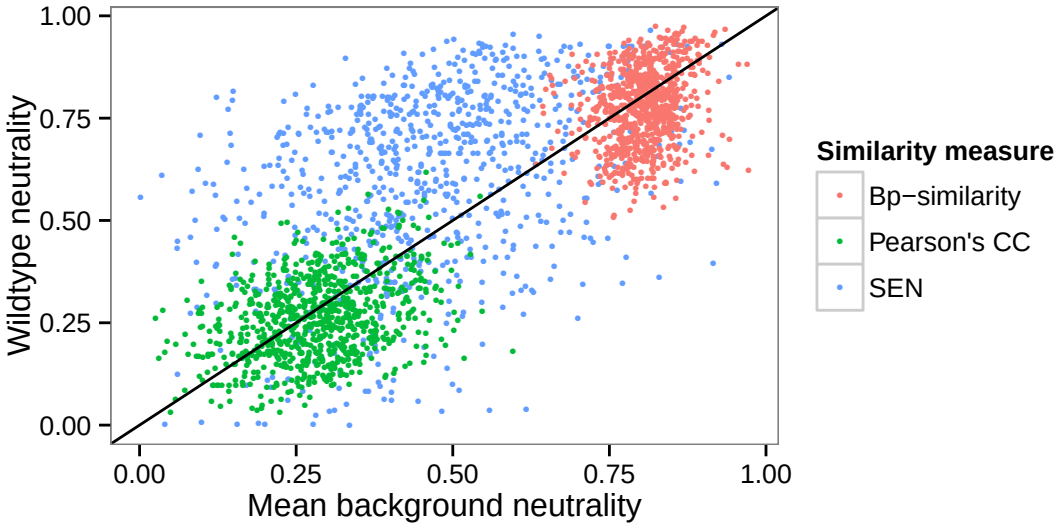

FIGURE 2.7: **SEN calls more sequences robust than other similarity measures.** The line represents wildtype sequence neutrality equal to mean background neutrality. If the wildtype sequence neutrality is higher than the mean background neutrality, the sequence is considered robust. Each point represents a sequence in the Dataset2 and it is compared against the average of the mean background neutrality. The SEN better detects mutational robustness of these sequences compared to PCC or bp-similarity.

TABLE 2.6: Fraction of robust sequences

Bp-similarity	0.405
PCC	0.412
SEN	0.749

SEN run time

SEN relies on the sampling of suboptimal structures from the ensemble of secondary structures. The run time is directly proportional to the number of sampled suboptimal structures and thus slower than traditional methods like bp-similarity. However, the calculation for each sample structure is identical so SEN calculations have been implemented to run in parallel, which can significantly reduce the run time. Code for calculating SEN is available at: https://github.com/ship561/sampled-ensemble-neutrality.

2.4 Conclusions

In this work, we show that RNA sequence neutrality is an effective feature for machine learning approaches to classify structured RNAs from various decoy sequences. We find that the most accurate classification occurs for neutrality measures that consider the ensemble of possible RNA structures rather than just the minimum free energy structure (PCC or SEN). Furthermore, neutrality used as the sole classifying feature is nearly as effective as existing SVMs [66, 68] indicating that current SVM features capture aspects of mutational robustness.

During the course of this work, we developed a novel measure of RNA sequence neutrality, the structural ensemble neutrality (SEN). The SEN differs from existing measures of neutrality in that it directly addresses several potential limitations. First, as a reference structure for neutrality calculation, SEN utilizes a consensus structure determined from an alignment of putative homologous sequences rather than an MFE structure, increasing the likelihood of utilizing a biologically relevant reference. Second, to assess the structure of the 1-mutant neighbors SEN considers not a single structure, but samples from the ensemble of potential low-energy structures. Finally, rather than consider all deviations from the reference structure equally deleterious, SEN only counts base pairs that are disrupted in the structure of the mutant sequence. This property renders SEN relatively robust to incomplete data that often degrades the quality of the predicted structure. The SEN is highly correlated with existing measures of neutrality (Table 2.3), but shows improved separation of structured and unstructured sequences in our data sets compared to these measures (Figure 2.1). While SEN's underlying model predicts structure disrupting mutations to occur in stems, this model does not completely explain experimental data (Figure 2.5) indicating there are other variables such as potential tertiary contacts to consider in such determinations. This result heavily depends on the similarity measure, accuracy of folding and inverse-folding algorithms. We found that PCC calculates lower neutrality for structured sequence, yet finds a similar proportion of sequences to be mutationally robust. SEN calculated neutrality indicates that many of the regulatory RNA structures in bacteria are mutationally robust (Table 2.6).

Chapter 3

Recognizing RNA structural motifs in HT-SELEX data for ribosomal protein S15 [1]

3.1 Background

RNA-binding proteins (RBPs) play essential cellular roles that range from co- and post-transcriptional regulation of mRNA transcripts [122, 123], to stabilization of macromolecular complexes such as the ribosome [124]. In this genomic era, the imperative to utilize primary sequence data to elucidate the relationship between an RBP, its recognition site, and its function, is only growing [125]. Identifying the binding sites for RBPs is an important task toward unraveling gene regulatory networks [126]. However, prediction of RBP interaction sites remains a challenge. Much of our understanding of RNA-protein binding motif identification comes from identifying transcription-factor binding sites. Following the assumption that RNA-protein interactions occur in single stranded regions, techniques to identify DNA-protein binding sites have been successfully applied to some RBPs. Unlike DNA-binding proteins (DBPs), RBPs may recognize features of single-stranded RNA, double-stranded RNA, or even three-dimensional tertiary

[1] Adapted from Pei, Shermin, Betty L. Slinger, and Michelle M. Meyer. "Recognizing RNA structural motifs in HT-SELEX data for ribosomal protein S15." BMC Bioinformatics. *Submitted*

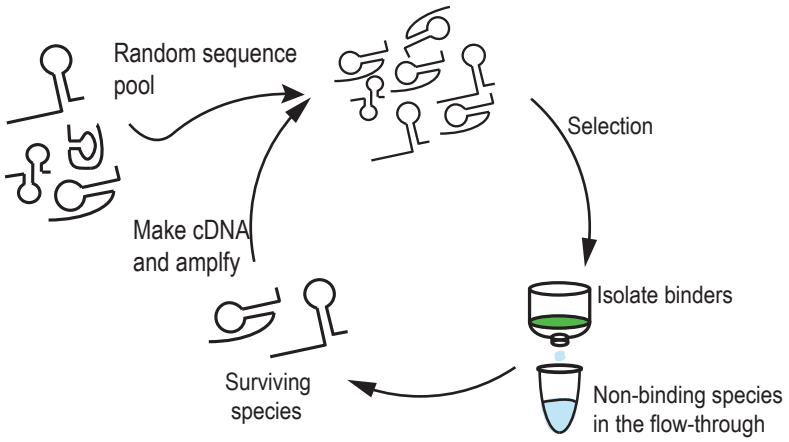

FIGURE 3.1: SELEX starts with a randomized sequence pool. The sequence pool undergoes selection to select binders and remove non-binders. The binders are reverse transcribed and PCR amplified. The cycle of selection and amplification is repeated until the sequence pool reaches sufficient binding affinity and diversity.

structures [127]. Therefore, RNA structure must be taken into account in assessment of potential binding-sites.

One method of experimentally identifying the constraints on a DBP or RBP recognition site is SELEX (Systematic Evolution of Ligands by Exponential Enrichment) [128, 129]. SELEX is an iterative *in vitro* selection technique that allows researchers to identify nucleic acids that interact with a target ligand. Starting with a random DNA or RNA sequence pool, each round in the selection is composed of a series of steps: 1) incubate the sequence pool with the target ligand, 2) remove non-binders, 3) elute binders, 4) reverse transcribe and PCR amplify the binders, 5) use binders as template for RNA into the next round of selection. The assumption is the aptamers will bind with varying affinity. As the number of rounds increases, those aptamers with the highest affinity will be selected for and become a larger fraction of the population. The SELEX process can continue for an indefinite number of rounds; however, over selection can lead to no sequence variation, and under selection results in a sequence pool that is not enriched with high affinity binders. Therefore, the number of SELEX rounds must balance binding affinity and aptamer diversity. Analysis of the sequences resulting from a SELEX experiment can be used to confirm the specificity of a binding site, or illuminate how RNA structural plasticity may enable multiple sequences to present a similar three-dimensional motif to the protein [130].

With the advent of next-generation sequencing, high-throughput sequencing-SELEX (HT-SELEX) has become an even more powerful approach to explore RNA-protein interactions. Sequence variation within the selected population gives insight into important residues, circumventing the need for laborious follow-up experiments to identify key regions of the selected sequences. The nucleotide differences between closely related sequences effectively explore local sequence space [131–135], and highly conserved areas point toward functionally important positions. Using such patterns of variation and conservation, information about the critical sequence motifs responsible for binding is revealed. Furthermore, sequencing intermediate rounds of the selection process allows ancestral sequences to be determined rather than inferred, and sequences that enrich over several SELEX rounds are more likely to be high affinity binders [136]. In addition, due to the high diversity of sequences undergoing selection, multiple possible and distinct binding motifs or structures can be discovered in a single experiment.

One downside of HT-SELEX approaches is the size and complexity of data that may be generated, especially from large randomized nucleotide populations. Typically, the RNA selection process starts with a pool of molecules on the order of $10^{12} - 10^{14}$ sequences, which can still be dwarfed by the total number of possible sequences ($4^{\text{sequence length}}$). In the ideal circumstance, over the course of a SELEX experiment, the sequence pool will converge on a small number of sequences that reflect a shared potential binding motif. If the entire sequence pool is sequenced, then these features should stand out as prevalent and enriching sequences within the population. In practice, given the size of the populations, under-sampling remains a significant hurdle. Thus, often only a sparse view of the RNA-binding pool is provided [132, 137, 138], potentially obscuring patterns that might be apparent from more thorough analysis.

Typical analysis of HT-SELEX data involves the identification of the RNA-protein binding motif. This analysis is distinct from transcription factor identification in that there can be multiple potential binding motifs and these motifs are likely to have a secondary structure context [139–141]. Programs found in the MEME suite [142] such as MEME, GLAM2 [143], and DREME [144] can be applied to the HTS data to identify binding motifs. MEME and DREME are designed to find contiguous sequence motifs. GLAM2 identifies motifs that can include short-gaps. However, there are a some of drawbacks

to using these tools. Due to their algorithmic complexity, MEME and GLAM2 are not equipped to use large magnitudes of sequence data [143, 145]. DREME's run time scales linearly with the data set size, but this is still not sufficient to keep pace with larger HTS data sets. Additionally, these programs ignore any potential secondary structure, which can hinder their ability to find the putative binding motifs.

To identify sequence-structure motifs, there are programs such as MEMERIS [139], RNAcontext [140, 146], AptaMotif [147], MPBind [148], Graphprot [149], RCK [85], AptaNI [150], and AptaTrace [151]. MEMERIS specifically identifies motifs found in the loop regions of the secondary structure, but like MEME, it is not designed for HTS data.. RNAContext and RCK use sequence and structure information to find RNA binding motifs, but they require a large number of both binder and non-binder motifs in order to determine the motif enrichment because it is assumed that the binding motif is contiguous and is present in majority of binders and not in the non-binders. MPbind uses a k-mer approach to identify contiguous binding motifs by identifying prominent subsequences that are enriched between selection rounds. Graphprot leverages secondary structure to identify binding motifs, but it also requires data on binders and non-binders alike. AptaMotif is designed to analyze low throughput SELEX data, but it has been extended in the form of AptaNI, which restricts the motif search to loop regions of the structure. AptaTrace is a state-of-the-art HT-SELEX motif identification tool that takes into both sequence and structure to identify binding motifs. Overall, many of these programs focus on identifying contiguous motifs while using secondary structure to restrict the search to single stranded regions.

HT-SELEX analysis techniques have been successfully applied to identify short sequence motifs responsible for RNA-protein interactions [152, 153], typically located in internal loop regions [154]. While this type of analysis is effective for many RBP binding-motifs, particularly those that involve recognition of single-stranded regions of RNA, not all RBPs conform to such recognition patterns [127]. In many cases an RBP may interact with complex tertiary structure motifs, and some cases with multiple complex structures. Some RNA binding proteins, such as ADAR or Staufen, specifically recognize double stranded RNA. These binding proteins target a structure containing 12 or 16 base-pairs, such as a single stem or co-axially stacked stems [155, 156].

In *Escherichia coli*, several ribosomal proteins interact not only with the rRNA, but also with structured portions of their own transcripts. These interactions allow stoichiometric production of ribosomal proteins by inhibiting transcription or translation [157]. While in some cases the mRNA structures are apparent mimics of the rRNA-binding sites, in other cases similarity is not obvious [158]. In addition, many of the mRNA structures responsible for this regulation in *E. coli* are narrowly distributed to only a few bacteria [159].

Ribosomal protein S15 is a particularly interesting example of ribosomal protein regulation. S15 is a conserved protein across bacterial phyla, and in some bacteria it is autoregulated at the translational level [160]. However, species within different bacterial phyla use distinct mRNA structures to accomplish the same regulatory task [159, 161, 162]. There are at least four distinct mRNA secondary structures that regulate in response to S15, each constrained to a single bacterial phyla. Each structure likely evolved independently, thus mRNA interactions with homologous S15 proteins are not necessarily conserved. In contrast, both the S15 protein and its 16S rRNA binding site are highly conserved among different lineages of bacteria. While previous work has identified the critical motifs in the 16S rRNA (a GU/GC within a paired region and a 3-helix junction) responsible for efficient S15 binding in *E. coli* and *Thermus thermophilus*, various mRNA structures can bind S15 despite containing some but not necessarily all of the 16S rRNA binding determinants [163–165]. Furthermore, not all homologous S15 proteins are interchangeable regulators between different bacterial species, indicating some target specificity [166]. Recently, we identified a set of SELEX derived RNA structures that bind *Geobacillus kaustophillus* S15 [167]. The identified RNAs are distinct from known natural regulators, but several still regulate gene expression in response to S15. Just as in nature, a high degree of sequence and structure diversity was found in this study, suggesting that the natural diversity of RNA regulation is not solely due to differences between S15 protein homologs.

In this work, we analyze the intermediate and final rounds of SELEX against *G. kaustophilus* S15 using high-throughput sequencing in order to better understand the diversity of potential RNA structures that interact with S15. The complex nature of the S15-binding site is a likely factor contributing to the high sequence diversity observed in

our data. To elucidate any sequence-structure motifs, we developed an analysis approach that simultaneously considers the sequence and structure to identify a discontinuous double-stranded binding motif. By treating RNA structure as a set of discrete substructures, we identify enriched structure elements associated with the RNA-S15 binding site. In particular, we find many potential binding motifs that are significantly enriched over the course of selection. Combining these motifs and experimentally validated binders, we build a model to separate specific and non-specific S15 binders. Overall, we find that S15 heavily relies on the structure for recognition of its target.

3.2 Results

Characterization of selected population

We characterized the reads resulting from sequencing reverse transcribed and amplified products of SELEX rounds 4, 9, 10, and 11 by examining read lengths, sequence enrichment, and diversity. There were 32,866,739 total pair-end reads of which 5,584,124 reads were forward strand and passed quality filters (Table 3.1) (See Methods: High-throughput sequencing). Most of the reads are the expected length of 87 nt (Figure 3.2A). The reads tend to become shorter in rounds 9, 10, and 11 compared to round 4. Additionally, we noticed there was an increase in fragments of approximately 79 nt length likely due to PCR amplification bias (Table 3.2). These shorter fragments were likely preferentially amplified during PCR compared to longer fragments. However, such individuals examined using filter-binding assays do not bind S15 specifically. We found that ≈2% of sequences from rounds 10 and 11 were enriched during the SELEX process (Figure 3.2B) indicating the selection is likely enriching for specifically binding sequences. Finally, there was incredible overall sequence diversity in the sequence pool. 95.33% of sequences appeared only once (singleton) and of the sequences that appeared more than once (multiton), 69.5% were seen fewer than 10 times (Figure 3.2C).

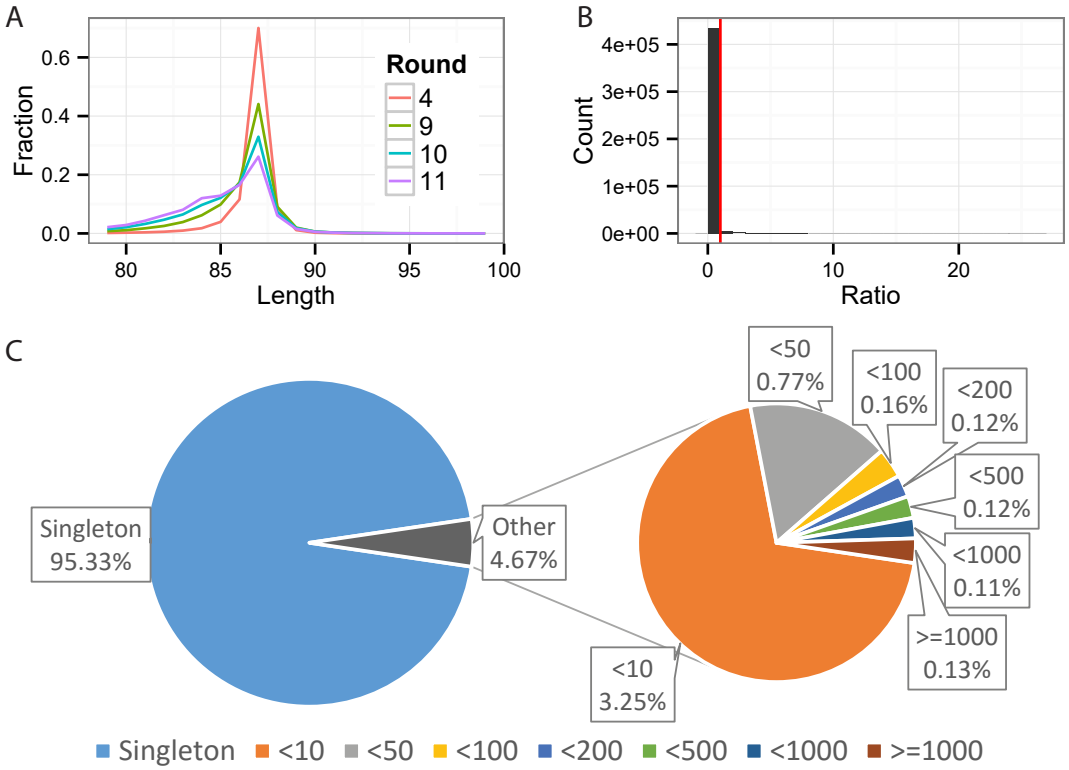

FIGURE 3.2: **A)** Distribution of read lengths shows most reads are the expected length of 87 nt. **B)** Distribution of sequence enrichment of multiton sequences in rounds 10 and 11. The enrichment is normalized to the total number of reads in the round. The red line indicates no enrichment (ratio = 1). **C)** Most sequences are singleton sequences.

TABLE 3.1: Total number reads by round before and after filtering.

Round	Unfiltered	Filtered
4	10,978,044	4,150,081
9	10,854,647	407,138
10	5,764,497	481,763
11	5,269,551	545,142
Total	32,866,739	5,584,124

Identification of global similarity between clusters

Despite the large number of singleton sequences, there may be a large number of similar or related sequences (similar primary or secondary structure) present within our data. Given that one of the hallmarks of homologous structured RNAs is that structure is often conserved when the sequence is quite variable [18], the high singleton frequency

TABLE 3.2: Percentage of rapid amplifier sequences in the SELEX sequence data separated by round.

Round	Percent rapid amplifier (%)
4	1.1
9	73.6
10	23.2
11	30.9

suggests the existence of a recognition motif in a common short sequence, and/or secondary structure. Due to the number of sequences, identification of common sequence or structure using pairwise comparisons is computationally prohibitive.

There are readily available programs that cluster based on sequence alone, such as CD-HIT [168], or cluster based on sequence and structure, such as RNAclust.pl + LocARNA [62]. Clustering using structure did not work as RNAclust.pl is designed to cluster $<$ 1000 sequences and LocARNA (and its derivatives LocARNA-P [63] and SPARSE [64]) are designed to simultaneously use sequence and structure to create multiple sequence alignments from homologous sequences, not the large and diverse set of sequences we obtained through SELEX. While CD-HIT only compares sequences, similar sequences are likely to fold into similar structure. Therefore, we used CD-HIT, which is a fast and widely-used program for nucleic acid clustering that utilizes heuristics to significantly reduce run time.

We established a clustering threshold by calculating the sequence similarity from the high frequency sequences. Examining the distribution of sequence distance shows a clear separation at 10%, which is equivalent to 90% similarity (Figure 3.3A). Clusters formed around the most frequent sequences are distinct, as seen by having lower within-cluster distance than between-cluster distance. This trend continues to be true for all high frequency sequences (Figure 3.3B). Because CD-HIT run time increases proportionally to the number of clusters, we use an 85% clustering threshold. However, to identify any global structure, we focus on those clusters with $>$ 90% similarity (Figure 3.3C).

Given the observed sequence diversity across our clusters, we also assessed whether any similar global secondary structures were shared between clusters. Clustering similar

TABLE 3.3: Clusters that have a mean structure distance less than the median intra-cluster distance of 0.0946 were also considered structurally similar.

Cluster1	Cluster1 size	Cluster2	Cluster2 size	Mean ensemble distance	belief
3543	253	21035	294	0.09324413	-1.744
5300	134	3543	253	0.09345742	-2.008
5300		72036	402	0.06441246	-2.698
5300		82519	911	0.05355445	-3.339
12222	209	82519	911	0.08549975	-2.031
72036	402	1290	601	0.08523696	-2.136
72036		21035	294	0.08567548	-1.920
82519	911	1290	601	0.08320683	-2.958
82519		2293	390	0.09279544	-2.823
82519		21035	294	0.08129743	-2.040

sequences together reduces the number of structure prediction operations because a representative cluster structure can be quickly determined by sampling and folding a small number of sequences (See Methods: Intra/inter-cluster ensemble distance). Using this method, we find that sequence clusters are also effective structure clusters because of the lower intra-cluster structure distance (median distance of 0.0898, Figure 3.4) compared to the inter-cluster structure distance (Figure 3.5). Additionally, pairwise comparisons of the clusters shows higher inter-cluster structure distance indicating there is no globally similar structure shared by any clusters. While some clusters appear to have similar structure (Figure 3.5B), upon closer inspection, this similarity is an artifact caused by comparing a limited number of structures from each cluster (See Method: Calculating belief for structure distances, Table 3.3, Figure 3.6).

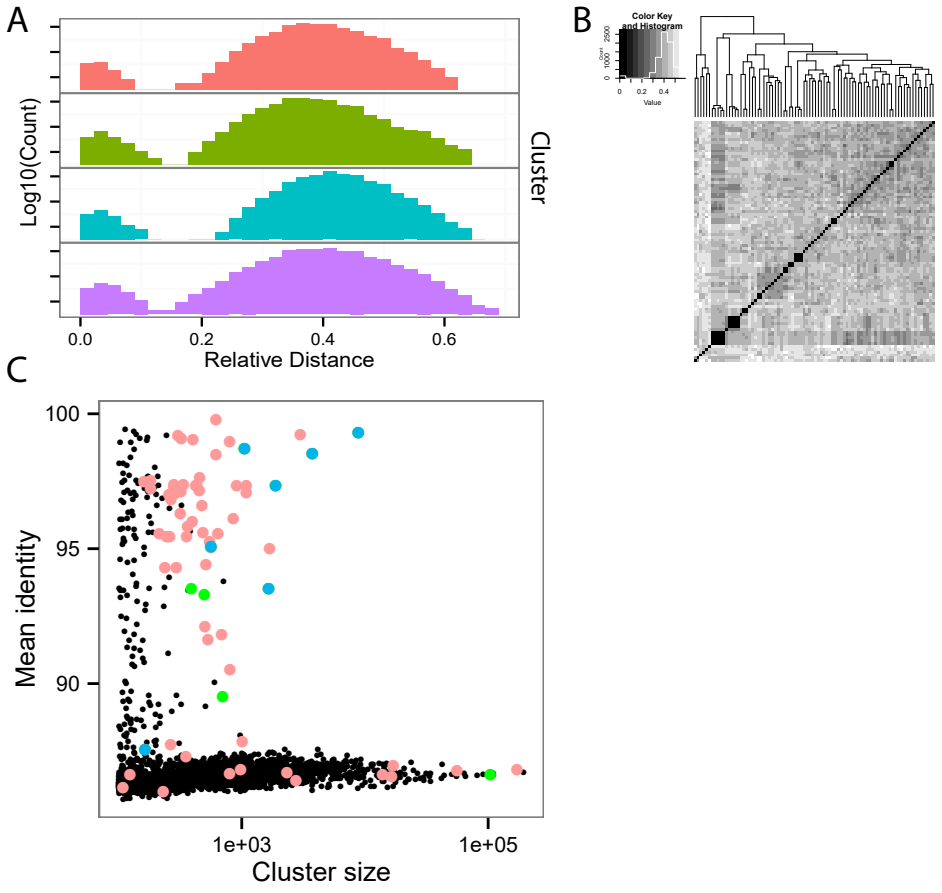

FIGURE 3.3: **A)** Distribution of Levenshtein distance around the top 4 multiton sequences shows a clear cluster cutoff at distance 0.1. Within the cluster, there is a decrease in the frequency of sequences further from the center indicating multiton clusters are valid. **B)** Heatmap of pairwise edit distance between the 84 multitons reveals very few multiton sequences can be grouped together (black). Majority of multitons are unrelated ($> 10\%$) to any other multiton sequence. Values are symmetrical across the diagonal. **C** Plot of the CD-HIT clustering data represented as cluster size vs mean percent identity to cluster seed (diffuseness). In red are the multiton clusters with more than 100 read counts. In blue are multiton clusters with more than 100 read counts, that have been experimentally examined (Table 3.12). In green are sequences experimentally tested that are not derived from the multiton clusters.

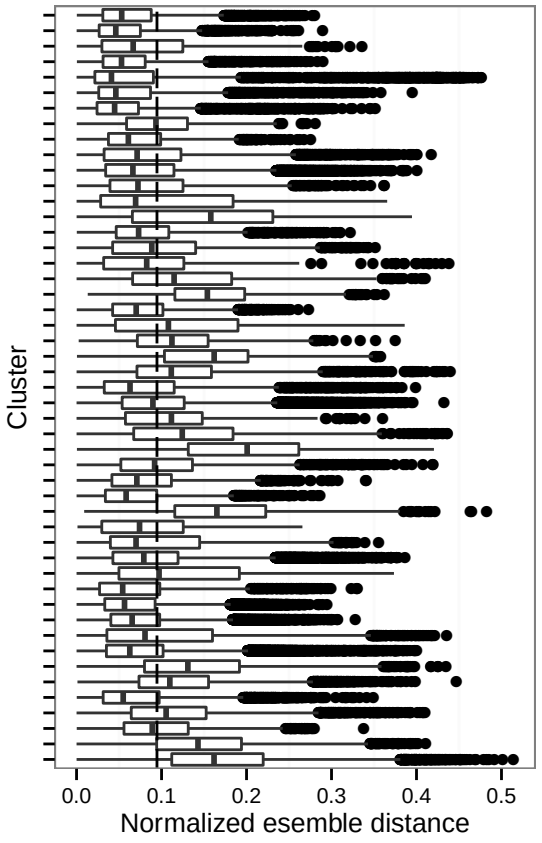

FIGURE 3.4: Distribution of intra-cluster ensemble distances by cluster. Box edges represent the first and third quartiles with the middle being the median. Clusters with the following criteria were selected: >100 sequences, and >90% mean identity to the seed. The line represents the median intra-cluster distance at 0.0898. The mean mean distance was 0.0946.

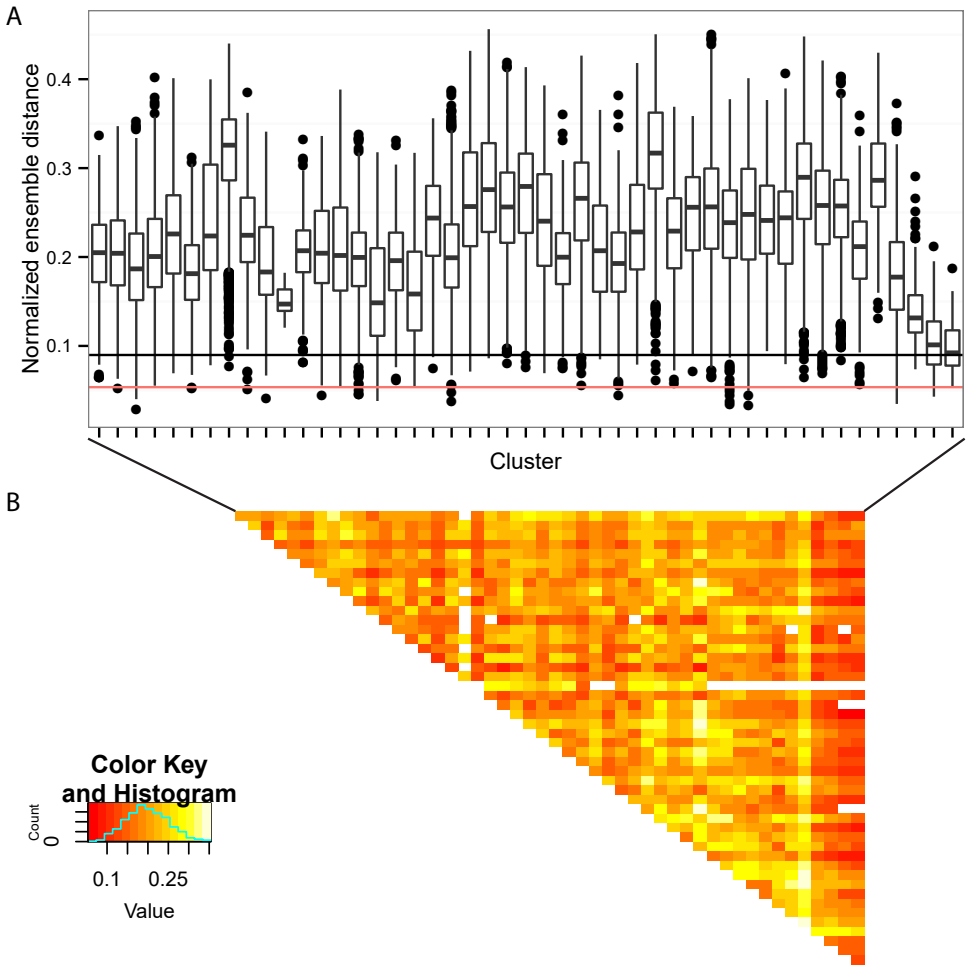

FIGURE 3.5: **A)** Distribution of inter-cluster ensemble distances from cluster 6062, which contains the most frequent sequence. Clusters selected for comparison included clusters with >100 distinct sequences, >90% mean identity to the seed. To get a distance distribution when comparing clusters to cluster 6062, individual sequences of the same length from the given cluster and cluster 6062 were compared in an all-against-all fashion. As a reference, the median intra-cluster distance for cluster 6062 was 0.0898 (black line) and the first-quartile was 0.0536 (red line). **B)** Representing all selected cluster pair-wise comparisons distance distributions in a heatmap shows that on average, clusters differ from other clusters by 0.2. In general, many of the structures are distinct from those of other cluster structures.

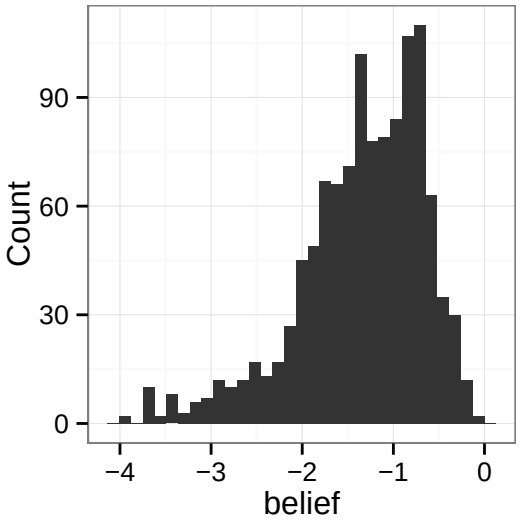

FIGURE 3.6: Distribution of belief in the evidence supporting our inter-cluster ensemble distance. The belief shows how much we believe in the evidence. Because certain clusters do not have enough pairwise comparisons, the average distance is somewhat biased to be low. The threshold for believable was set at the beginning of the left-tail, such that belief ≥ -2.

Identification of local sequence/structure motifs

Sequence

The high cluster count made it difficult to extract meaningful patterns from the sequences. In order to identify any common short sequence motifs, we started with sequence based approaches for motif identification because there are a variety of existing tools (summarized in Table 3.4). Many tools for motif identification are found in the MEME suite (MEME, GLAM2, DREME). These programs are not designed to process HTS data, so we reduced the data set by sampling 10^5 sequences from each of our SELEX rounds. While MEME is powerful and can identify transcription factor binding sites, in practice the algorithmic complexity limits the data to < 1000 sequences [145]. GLAM2 is able to identify gapped motifs and tolerates larger data sets, but it does not find any significant motifs (E-value = 1) in our data (Figure 3.7). We also applied DREME to find short k-mers ($3 \leq k \leq 8$), and some of the top motifs with more than 10^4 occurrences are significant (Table 3.5). They are repeatedly found in multiple samples; however, they are only found in 1.2%-5% of the total sequence pool.

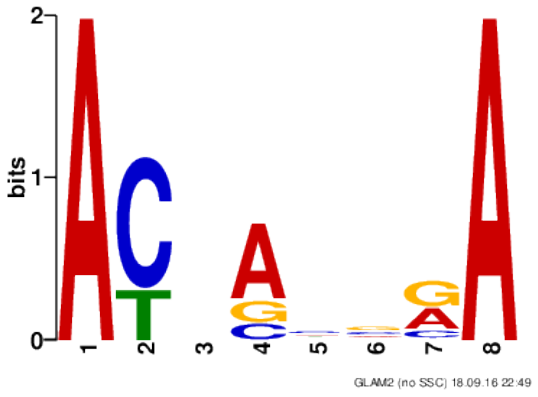

FIGURE 3.7: A logo representing the top motif from running GLAM2 over our sampled data set. This motif is unlikely to be the S15 recognition motif because of the low conservation of majority of the motif other than the two A's. Additionally, this motif is not significant (E-value=1).

Additionally, we applied other state of the art programs for identifying binding motifs in HT-SELEX data. Due to our lack of non-binder data, we could not use RNA-context or

TABLE 3.4: Comparison to existing tools

Software	Run Time	Motifs identified
MEME	N/A	N/A
DREME	≈3 hrs	top 10
GLAM2	≈1 week	top 10
AptaTrace	3+ weeks	N/A
AptaNI	N/A	N/A
NCM	≈10 hrs	N/A

TABLE 3.5: Top DREME motifs with $>10^4$ observations

Motif	E-Value	Percent sequences containing motif (%)
YACTGCT	2.4e-2784	1.2
WTAYGGA	5.6e-1525	1.5
WCCRAG	1.3e-515	5.0

Where R = A or G; Y = C or T; W = A or T;

RCK. We applied AptaNI, which searches loop regions for potential motifs, but the program did not find a single common motif in a large fraction of our data. We also applied AptaTrace, which identifies multiple motifs but the program did not finish, possibly due to the high sequence diversity of our data.

TABLE 3.6: K-mer (k=5) enrichment of round 11 variable region compared to a background set (bg).

K-mer	ratio enrichment 11 vs bg
AACCA	3.653
AACGA	3.557
ACCGA	3.403
ACCAA	3.302
AACAA	3.274
ACGGA	3.244

K-mer analysis is a common method for motif identification, which works by identifying enriched short sequences of length k (k-mers). Several programs implement this method by considering all k-mers (MPBind) or those restricted to loop regions (Aptacluster). MPBind compares sequences from different SELEX rounds with the expectation that short motifs increasing in frequency are being selected for [169]. We calculate k-mer enrichment of round 11 relative to a uniform probability background set because the

potential motifs may be well represented by SELEX round 4. We found that k-mers of length 5 are enriched compared to the background set (Table 3.6). Using the enriched k-mers (ACCGA, ACGGA, AACGA, AACCA, ACCAA), we created a regular expression (A[AC][CG][ACG]A) to search our data for sequences that have these k-mers. Sequences containing a combination of these k-mers were tested for binding to the *G. kaustophilus* S15 (Table 3.7). The number of appearances of the k-mers does not impact the binding affinity (4 appearances - 99.43 nM, 2 appearances - 123.74 nM, and 5 appearances - 77.5 nM) (Table 3.7). K-mers outside the variable regions were not counted because it is unlikely that the binding motifs are constant and already found in the initial sequence thus common to the entire sequence pool.

TABLE 3.7: Summary of experimentally tested sequences and their binding affinity. Sequences were chosen based on over-representation of k-mers of length 5.

Seq. Id	Cluster Id	K_d (nM)	Reason
46474	63331	99.43	4 appearances of motif
355069	1307	123.74	2 appearances of motif
279047	70316	77.5	5 appearances of motif

In order to determine if the structure context was important to the motif, we also calculated any k-contexts for enrichment relative to round 4. We obtain the structure context by sampling 1000 suboptimal structures and determining the most probable sequence of paired (p), unpaired (u), or loop (l) for a given k length subsequence. This technique readily identifies k-mers located in single stranded regions of the RNA structure. We use k-mers of length 4 because tetramers form stable loops in RNA secondary structure. The top 10 enriched k-contexts show very high enrichment when comparing SELEX round 11 to round 4 (Table 3.8). The high enrichment value is likely due to the sparsity of the k-context vector as k-mer enrichment alone on the data set only shows $\approx$ 2 fold enrichment (Table 3.9). We also established that the k-mer was not likely to be located in a single stranded region of the RNA (Table 3.10). Applying k-mer analysis with or without structure context did not reveal a specific k-mers above background (uniform and round 4), which suggests the protein is not interacting with the RNA in a sequence specific or loop region.

Table 3.8: Top 10 enriched K-context using k=4 comparing the variable regions of sampled sequences from rounds 11 and 4

K-context (kmer/context)	Enrichment
CTGC/uupp	41.684
ACTG/uuuu	30.512
TGCT/uuuu	25.634
CTGC/ulpp	25.095
TTCT/uuul	22.164
CGAC/plpl	22.062
TGCT/upuu	21.540
CTGC/uuuu	21.364
TGCT/uppu	20.916
CTGC/uupu	20.808

u=unpaired, p=paired, l=loop

Table 3.9: Top 10 enriched K-mers using k=4 comparing the variable regions of sampled sequences from rounds 11 and 4

K-context (kmer)	Enrichment
TGGA	2.275
TTTG	2.237
GGTG	2.187
TTGG	2.099
TGCT	2.005
GTTT	1.975
TGGT	1.836
ATGG	1.814
GGTT	1.779
TATG	1.778

Table 3.10: Top 10 enriched K-mers using k=4, using joint probability of the k-mer being in an upaired region comparing the variable regions of sampled sequences from rounds 11 and 4

K-context (k-mer)	Enrichment
TTCG	2.731
GTAT	2.503
TTTG	2.495
GAAG	2.218
GTAC	2.133
TTAG	2.108
TTGG	2.108
TTGA	2.067
TGCT	2.065
CGCA	2.040

Structure

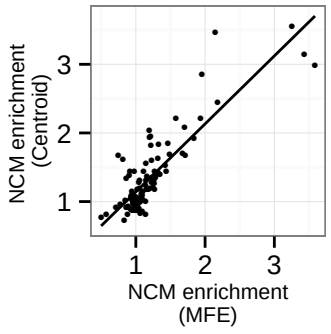

FIGURE 3.8: NCM enrichment as calculated using the minimum free energy (MFE) or the centroid structure. The calculated enrichment values show moderate correlation (r^2=0.771).

The lack of enriched sequence motifs and global secondary structure conservation indicates the binding occurs in a substructure of the selected RNA sequences. The existence of a substructure is further supported by data showing motifs identified by existing motif finders only account for a small fraction of the sequence pool. To identify potentially important substructures, we developed a novel approach that differs from existing methods by specifically focusing on stacking base-pairs. We represent stacked base-pairs as nucleotide cyclic motifs (NCM) [170]. This representation is advantageous because NCMs discretize the secondary structure into smaller components and they have been used to great effect in improving RNA tertiary structure predictions.

Since our approach depends on structure predictions, we calculate NCM enrichment using both the minimum free energy (MFE) and the centroid structure, which better represents the ensemble of structures. Both representations capture trends such increasing stability in later rounds. The NCM enrichment values derived from using the MFE structure or the centroid structure are moderately correlated (Figure 3.8). Using the centroid structure reduces the NCM frequency, but the reduced frequency has small impact on enrichment. Therefore, we carried out the remaining enrichment analysis using the MFE structure.

NCM enrichment is calculated by taking the ratio of the mean NCM frequency per round. To overcome the large number of sequences and differences in round size, we calculate the

mean enrichment by repeatedly sampling 10^5 sequences from each round. This sample size represents approximately 20% of rounds 9, 10, and 11, but only 2.5% of round 4. Despite the lower percentage of round 4 sequences sampled, the enrichment analysis is robust to the sampling and identifies similar enriched, depleted, and unchanged NCMs relative to round 4 (Figure 3.9A). GU/GU and UG/GU appear to be highly enriched and have larger error bars. However, these NCMs are not significantly greater than background, and the high variability is due to low frequency, thus these are considered spuriously enriched NCMs.

To identify significantly enriched NCMs, we also calculated the expected enrichment by comparing the NCM frequencies of the sampled sequences to background sequences, either created using uniform base frequencies (BG_{uni}) or base frequencies based on our total sequence pool (BG_{samp}) (See Methods: Background set construction). Our criteria for enrichment is that the NCM ratio of round 11 to round 4 must be significantly greater than the ratio of round 11 to background. Many NCMs are significantly enriched (AU/GU, AU/UG, CG/GC, CG/GU, GC/GU, GU/AU, GU/CG, GU/UA, UG/CG, UG/GC, UG/UG), while some are depleted (AU/UA, GC/GC, GC/UA) when compared against BG_{samp} (Figure 3.9B). There is significant overlap of enriched and depleted NCMs when comparing against BG_{uni} (Figure 3.10). Interestingly, many of the enriched motifs contain a GU wobble pair, which could be a potential recapitulation of the natural binding site.

The NCM enrichment in later rounds suggests selection for particular motifs. By treating clusters as "sequence families," we used LASSO logistic regression to identify NCMs associated with cluster enrichment. Since the analysis depends the clustering, we re-clustered our sequence pool multiple times and found the clustering is relatively stable (Figure 3.11). For each repeated clustering, we carried out LASSO regression and reduced our NCM predictors to those that appeared in majority of the models with p-value $<$ 0.01. Using this method on both round 4 to round 11, and round 4 to round 10, we identified positive predictors CG/GU and GU/GC as well as negative predictors AU/GC and CG/UA that are found in both models (Table 3.11). CG/GU was identified by enrichment analysis as well, further indicating its importance.

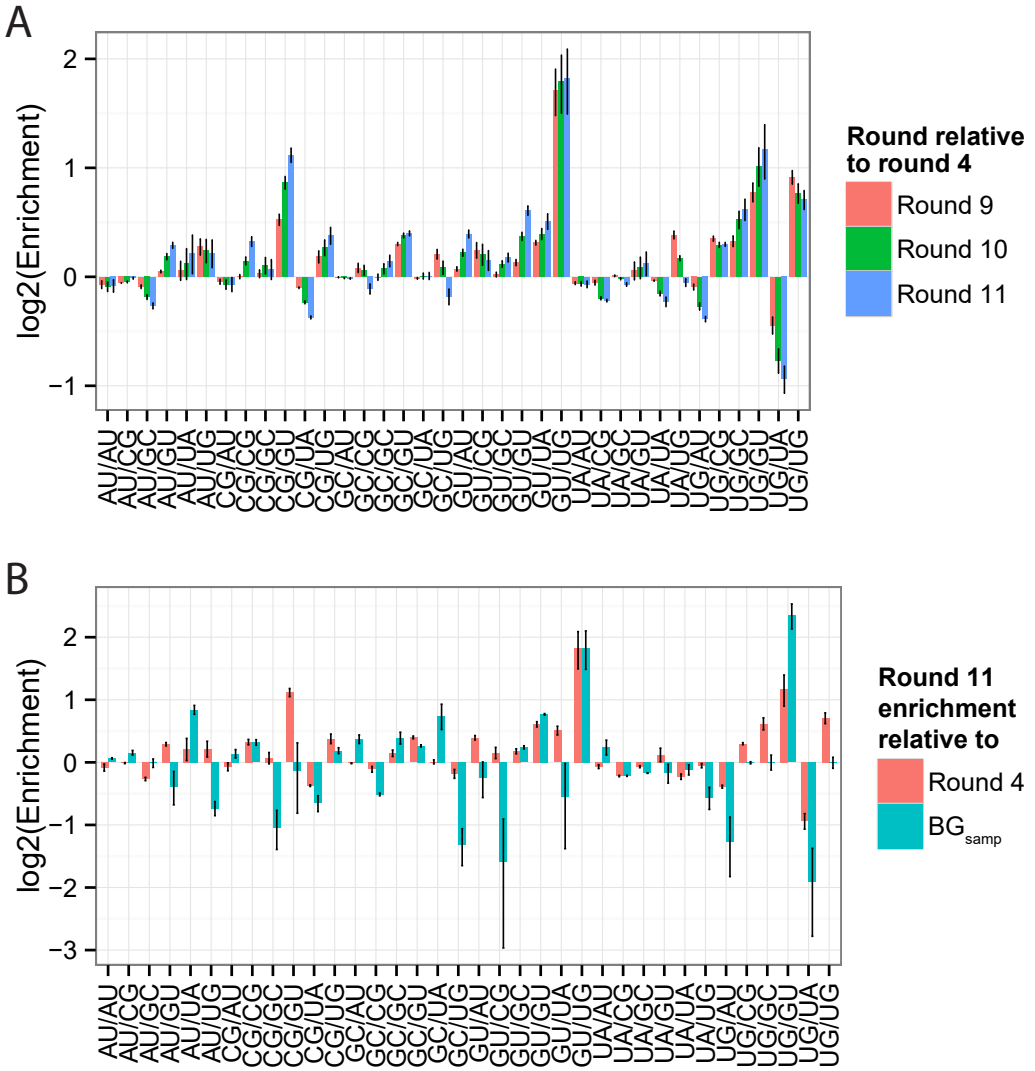

FIGURE 3.9: **A)** Log2 fold change of NCMs averaged over 11 re-samplings. The round 11 enrichment trends are consistent with the round 9 and round 10 enrichment. **B)** Log2 fold change of NCMs averaged over 11 re-samplings comparing the enrichment of round 11 vs. round 4 and round 11 vs. background created with sampled base frequency (BG_{samp}). Error bars represent standard error.

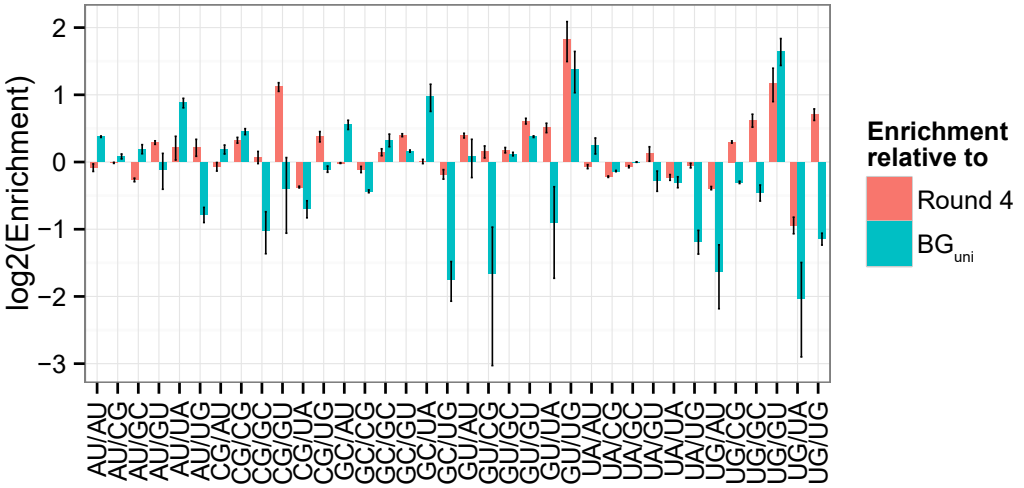

FIGURE 3.10: Log2 fold change of NCMs averaged over 11 re-samplings comparing the enrichment of round 11 vs. round 4 and round 11 vs. BG_{uni}. Error bars represent standard error.

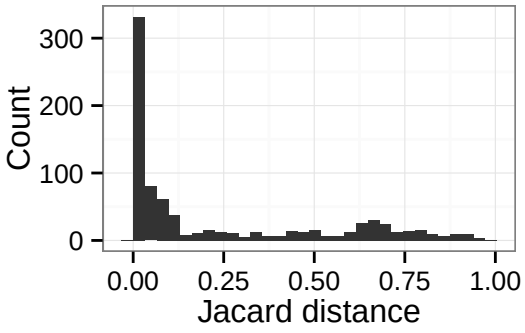

FIGURE 3.11: The CD-HIT clusters for high frequency sequences are relatively stable and many sequences often appear together despite multiple clustering runs. We show the distribution of the average the Jacard distance over 5 clustering runs. Occasionally, sequences are put into a different cluster, represented by high Jacard distance, but it is rare compared to being put into a similar cluster.

TABLE 3.11: Representative NCMs that are significantly associated with cluster enrichment from first clustering

Rounds compared	NCM	Log odds (95% CI)	P-value
11 to 4	AU/GC	-3.39 (-4.95 - -1.85)	1.87e-5
	CG/GU	8.04 (3.07 - 13.16)	1.75e-3
	CG/UA	-5.53 (-7.75 - -3.37)	6.86e-7
	GU/GC	4.66 (2.08 - 7.27)	4.36e-4
10 to 4	AU/GC	-1.89 (-3.37 - -0.44)	0.0112
	CG/GC	3.11 (0.523 - 5.74)	0.0194
	CG/GU	9.45 (4.08 - 15.00)	6.78e-4
	CG/UA	-3.18 (-5.36 - -1.13)	2.47e-3
	GU/GC	4.51 (1.98 - 7.10)	5.61e-4
	UA/UA	-5.96 (-8.67 - -3.31)	1.31e-5

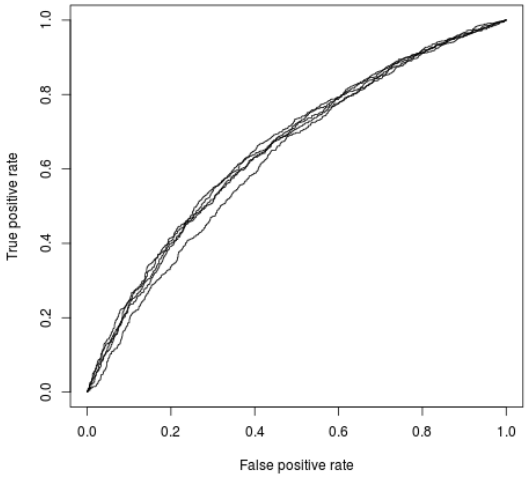

FIGURE 3.12: Receiver operator characteristic (ROC) curves showing the model performance on classifying clusters as enriched for later round sequences. The LASSO logistic regression model applied to each CD-HIT clustering run shows similar classifier performance across all runs. The mean area under the curve (AUC) is 0.651.

Given the overlap of predictors, we tested whether the logistic regression model for round 10 enrichment could predict future cluster enrichment (i.e. round 11 enrichment). Ideally, the same NCMs are selected throughout the SELEX process. After training on round 10 enrichment data, we tested the model by using cluster enrichment from each of the re-clustered data sets. However, this model offers a limited prediction accuracy (mean AUC=0.651), indicating some predictors are not readily identified (Figure 3.12).

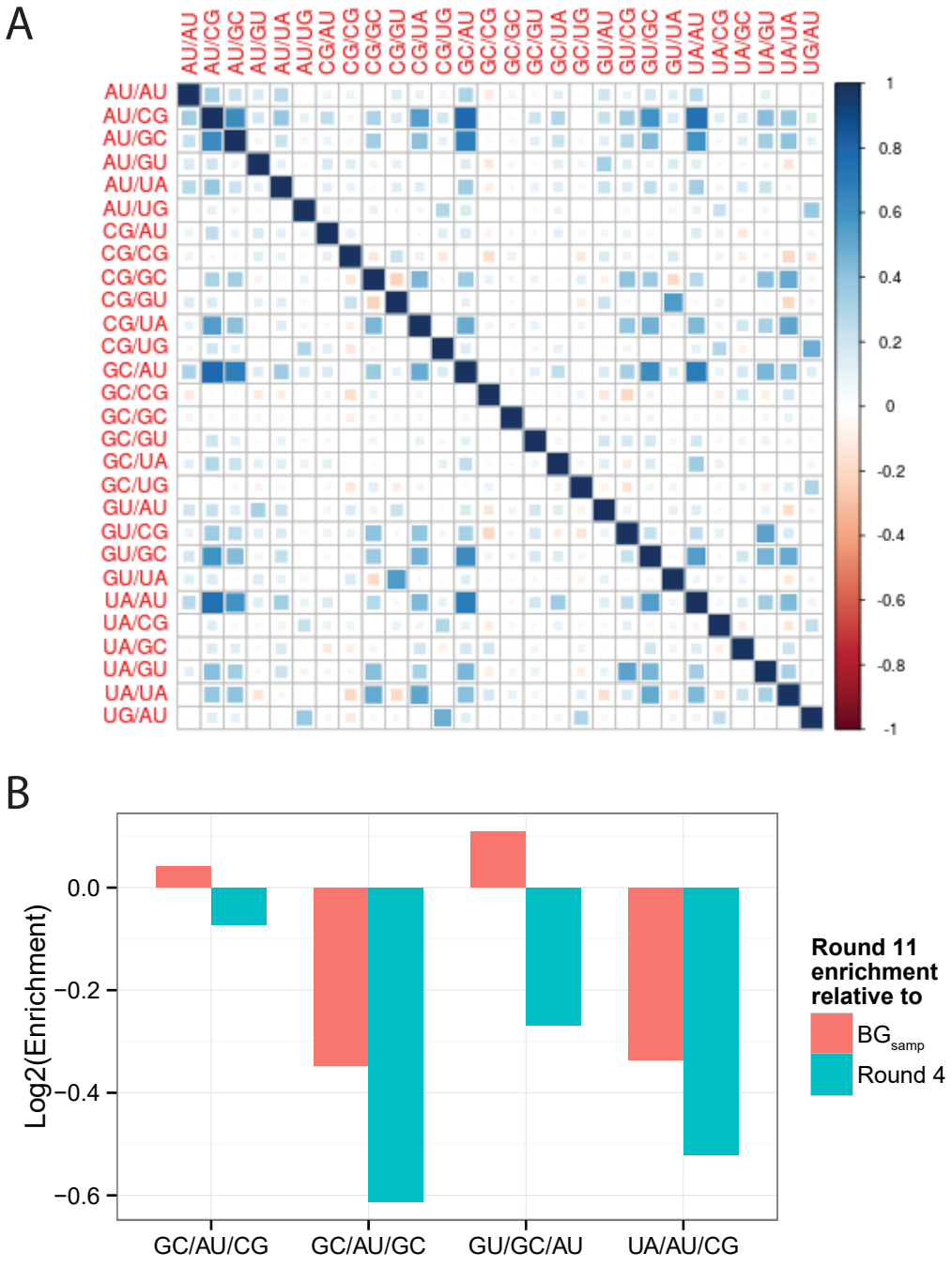

FIGURE 3.13: **A)**Spearman correlation matrix between all 2_2 NCMs with greater than 10k counts per round. The matrix is symmetrical around the diagonal. The larger and darker squares indicate stronger correlation. Positive correlation indicates the NCM pair could be combined as part of a larger binding motif. The correlation between GTTA and CGGT is 0.647. The correlation between TATA and CGTA is 0.522. **B)** The enrichment/depletion of 3_3 NCMs as compared to round 4 or BG_{samp}. These NCMs are composed of 2_2 NCMs that often appear together suggesting a potentially larger motif. The ratio suggests these larger motifs are more depleted in later rounds than expected.

In order to ensure the 2_2 NCM was not part of a larger base-pair stack, we used Spearman correlation to identify any NCMs that often appear with each other. There is moderate correlation between some NCMs ($\rho > 0.6$) (Figure 3.13A). However, this correlation is most likely spurious because repeated analysis with 3_3 NCMs does not show higher enrichment of these NCMs relative to BG_{samp} (Figure 3.13B).

Experimental assessment of S15 binding affinity

TABLE 3.12: Summary of experimentally tested sequences and their binding affinity

	Seq. Id	Cluster Id	K_d (nM)	Reason
A	98	52739	85	High freq.; High mean pairwise identity ($> 90\%$)
B	101	6062	42	Most freq.; High mean pairwise identity ($> 90\%$)
C	575	2903	62	High freq.; High mean pairwise identity ($> 90\%$)
D	669	1792	25	High freq.; High mean pairwise identity ($> 90\%$)
E	4778	851	19	High freq.; High mean pairwise identity ($> 90\%$)
F	27773	517	2.8	High freq.; High mean pairwise identity ($> 90\%$)
G	46474	63331	99	Singleton; Small cluster (≤ 100 seqs.)
H	355069	1307	123	Singleton; Small cluster (≤ 100 seqs.)
I	244064	4454	62	Singleton; Medium cluster ($100 <$ seqs. < 1000)
J	158254	91212	31	Singleton; Singleton cluster ($= 1$ seq.)
K	279047	70316	77	Singleton; Singleton cluster ($= 1$ seq.)
L	4077	68	9.8	Singleton; Large cluster (≥ 1000 seqs.); Low mean pairwise identity cluster ($< 90\%$)
M	170365	2293	Non-specific	Singleton; Low mean pairwise identity cluster ($< 90\%$)
N	192209	3606	Non-specific	Singleton; Low mean pairwise identity cluster ($< 90\%$)
O	4650	3969	38	Depleted; Medium cluster ($100 <$ seqs. < 1000); low pairwise identity cluster ($< 90\%$)
P	315173	5799	28	Depleted; Previously identified regulator [167]

In order to ensure our SELEX data provided an accurate reflection of binding sequences, we assayed a variety of sequences for binding affinity for S15 (Summarized in Table 3.12). We find many high frequency sequences had moderate affinity for S15 ranging from 19-85.6 nM (Table 3.12 A-F). Given the high diversity of the sequence pool, we also tested singleton sequences for binding, which revealed 6 of 8 singleton sequences also bind S15 (Table 3.12 G-N). Previous literature suggests that enrichment is a better predictor or binding affinity [136]. We find that there is no correlation between the degree of enrichment and binding affinity (Figure 3.14). Both depleted sequences bind S15 with moderate affinity (Table 3.12 O-P).

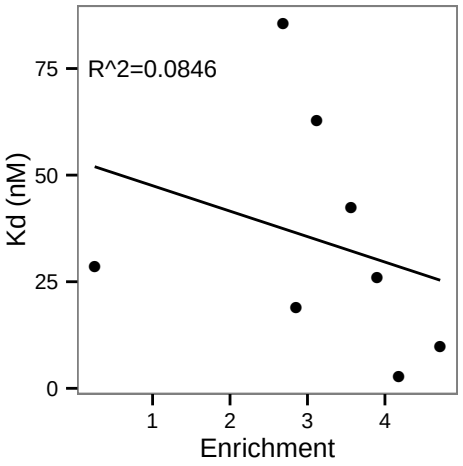

FIGURE 3.14: Linear regression showed that the sequence enrichment did not predict the K_d. The regression line has an R^2 value of 0.0846.

We also tested sequences from clusters that are centered on high frequency sequences. When a sequence represents a large fraction of the cluster, we hypothesize that this sequence binds while the remaining sequences "explore" the local sequence space. Fitting with our hypothesis, many high frequency sequences specifically bind S15 and are found in high mean pairwise identity cluster (Table 3.12 A-F). As a control, sequences from clusters with low mean pairwise identity not centered on high frequency sequences were also examined (Table 3.12 L-O). We find half of these sequences bind specifically, which suggests high identity clusters are more likely to contain S15 binders.

We use the enriched/depleted NCMs with our experimental data to build a model to identify potential binders (See Methods: Classifying S15 binders using the NCM model).

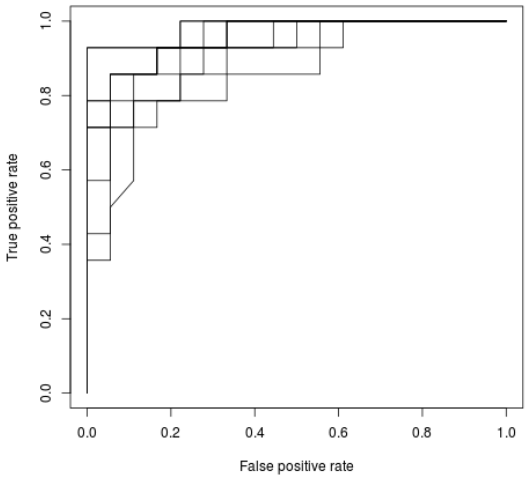

FIGURE 3.15: The logistic regression model using significantly enriched/depleted NCMs as predictors is applied to multiple re-sampled data sets. Each data set is composed of the experimentally tested sequences and non-binder sequences from BG_{samp}. Receiver operator characteristic (ROC) curves showing the model performance on classifying sequences as either "binders" or "non-specific binders". The mean area under the curve (AUC) is 0.921.

Due to the limited number of negative test cases, we use additional sequences from our background set to build a logistic regression model. The model suggests using enriched and depleted NCMs are good predictors of binding (mean AUC = 0.921) (Figure 3.15)

3.3 DISCUSSION

The RNA binding sites of many proteins are complex in terms of both sequence and structure. In this work we sought to understand the pool of potential RNA-binding sites for *G. kaustophilus* ribosomal protein S15 using *in vitro* selection coupled with high-throughput sequencing (HT-SELEX). To our surprise, the high-throughput sequencing revealed an extraordinarily large pool of potential binding sites with over 95.3% of our sequences appearing only once in the population. We were able to cluster our data using a number of different methods. However, the large number of unique clusters did not share any obvious global structure, or sequence characteristics. Existing strategies that have been applied to the analysis of other RBPs were unsuccessful at identifying any

features that would explain a significant portion of our data. Many programs are not designed for the number or diversity of our sequence data.

We developed a novel approach to analyzing HT-SELEX data for motifs that incorporate RNA structures. Our approach borrows from three-dimensional structure prediction [170], by considering all potential substructures or nucleotide cyclic motifs (NCMs) of a certain length. This approach is further necessitated by the complexity of the known RNA binding sites for S15 [160, 163, 164]. We repeatedly sampled sequences from each round to carry out our analysis. There are many enriched or depleted 2_2 NCMs relative to earlier rounds, with many of the enriched NCMs containing a GU wobble base-pair, which could be a potential recapitulation of the natural binding motif. By using LASSO regression, we effectively reduced the number of NCMs to potential predictors of enrichment.

Our algorithm is easily parallelized and the run time is increased proportionally to the number of secondary structure predictions. The run time falls on the shorter end of the spectrum compared to existing software, which can sometimes take a week to finish. We have demonstrated that the algorithm is robust to structure representation. Additionally, the NCM data is easily integrated into models to predict potential binders. Despite a limited number of validated binders and non-binders, the model accurately distinguish binders from background sequences. Surprisingly, our limited model classifies only 15.7% of the total sequence pool as potential S15 binders, suggesting many potential non-binders. Considering the proportion of binders found within our limited population of verified binder sequences, it appears that only a subpopulation of binding sequences can be identified using NCMs alone and that S15 likely can recognize additional features that are not captured by this data.

3.4 CONCLUSION

Our analysis of the HT-SELEX data for the *G. kaustophilus* S15 suggests that this protein can bind a large diversity of sequences *in vitro* and our previous work demonstrated that half of the RNAs examined allowed regulation [167]. The analysis also suggests

that the recognition motif is located in a combination of structure elements with little requirement on the sequence itself. This finding also illuminates a possible reason for the large sequence and structure diversity in natural S15 mRNA secondary structures. The approach we developed to analyze our data is broadly applicable to many other RBPs that have complex noncontiguous recognition motifs. By considering RNA secondary structure elements as building blocks (NCMs), we bring a novel approach to analyzing *in vitro* selection data for RNA-protein interactions that may primarily rely on specific local features in the context of a larger secondary structure.

3.5 Material and Methods

High-throughput sequencing

We previously identified S15 binders using 11 rounds of SELEX [167]. Expecting less diversity, we initially sequenced cDNA pools resulting from reverse transcription of the selected sequence pools after rounds 4 and 9. But after a brief data analysis, we further sequenced rounds 10, and 11 as they were the final rounds of selection. The sequence pools were sequenced using Illumina short read 100 nucleotide (nt) paired-end sequencing (Otogenetics Corporation). The expected length of the aptamer was 87 nt, composed of 30 nt PCR primers, 30 nt variable region, and 27 nt non-constant region.

Sequences were filtered by having the correct primers, standard nucleotides (A,C,G,T), forward strand, and every nucleotide's PHRED quality score of greater than or equal to 20. Any sequences shorter than 79 nt or containing duplicated T7 promoter sequence were removed. These sequences are considered rapid amplifier sequences because they only contain T7, 5', and 3' sequences (See Methods: Rapid amplifiers).

The libraries are stored in separate FASTQ files for each round. The remaining sequences were stored in a MySQL database for speed and ease of access. For subsequent analysis, only the sequence contained between and including perfect primers was used. When calculating enrichment, the sequence counts were normalized to the total number of usable reads in that round.

Rapid amplifier

We remove short sequences of 79 nt that are only composed of the T7 promoter, 5' primer, and 3' primer. These sequences do not contain a variable region and appear to rapidly amplify during the PCR amplification step. Experimental results show these sequences do not specifically bind to the *G. kaustophilus* S15; therefore, we remove them from our sequence pool during analysis.

Clustering

Sequence

In order to determine a cluster threshold, sequences from rounds 10 and 11 with > 100 total counts were used as initial cluster centroids to compare to the remaining sequences. The distance metric (Levenshtein distance) was normalized to the length of the longer sequence. As an optimization, the primer regions were removed for the purposes of sequence comparison.

CD-HIT-est [168] was used for nucleotide clustering with the following options: compare positive strand only (-r 0), mismatch penalty -1, gap penalty -1, gap extension 0 and cluster threshold of 85% (-c 0.85). The mismatch penalty and gap open penalty are both the same value to minimize the effect of single base variation or deletions in the variable region. The gap extension is set to 0 because it heavily penalized short stretches of base differences in the variable region thus creating many more singleton clusters. Only the non-primer regions were compared. The output from CD-HIT was imported into a MySQL database for speed and ease of access.

Structure

RNAclust.pl + LocARNA will cluster sequences based on sequence and structure. We used the default parameters, 8 CPU threads and "–sparse" for the LocARNA option.

Intra/inter-cluster ensemble distance

The clusters used for analysis were selected from the CD-HIT clusters using the following criteria: $>$ 100 sequences and $>$ 90% mean identity to the CD-HIT cluster seed. Secondary structure prediction was done using the Vienna RNAfold package [31]. The ensemble distance was calculated by first predicting the secondary structure ensemble using 'RNAfold -p'. The ensemble distance is the mean base-pair distance between all possible structures of two input sequences [101]:

$$\frac{1}{|A|} \sum_{(i,j) \in A \cup B} (P_{ij}^{A} - P_{ij}^{B})^2 \tag{3.1}$$

where $i < j$ and P_{ij} is the probability of a nucleotide at position i paired to a nucleotide at position j and $|A|$ is the length of structure A. Structures A and B must be the same length.

Intra-cluster distance was calculated by taking 1,000 (or fewer) distinct sequences from each of the clusters meeting our criteria. Then ensemble distance was calculated in a pairwise fashion.

Inter-cluster distance was calculated using the top 100 most frequent sequences from each cluster. Structures in each cluster were compared in a pairwise manner to structures in the other cluster.

Calculating belief for structure distances

Since only structures of the same length are comparable using ensemble distance, To identify cluster pairs impacted by this artifact, we calculated the ratio of actual comparisons versus the number of possible comparisons (Figure 3.6). If the ratio was less than 0.01, the ensemble distance was ignored because it was likely calculated using only one sequence per cluster. This additional screen removed all clusters that appear to be similar.

$$belief = \log \frac{count}{min(10000, (cid1size * cid2size))} \tag{3.2}$$

where *count* is the number of pairwise comparisons between 2 clusters (cid1, cid2) and $cid1size$ is the size of cluster1 and $cid2size$ is the size of cluster2.

Sequence motif identification

We applied a variety of existing motif finder programs to our sequence pool: DREME, GLAM2, AptaNI, and AptaTrace. For all programs, we used the same sample, which is created by sampling 10^5 sequences from each round of selection. The parameters for DREME were motifs of length k such that $3 \leq k \leq 8$, no reverse complement, and stop after the top 10 motifs are identified. GLAM2 parameters: motifs of length k such that $3 \leq k \leq 8$, and 50000 iterations. AptaNI was run with the default parameters. AptaTrace was run with default parameters, using SFold [110] as the RNA folding program.

K-mer and k-context

K-mers represent all possible subsequences of length k. Given a sequence set, k-mers counts are normalized to the sequence length - 1 and the size of the sequence set. The k-mer counts from the primer regions are excluded.

The k-contexts represent all possible tuples of subsequence and substructure of length k. To minimize the effects of inaccurate base-pairs, we sample 1000 suboptimal structures to estimate the probability the nucleotide's context is paired, unpaired, or in a loop. Using the context probability, the resulting k length substructure is the most likely sequence of contexts.

We also calculate the k-mer enrichment of k-contexts only located in unpaired or loop positions. This calculation differs slightly from the k-context in that the nucleotide context of a position i is not independent of its neighbors. Therefore, the joint probability of k nucleotides being unpaired is calculated using "RNAplfold". K-mers are counted only if every position in the context is likely to be unpaired (i.e Pr(unpaired) > 0.5).

Identifying enriched/depleted secondary structure motifs

The structural motifs we identify are derived from the 2_2 and 3_3 nucleotide cyclic motifs (NCM) [170]. We modified the naming convention to be more base-pair centric — N1_N2 <sequence> such that the N1 and N2 designate the length of the 5' and 3' strands, respectively. The <sequence> represents the order of stacking base-pairs starting at the 5' end.

To detect NCM enrichment, NCMs are counted by sampling 10^5 distinct sequences without replacement from each round. For each sequence, the MFE or centroid structure is predicted using Vienna RNAfold [31] and each possible 2_2 or 3_3 NCM stack is counted. Similar to calculating k-mer frequency, NCM frequency is calculated by normalizing the NCM count to the total number of NCMs per sequence and number of sequences sampled. NCM enrichment/depletion is calculated by the ratio of the average NCM frequency between rounds.

In order to identify enriched NCMs, we repeatedly calculate NCM enrichment relative to both round 4 and a background set. The NCM enrichment relative to background provides an "expected" baseline enrichment value. NCMs are considered significantly enriched when the average NCM enrichment relative to round 4 is higher than average expected NCM enrichment (p-value < 0.001). Significance is calculated using a one-sided T-test [171].

Background set construction

The background sequence set variable region was created using either a uniform (BG_{uni}) or a sampled base distribution (BG_{samp}). The sampled base frequency is determined using the variable regions from the sequence pool. The variable region was identified by minimizing the Levenshtein distance between our known non-constant region sequence (TCATTCTATATACTTTGGAGTTTTAAA) and a sliding window of length 20 along the given input sequence.

TABLE 3.13: Number of enriched clusters from each clustering run.

Cluster run	Depleted	Enriched	Total
1	1140	2079	3219
2	1151	2073	3224
3	1140	2083	3223
4	1152	2051	3203
5	1162	2062	3224

Any mutations to non-variable and non-primer regions were simulated using the "mutation rate" derived from the non-constant region of round 11 sequences. The mutations were categorized as point mutation, insertion, or deletion. The synthetic constant region was simulated by choosing the site(s), which is governed by the Poisson distribution, and type(s) of mutation in a sequence based on the overall mutation frequency. Then the resulting mutation is selected based on the observed mutational frequency. The synthetic sequence was generated by concatenating the primers, a simulated variable region (30 bases chosen with uniform or observed probability) and a simulated non-constant region in the proper order.

LASSO Logistic regression models

Logistic regressions and LASSO were done in the R project [171]. Only clusters with $>$ 100 sequences were used, as these clusters are likely to contain sequences from different rounds. Clusters are considered enriched if the ratio of sequence frequency from later to earlier rounds were real numbers and exceeded a certain threshold. This threshold is determined by calculating ratio of total round counts (i.e. $\frac{\text{round 4 size}}{\text{round 11 size}}$). For round 11 (r11) sequences to be considered enriched, the ratios r11:r4 $>$ 7.61 or r11:r9 $>$ 0.7468. For round 10 (r10) sequences to be considered enriched, the ratios r10:r4 $>$ 8.61 or r10:r9 $>$ 0.8451. For the training set, a 1:1 ratio of enriched vs depleted clusters were used. The number of enriched and depleted clusters for each re-clustering is summarized in Table 3.13.

We re-cluster multiple times using CD-HIT. For each CD-HIT re-clustering, NCM predictors are selected automatically by LASSO logistic regression. Predictors are retained if they appear in 3 out of 5 re-clusters with a significant p-value $<$ 0.01.

Classifying S15 binders using the NCM model

When predicting potential binders, we use our experimentally validated sequences as positives (14 positive, 2 negatives) and 16 sampled sequences from BG_{samp} as negatives. For each sequence, we calculated the NCM frequency. The model uses the established enriched/depleted NCMs as predictors (AU/GU, AU/UG, CG/GC, CG/GU, GU/AU, GU/CG, GU/UA, UG/CG, UG/GC, GC/GC, GC/UA). Some NCMs are removed because of singularities. Since our data set is small, we re-sampled background sequences and average the performance over multiple samples.

RNA/Protein preparation

The aptamer sequence was synthesized using assembly PCR from overlapping oligos (from IDT) with the T7-promoter sequence added within the forward primer sequence. T7 RNA polymerase [172] was used to transcribe RNA and transcription reactions were purified by 6% denaturing PAGE. Bands were visualized using UV shadow, excised, and the RNA eluted (in 200 mM NaCl, 1 mM EDTA ph 8, 10 mM Tris-HCl pH 7.5) and ethanol precipitated. Purified RNA (10 pmol) was 5'-labeled with 32P-ATP and purified as previously described [173]. Protein expression and purification was conducted as described previously [161].

Filter binding assay

As done in Slinger *et. al* 2015, a fixed amount of 5'^{32}P-labeled RNA (1000 cpm, <1 nM) was renatured for 15 minutes at 42°C, then incubated with serial dilution of *G. kaustophilus* S15 in Buffer A (50 mM-Tris/Acetate, pH 7.5, 20 mM Mg-acetate, 270 mM KCl, 5 mM dithiothreitol, 0.02% bovine serum albumin[63]) for 30 minutes at 25°C. Nitrocellulose membrane (GE Healthcare) was used to collect RNA-S15 complexes and positively charged nylon membrane (GE Healthcare) was used to collect unbound RNA under suction in a filter binding apparatus. Membranes were air-dried 5 minutes and the fraction bound quantified by imaging membranes on a phosphorimager screen. Radioactivity counts per sample on each membrane were measured using GE Healthcare

STORM 820 phosphorimager and ImageQuant. For each sample the fraction bound (Fb) corresponds to

$$Fb = \frac{\text{counts nitrocellulose}}{\text{counts nitrocellulose} + \text{counts nylon}} \tag{3.3}$$

Since Fb is known, to determine the K_d and the Hill coefficient (n), the resulting values were fit to the equation:

$$Fb = Min\% + \frac{Max\% - Min\%}{1 + (\frac{K_d}{[S15]})^n} \tag{3.4}$$

where [S15] corresponds to the concentration of S15 in the reaction and $Min\%$ and $Max\%$ correspond to the minimum and maximum fraction bound, respectively. The residuals were minimized using the nonlinear least squares estimate (nls) in R to find both the Hill coefficient and the K_d.

Availability of data and materials

The code for the analysis is publicly available on github https://github.com/ship561/hts-exploration. The HT-SELEX data is publicly available on the SRA with the ID: SRP077756.

Chapter 4

Discussion

Identifying *de novo* RNA structures in sequence data is an ongoing challenge. The difficulties center around poor alignment quality and our inability to distinguish functional elements within the global structure. The structure conservation that is the hallmark of many conserved ncRNAs is non-trivial to identify. Aligning based on sequence alone causes base-pairing information to be lost, and aligning based on structure alone requires a very high degree of structural similarity. Therefore, poor alignment quality directly limits our ability to distinguish the functional local structure from the global structure. For example, a functional minimal structure could be difficult to align to its homologous ncRNA because it can have low global structure similarity but high local structure similarity. This problem highlights the difficulty of determining conserved structure and thus novel ncRNA structure from limited data.

This thesis approaches the challenge of identifying functional RNA structures from opposite sides: *de novo* detection of natural RNAs, and comparison of synthetic RNAs from HTS of *in vitro* selected populations. Despite the difference in sequence source, a common theme to my projects is the development of novel methods to emphasize the functional portions of the structure. In chapter 2, we develop a novel measure of neutrality that emphasizes maintaining the existing structure determined from alignments of known ncRNA structures. Here, the hypothesis is driven by the evolutionary theory that suggests evolved ncRNAs maintain their structure in the face of genomic change thus

making them mutationally robust. In chapter 3, we develop a method for identifying enriched substructures found in an *in vitro* selected sequence pool. Our data contain high structural diversity of binders. Therefore, we hypothesize that the RNA-protein interaction occurs in a substructure or subsequence, and there can be multiple conserved motifs that contribute to binding.

Described in chapter 2, the SEN project focuses on using neutrality as a novel feature in RNA classifiers. As a solo feature, SEN has similar performance to existing RNA classifiers. Surprisingly, inclusion of SEN into the feature set only marginally improved performance. Our model performance data indicate that neutrality is a generalization of the existing feature set where features such as structure stability or co-varying mutations are specific properties of structures selected for during RNA evolution. Additionally, we determined that many naturally occurring structured RNAs are mutationally robust.

There are limitations to using SEN calculated neutrality as a feature. A fundamental assumption made in the neutrality calculation is that changes to the structure are deleterious. This assumption leads SEN to underestimate some mutations to loops. This limitation can be problematic because there are many structured RNAs that interact with their ligand in loops and any mutation to the motif render the RNA non-functional. Despite the importance of loop interactions, we found that mutations in loop regions are generally more neutral than mutations found in base-pairs. Another limitation is the runtime for calculating SEN is proportional to sequence length and the number of sampled structures. The increase in the number of structure comparisons makes neutrality much slower than calculating features such as MFE and mutual information of stems.

Described in chapter 3, the SELEX analysis project focuses on elucidating potential RNA binding motifs responsible for the RNA-S15 interaction. Our analysis of the HTS-SELEX data showed high sequence and structure diversity leading us to search for common subsequence and substructure. This data analysis highlights the diversity of RNA-protein interactions and that not all RNA-protein interactions occur in loop regions. We designed an algorithm that treats a structure as a set of small building blocks that may be treated analogously to k-mers for sequence analysis. Our approach considers both

sequence and structure components in the binding motif. Identifying enriched substructures showed there were common, positively selected motifs, possibly involved in RNA-protein interaction. Because the algorithm considers base-pairs, it also considers long range non-contiguous interactions that are not normally considered in sequence motif identification algorithms.

The limitations of using our NCM approach are similar to the limitations of k-mer analysis. In order to determine enrichment, we need a proper background sequence set to compare our sequence pool against. Determining the size of the motif is done through trial and error. In addition to the k-mer limitations, our algorithm has a slower runtime because we must fold the RNA sequences.

Bibliography

[1] S R Eddy. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, 2:919–929, 2001. ISSN 1471-0056.

[2] Wade C Winkler and Ronald R Breaker. Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.*, 59:487–517, 2005.

[3] Jane N Kim and Ronald R Breaker. Purine sensing by riboswitches. *Biology of the Cell*, 100(1):1–11, 2008.

[4] Stefanie A Mortimer, Mary Anne Kidwell, and Jennifer A Doudna. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.*, 15(7): 469–479, 2014.

[5] Thomas R. Cech and Joan A. Steitz. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell*, 157(1):77–94, March 2014. ISSN 00928674.

[6] Matthew Halvorsen, Joshua S Martin, Sam Broadaway, and Alain Laederach. Disease-Associated Mutations That Alter the RNA Structural Ensemble. *PLoS Genetics*, 6(8):11, August 2010. ISSN 1553-7404.

[7] Lynge C Christiansen, Simon Schou, Per Nygaard, and Hans H Saxild. Xanthine metabolism in Bacillus subtilis: characterization of the xpt-pbuX operon and evidence for purine-and nitrogen-controlled expression of genes involved in xanthine salvage and catabolism. *J. Bacteriol.*, 179(8):2540–2550, 1997.

[8] Gisela Storz, Jörg Vogel, and Karen M Wassarman. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell*, 43(6):880–891, 2011.

[9] Minju Ha and V Narry Kim. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, 15(8):509–524, 2014.

[10] Stinus Lindgreen, Sinan Uğur Umu, Alicia Sook-Wei Lai, Hisham Eldai, Wenting Liu, Stephanie McGimpsey, Nicole E Wheeler, Patrick J Biggs, Nick R Thomson, Lars Barquist, et al. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.*, 10(10): e1003907, 2014.

[11] P Schuster, W Fontana, P F Stadler, and I L Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. [Biol]*, 255(1344):279–284, 1994. ISSN 0962-8452.

[12] Raheleh Salari, Chava Kimchi-Sarfaty, Michael M Gottesman, and Teresa M Przytycka. Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. *Nucleic Acids Res.*, 41(1): 44–53, 2013.

[13] R R Gutell, N Larsen, and C R Woese. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews*, 58: 10–26, 1994.

[14] G E Fox and C R Woese. The architecture of 5S rRNA and its relation to function. *J. Mol. Evol.*, 6:61–76, 1975.

[15] G M Gongadze. 5S rRNA and ribosome. *Biochemistry*, 76:1450–64, 2011.

[16] R R Gutell, A Power, G Z Hertz, E J Putz, and G D Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, 20(21): 5785–95, November 1992. ISSN 0305-1048.

[17] J Parsch, J M Braverman, and W Stephan. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics*, 154(2):909–21, February 2000. ISSN 0016-6731.

[18] Jan Gorodkin, Ivo L Hofacker, Elfar Torarinsson, Zizhen Yao, Jakob H Havgaard, and Walter L Ruzzo. De novo prediction of structured RNAs from genomic sequences. *Trends in Biotechnology*, 28(1):9–19, 2010.

[19] A G Vitreschak, D A Rodionov, A A Mironov, and M S Gelfand. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends in Genetics*, 20:44–50, 2004.

[20] RW W Siegel, AB B Banta, ES S Haas, JW W Brown, and NR R Pace. Mycoplasma fermentans simplifies our view of the catalytic core of ribonuclease P RNA. *RNA*, 2(5):452–62, May 1996. ISSN 1355-8382.

[21] Alexei V Kazantsev and Norman R Pace. Bacterial RNase P: a new view of an ancient enzyme. *Nat. Rev. Microbiol.*, 4(10):729–40, October 2006. ISSN 1740-1534.

[22] Irmtraud M Meyer. A practical guide to the art of RNA gene prediction. *Briefings in Bioinformatics*, 8(6):396–414, November 2007. ISSN 1477-4054.

[23] Michael Zuker and David Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, 1984.

[24] Susan M Freier, Ryszard Kierzek, John A Jaeger, Naoki Sugimoto, Marvin H Caruthers, Thomas Neilson, and Douglas H Turner. Improved free-energy parameters for predictions of RNA duplex stability. *PNAS*, 83(24):9373–9377, 1986.

[25] David H Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288(5):911–940, 1999.

[26] Sean R Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Res.*, 22(11):2079–2088, 1994.

[27] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978.

[28] Ruth Nussinov and Ann B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *PNAS*, 77(11):6309–6313, 1980.

[29] Ignacio Tinoco, Olke C Uhlenbeck, and Mark D Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367, 1971.

[30] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9 (1):133–148, 1981.

[31] Ronny Lorenz, Stephan HF Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, Ivo L Hofacker, et al. ViennaRNA package 2.0. *Algorithms for Mol. Biol.*, 6(1):26, 2011.

[32] Michael Zuker. Prediction of rna secondary structure by energy minimization. In *Computer Analysis of Sequence Data: Part II*, pages 267–294. Springer New York, Totowa, NJ, 1994. ISBN 978-1-59259-512-9.

[33] Nicholas R Markham and Michael Zuker. UNAFold. *Bioinformatics: Structure, Function and Applications*, pages 3–31, 2008.

[34] Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):1, 2010.

[35] Rune B Lyngsø and Christian NS Pedersen. RNA pseudoknot prediction in energy-based models. *J. Comp. Biol.*, 7(3-4):409–427, 2000.

[36] YE Ding, Chi Yu Chan, and Charles E Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11(8):1157–1166, 2005.

[37] David H Mathews. Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, 359(3):526–532, 2006.

[38] Mirela Andronescu, Anne Condon, Holger H Hoos, David H Mathews, and Kevin P Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13):i19–i28, 2007.

[39] Mirela Andronescu, Anne Condon, Holger H Hoos, David H Mathews, and Kevin P Murphy. Computational approaches for RNA energy parameter estimation. *RNA*, 16(12):2304–2318, 2010.

[40] Elena Rivas, Raymond Lang, and Sean R Eddy. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, 18(2):193–212, 2012.

[41] John S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990. ISSN 0006-3525.

[42] David H Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8): 1178–90, August 2004. ISSN 1355-8382.

[43] Monir Hajiaghayi, Anne Condon, and Holger H Hoos. Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics*, 13(1): 1, 2012.

[44] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. Contrafold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14): e90–e98, 2006.

[45] Zhi John Lu, Jason W Gloor, and David H Mathews. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, 15(10):1805–1813, 2009.

[46] S Wuchty, W Fontana, I L Hofacker, and P Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999. ISSN 0006-3525.

[47] Michiaki Hamada, Hisanori Kiryu, Kengo Sato, Toutai Mituyama, and Kiyoshi Asai. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, 25(4):465–473, 2009.

[48] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Res.*, 32(16):4843–4851, 2004.

[49] Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. *J. Comp. Biol.*, 1(4):337–348, 1994.

[50] Mark A Larkin, Gordon Blackshields, NP Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–8, November 2007.

[51] Paul P Gardner, Andreas Wilm, and Stefan Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, 33(8): 2433–2439, 2005.

[52] Stephan H Bernhart, Ivo L Hofacker, Sebastian Will, Andreas R Gruber, and Peter F Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:1, 2008.

[53] Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

[54] Xing Xu, Yongmei Ji, and Gary D Stormo. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics*, 23(15):1883–1891, 2007.

[55] D Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.

[56] David H Mathews and Douglas H Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, 317(2):191–203, 2002.

[57] Arif Ozgun Harmanci, Gaurav Sharma, and David H Mathews. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*, 8(1):1, 2007.

[58] Jakob H Havgaard, Elfar Torarinsson, and Jan Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, 3(10):e193, 2007.

[59] Ivo L Hofacker, Stephan HF Bernhart, and Peter F Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–2227, 2004.

[60] Ian Holmes. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, 6(1):1, 2005.

[61] Yasuo Tabei, Koji Tsuda, Taishin Kin, and Kiyoshi Asai. ScaRNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *Bioinformatics*, 22(14):1723–1729, 2006.

[62] Sebastian Will, Kristin Reiche, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3(4):e65, 2007.

[63] Sebastian Will, Tejal Joshi, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18(5):900–914, 2012.

[64] Sebastian Will, Christina Otto, Milad Miladi, Mathias Mohl, and Rolf Backofen. SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, page btv185, 2015.

[65] Elena Rivas and Sean R Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(1):8, 2001.

[66] Stefan Washietl, Ivo L Hofacker, and Peter F Stadler. Fast and reliable prediction of noncoding RNAs. *PNAS*, 102(7):2454–9, February 2005. ISSN 0027-8424.

[67] Zizhen Yao, Zasha Weinberg, and Walter L Ruzzo. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–52, February 2006. ISSN 1367-4803.

[68] Xing Xu, Yongmei Ji, and Gary D Stormo. Discovering cis-regulatory RNAs in Shewanella genomes by Support Vector Machines. *PLoS Comput. Biol.*, 5(4):e1000338, April 2009. ISSN 1553-7358.

[69] Elena Rivas and Sean R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7): 583–605, July 2000. ISSN 1367-4803.

[70] Paul P Gardner, Jennifer Daub, John G Tate, Eric P Nawrocki, Diana L Kolbe, Stinus Lindgreen, Adam C Wilkinson, Robert D Finn, Sam Griffiths-Jones, Sean R Eddy, and Alex Bateman. Rfam: updates to the RNA families database. *Nucleic Acids Res.*, 37(suppl 1):D136–D140, 2009.

[71] Eric P Nawrocki, Sarah W Burge, Alex Bateman, Jennifer Daub, Ruth Y Eberhardt, Sean R Eddy, Evan W Floden, Paul P Gardner, Thomas A Jones, John Tate, et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, page gku1063, 2014.

[72] W Seffens and D Digby. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.*, 27(7):1578–84, April 1999. ISSN 0305-1048.

[73] Christopher Workman and Anders Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, 27(24):4816–4822, 1999.

[74] Peter Clote and Evangelos Kranakis. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, pages 578–591, 2005.

[75] Robin D Dowell and Sean R Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5(1):1, 2004.

[76] Bjarne Knudsen and Jotun Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6): 446–454, 1999.

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

[78] Stefan E Seemann, Jan Gorodkin, and Rolf Backofen. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, 36(20):6355–62, November 2008. ISSN 1362-4962.

[79] Jakob Skou Pedersen, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S Lander, Jim Kent, Webb Miller, and David Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, 2(4):e33, 2006.

[80] Richard Hughey and Anders Krogh. Hidden markov models for sequence analysis: extension and analysis of the basic method. *CABIOS*, 12(2):95–107, 1996.

[81] Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[82] Todd M Lowe and Sean R Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, 25(5):955–964, 1997.

[83] Osamu Gotoh. Optimal alignment between groups of sequences and its application to multiple sequence alignment. *CABIOS*, 9(3):361–370, 1993.

[84] Matthew C. Cowperthwaite and Lauren Ancel Meyers. How Mutational Networks Shape Evolution: Lessons from RNA Models. *Annu. Rev. Ecol. Evol. Syst.*, 38(1): 203–230, December 2007. ISSN 1543-592X.

[85] Yaron Orenstein, Yuhao Wang, and Bonnie Berger. RCK: accurate and efficient inference of sequence-and structure-based protein–RNA binding models from RNAcompete data. *Bioinformatics*, 32(12):i351–i359, 2016.

[86] E Van Nimwegen, J P Crutchfield, and M Huynen. Neutral evolution of mutational robustness. *PNAS*, 96(17):9716–9720, 1999. ISSN 00278424.

[87] Matthew C Cowperthwaite, J J Bull, and Lauren Ancel Meyers. From bad to good: Fitness reversals and the ascent of deleterious mutations. *PLoS Comput. Biol.*, 2 (10):e141, October 2006. ISSN 1553-7358.

[88] C O Wilke, J L Wang, C Ofria, R E Lenski, and C Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412 (6844):331–3, July 2001. ISSN 0028-0836.

[89] Andreas Wagner. Robustness and evolvability: a paradox resolved. *Proc. R. Soc. Lond. [Biol]*, 275(1630):91–100, January 2008. ISSN 0962-8452.

[90] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, volume 1. na, 1932.

[91] Jan Gorodkin, Laurie J. Heyer, Soeren Brunak, and GD Storomo. Displaying the in formation contents of structural RNA alignments: the structure logos. *CABIOS*, 13(6):583–586, 1997.

[92] Lauren Ancel Meyers, Jennifer F Lee, Matthew Cowperthwaite, and Andrew D Ellington. The robustness of naturally and artificially selected nucleic acid secondary structures. *J. Mol. Biol.*, 58(6):681–91, June 2004. ISSN 0022-2844.

[93] M A Huynen, P F Stadler, and W Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *PNAS*, 93(1):397–401, 1996. ISSN 00278424.

[94] Rafael Sanjuán, Javier Forment, and Santiago F Elena. In silico predicted robustness of viroid RNA secondary structures. II. Interaction between mutation pairs. *Mol. Biol. and Evol.*, 23(11):2123–30, November 2006. ISSN 0737-4038.

[95] W Grüner, R Giegerich, D Strothmann, C Reidys, J Weber, I L Hofacker, P F Stadler, and Peter Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration I. Neutral networks. *Monatshefte für Chemie Chemical Monthly*, 127(4):355–374, 1996. ISSN 0026-9247.

[96] Elhanan Borenstein and Eytan Ruppin. Direct evolution of genetic robustness in microRNA. *PNAS*, 103(17):6593–6598, April 2006. ISSN 0027-8424.

[97] Alexander Churkin, Moriah Cohen, Yonat Shemer-Avni, and Danny Barash. Bioinformatic Analysis of the Neutrality of Rna Secondary Structure Elements Across Genotypes Reveals Evidence for Direct Evolution of Genetic Robustness in Hcv. *J. Bioinform. Comput. Biol.*, 08(06):1013–1026, December 2010. ISSN 0219-7200.

[98] Guillermo Rodrigo and Mario a Fares. Describing the structural robustness landscape of bacterial small RNAs. *BMC Evol. Biol.*, 12(1):52, January 2012. ISSN 1471-2148.

[99] Guillermo Rodrigo and Santiago F Elena. MicroRNA precursors are not structurally robust but plastic. *Genome Biology and Evolution*, 5(1):181–6, January 2013. ISSN 1759-6653.

[100] Juan Antonio Garcia-Martin, Amir H Bayegan, Ivan Dotu, and Peter Clote. RNAdualPF: software to compute the dual partition function with sample applications in molecular evolution theory. *BMC bioinformatics*, 17(1):424, 2016.

[101] Andreas R Gruber, Stephan H Bernhart, Ivo L Hofacker, and Stefan Washietl. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, 9(1):122, 2008.

[102] Alexander Churkin and Danny Barash. RNAmute: RNA secondary structure mutation analysis tool. *BMC Bioinformatics*, 7:221, January 2006. ISSN 1471-2105.

[103] Wenjie Shu, Xiaochen Bo, Rujia Liu, Dongsheng Zhao, Zhiqiang Zheng, and Shengqi Wang. RDMAS: a web server for RNA deleterious mutation analysis. *BMC Bioinformatics*, 7:404, January 2006. ISSN 1471-2105.

[104] Wenjie Shu, Xiaochen Bo, Zhiqiang Zheng, and Shengqi Wang. RSRE: RNA structural robustness evaluator. *Nucleic Acids Res.*, 35(Web Server issue):W314–9, July 2007. ISSN 1362-4962.

[105] Jérôme Waldispühl, Srinivas Devadas, Bonnie Berger, and Peter Clote. Efficient algorithms for probing the RNA mutation landscape. *PLoS Comput. Biol.*, 4(8): e1000124, January 2008. ISSN 1553-7358.

[106] Radhakrishnan Sabarinathan, Hakim Tafer, Stefan E Seemann, Ivo L Hofacker, Peter F Stadler, and Jan Gorodkin. RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Human Mutation*, 34(4):546–556, 2013.

[107] Hisanori Kiryu and Kiyoshi Asai. Rchange: algorithms for computing energy changes of RNA secondary structures in response to base mutations. *Bioinformatics*, 28(8):1093–101, April 2012. ISSN 1367-4811.

[108] Justin Ritz, Joshua S Martin, and Alain Laederach. Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC Genomics*, 13 Suppl 4 (Suppl 4):S6, January 2012. ISSN 1471-2164.

[109] Cédric Notredame. Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLoS Comput. Biol.*, 3(8):4, August 2007. ISSN 1553-7358.

[110] Ye Ding and Charles E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31(24):7280–7301, 2003.

[111] Peter Menzel, J A N Gorodkin, and Peter F Stadler. The tedious task of finding homologous noncoding RNA genes. *RNA*, 15(12):2075–2082, 2009.

[112] S W Burge, J Daub, R Eberhardt, J Tate, L Barquist, E P Nawrocki, S R Eddy, P P Gardner, and A Bateman. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res*, 41:D226–32, 2013. ISSN 1362-4962.

[113] Tanja Gesell and Stefan Washietl. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics*, 9(1):248, 2008.

[114] Yasuo Tabei, Hisanori Kiryu, Taishin Kin, and Kiyoshi Asai. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, 9:33, 2008.

[115] I.L. Hofacker, W. Fontana, P.F. F. Stadler, L.S. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Chemical Monthly*, 125(2):167–188, February 1994. ISSN 0026-9247.

[116] Juan Antonio Garcia-Martin, Peter Clote, and Ivan Dotu. RNAiFOLD: a constraint programming algorithm for RNA inverse folding and molecular design. *J. Bioinform. Comput. Biol.*, 11(02):1350001, 2013.

[117] Juan Antonio Garcia-Martin, Ivan Dotu, and Peter Clote. RNAifold 2.0: a web server and software to design custom and rfam-based RNA molecules. *Nucleic Acids Res.*, 43(W1):W513–W521, 2015.

[118] Gergely J Szöllosi and Imre Derényi. Congruent evolution of genetic and environmental robustness in micro-RNA. *Mol. Biol. and Evol.*, 26(4):867–74, April 2009.

[119] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27):1–27, 2011.

[120] Alexander Churkin and Danny Barash. An efficient method for the prediction of deleterious multiple-point mutations in the secondary structure of RNAs using suboptimal folding solutions. *BMC Bioinformatics*, 9:222, January 2008. ISSN 1471-2105.

[121] Pablo Cordero, Julius B Lucks, and Rhiju Das. An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics*, 28:3006–8, 2012. ISSN 1367-4811.

[122] Angela Re, Tejal Joshi, Eleonora Kulberkyte, Quaid Morris, and Christopher T Workman. RNA–Protein Interactions: An Overview. In *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, pages 491–521. Springer, 2014.

[123] Elke Van Assche, Sandra Van Puyvelde, Jos Vanderleyden, and Hans P Steenackers. RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Frontiers in Microbiology*, 6, 2015.

[124] Zahra Shajani, Michael T Sykes, and James R Williamson. Assembly of bacterial ribosomes. *Ann. Rev. of Biochem.*, 80:501–526, 2011.

[125] Janosch Hennig and Michael Sattler. Deciphering the protein-RNA recognition code: Combining large-scale quantitative methods with structural biology. *BioEssays*, 37(8):899–908, 2015.

[126] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499 (7457):172–177, 2013.

[127] Grégoire Masliah, Pierre Barraud, and Frédéric H-T Allain. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cellular and Molecular Life Sciences*, 70(11):1875–1895, 2013.

[128] Craig Tuerk and Larry Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968): 505–510, 1990.

[129] Andrew D Ellington and Jack W Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287):818–822, 1990.

[130] HERVE Moine, C Cachia, E Westhof, B Ehresmann, and C Ehresmann. The RNA binding site of S8 ribosomal protein of Escherichia coli: Selex and hydroxyl radical probing studies. *RNA*, 3(3):255–268, 1997.

[131] Jan Hoinka, Alexey Berezhnoy, Phuong Dao, Zuben E Sauna, Eli Gilboa, and Teresa M Przytycka. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res.*, 43(12):5699–5707, 2015.

[132] Sandeep Ameta, Marie-Luise Winz, Christopher Previti, and Andres Jäschke. Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic Acids Res.*, 42(2):1303–1310, 2014.

[133] Mark A Ditzler, Margaret J Lange, Debojit Bose, Christopher A Bottoms, Katherine F Virkler, Andrew W Sawyer, Angela S Whatley, William Spollen, Scott A Givan, and Donald H Burke. High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Res.*, 41(3):1873–1884, 2013.

[134] Alexey Berezhnoy, C Andrew Stewart, James O Mcnamara II, William Thiel, Paloma Giangrande, Giorgio Trinchieri, and Eli Gilboa. Isolation and optimization of murine IL-10 receptor blocking oligonucleotide aptamers using high-throughput sequencing. *Molecular Therapy*, 20(6):1242–1250, 2012.

[135] Gillian V Kupakuwana, James E Crill, Mark P McPike, and Philip N Borer. Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. *PloS One*, 6(5):e19395–e19395, 2011.

[136] Minseon Cho, Yi Xiao, Jeff Nie, Ron Stewart, Andrew T. Csordas, Seung Soo Oh, James A. Thomson, and H. Tom Soh. Quantitative selection of dna aptamers through microfluidic selection and high-throughput sequencing. *PNAS*, 107(35): 15373–15378, 2010.

[137] José I Jiménez, Ramon Xulvi-Brunet, Gregory W Campbell, Rebecca Turk-MacLeod, and Irene A Chen. Comprehensive experimental fitness landscape and evolutionary network for small RNA. *PNAS*, 110(37):14984–14989, 2013.

[138] Jason N Pitt and Adrian R Ferré-D'Amaré. Rapid construction of empirical RNA fitness landscapes. *Science*, 330(6002):376–379, 2010.

[139] Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, 34(17):e117–e117, 2006.

[140] Xiao Li, Gerald Quon, Howard D Lipshitz, and Quaid Morris. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, 16 (6):1096–1107, 2010.

[141] Kate B Cook, Timothy R Hughes, and Quaid D Morris. High-throughput characterization of protein–RNA interactions. *Briefings in Functional Genomics*, 14(1): 74–89, 2015.

[142] Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The MEME Suite. *Nucleic Acids Res.*, 43(W1):W39–W49, 2015.

[143] Martin C Frith, Neil FW Saunders, Bostjan Kobe, and Timothy L Bailey. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, 4(5):e1000071, 2008.

[144] Timothy L Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[145] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 1994.

[146] Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, 6(7):e1000832, 2010.

[147] Jan Hoinka, Elena Zotenko, Adam Friedman, Zuben E Sauna, and Teresa M Przytycka. Identification of sequence–structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics*, 28(12):i215–i223, 2012.

[148] Peng Jiang, Zhonggang Hou, Nicholas E Propson, H Tom Soh, James A Thomson, and Ron Stewart. MPBind: A Meta-Motif Based Statistical Framework and Pipeline to Predict Binding Potential of SELEX-derived Aptamers. *Bioinformatics*, 30(18):2665–2667, 2014.

[149] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology*, 15(1):1, 2014.

[150] J Caroli, C Taccioli, A De La Fuente, Paolo Serafini, and S Bicciato. APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data. *Bioinformatics*, 32(2):161–164, 2016.

[151] Phuong Dao, Jan Hoinka, Yijie Wang, Mayumi Takahashi, Jiehua Zhou, Fabrizio Costa, John Rossi, John Burnett, Rolf Backofen, and Teresa M Przytycka. AptaTRACE: Elucidating Sequence-Structure Binding Motifs by Uncovering Selection Trends in HT-SELEX Experiments. *Cell Systems*, 3(1):62–70, 2016.

[152] Daniel C Reid, Brian L Chang, Samuel I Gunderson, Lauren Alpert, William A Thompson, and William G Fairbrother. Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. *RNA*, 15(12):2385–2397, 2009.

[153] Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Peña Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnol.*, 27(7):667–670, 2009.

[154] Christian Schudoma, Patrick May, Viktoria Nikiforova, and Dirk Walther. Sequence–structure relationships in RNA loops: establishing the basis for loop homology modeling. *Nucleic Acids Res.*, 38(3):970–980, 2010.

[155] Jodi M Ryter and Steve C Schultz. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *The EMBO Journal*, 17(24):7505–7513, 1998.

[156] Andres Ramos, Stefan Grünert, Jan Adams, David R Micklem, Mark R Proctor, Stefan Freund, Mark Bycroft, Daniel St Johnston, and Gabriele Varani. RNA recognition by a staufen double-stranded RNA-binding domain. *The EMBO Journal*, 19(5):997–1009, 2000.

[157] Janice M Zengel and Lasse Lindahl. Diverse mechanisms for regulating ribosomal protein synthesis in Escherichia coli. *Progress in Nucleic Acid Research and Molecular Biology*, 47:331–370, 1994.

[158] Ulrich Stelzl, Janice M Zengel, Marina Tovbina, Marquis Walker, Knud H Nierhaus, Lasse Lindahl, and Dinshaw J Patel. RNA-structural mimicry in Escherichia coli ribosomal protein L4-dependent regulation of the S10 operon. *J. Biol. Chem.*, 278 (30):28237–28245, 2003.

[159] Yang Fu, Kaila Deiorio-Haggar, Jon Anthony, and Michelle M Meyer. Most RNAs regulating ribosomal protein biosynthesis in Escherichia coli are narrowly distributed to Gammaproteobacteria. *Nucleic Acids Res.*, 41(6):3491–3503, 2013.

[160] Alexander Serganov, Ann Polonskaia, Bernard Ehresmann, Chantal Ehresmann, and Dinshaw J Patel. Ribosomal protein S15 represses its own translation via adaptation of an rRNA-like fold within its mRNA. *The EMBO Journal*, 22(8): 1898–1908, 2003.

[161] Betty L Slinger, Kaila Deiorio-Haggar, Jon S Anthony, Molly M Gilligan, and Michelle M Meyer. Discovery and validation of novel and distinct RNA regulators for ribosomal protein S15 in diverse bacterial phyla. *BMC Genomics*, 15(1):657, 2014.

[162] Kaila Deiorio-Haggar, Jon Anthony, and Michelle M Meyer. RNA structures regulating ribosomal protein biosynthesis in bacilli. *RNA Biology*, 10(7):1180–1184, 2013.

[163] AA Serganov, B Masquida, E Westhof, C Cachia, C Portier, M Garber, B Ehresmann, and C Ehresmann. The 16S rRNA binding site of Thermus thermophilus ribosomal protein S15: comparison with Escherichia coli S15, minimum site and structure. *RNA*, 2(11):1124, 1996.

[164] Lionel Bénard, Nathalie Mathy, Marianne Grunberg-Manago, Bernard Ehresmann, Chantal Ehresmann, and Claude Portier. Identification in a pseudoknot of a UG motif essential for the regulation of the expression of ribosomal protein S15. *PNAS*, 95(5):2564–2567, 1998.

[165] Lincoln G Scott and James R Williamson. The binding interface between Bacillus stearothermophilus ribosomal protein S15 and its 5'-translational operator mRNA. *J. Mol. Biol.*, 351(2):280–290, 2005.

[166] Betty L Slinger, Hunter Newman, Younghan Lee, Shermin Pei, and Michelle M Meyer. Co-evolution of Bacterial Ribosomal Protein S15 with Diverse mRNA Regulatory Structures. *PLoS Genetics*, 11(12):e1005720, 2015.

[167] Betty L Slinger and Michelle M Meyer. RNA regulators responding to ribosomal protein S15 are frequent in sequence space. *Nucleic Acids Res.*, page gkw754, 2016.

[168] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28 (23):3150–3152, 2012.

[169] Shawn Hoon, Bin Zhou, Kim D Janda, Sydney Brenner, and Jonathan Scolnick. Aptamer selection by high-throughput sequencing and informatic analysis. *Biotechniques*, 51(6):413–416, 2011.

[170] Marc Parisien and Francois Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–55, 2008.

[171] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/.

[172] John F Milligan, Duncan R Groebe, Gary W Witherell, and Olke C Uhlenbeck. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic dna templates. *Nucleic Acids Res.*, 15(21):8783–8798, 1987.

[173] Elizabeth E Regulski and Ronald R Breaker. In-line probing analysis of riboswitches. In *Post-Transcriptional Gene Regulation*, pages 53–67. Springer, 2008. ISBN 978-1-59745-033-1.