Ab initio Identification of Regulatory RNAs Using Information-theoretic Uncertainty

by

Amirhossein Manzourolajdad

(Under the Direction of Dr. Jonathan Arnold)

Abstract

RNA regulatory elements play a significant role in gene regulation. Riboswitches are regulatory elements which function by forming a ligand-induced alternative fold that controls access to ribosome binding sites or other regulatory sites in RNA. Traditionally, riboswitches have been identified based on sequence and structural homology. In this work, in an attempt to devise an *ab initio* method for identification of regulatory elements, mainly riboswitches, we derive and implement Shannon's entropy of the SCFG ensemble on an RNA sequence in polynomial time for both structurally ambiguous and unambiguous grammars. We then evaluate the significance of this new measure of structural entropy in identifying riboswitches. Finally, simple lightweight stochastic context-free grammar folding models assign significant values to long extensive secondary structures in *Bacillus subtilis*.

Index words: Riboswitch, RNA Secondary Structure, Stochastic Context-free Grammars, Information Theory

Ab initio IDENTIFICATION OF REGULATORY RNAS USING INFORMATION-THEORETIC UNCERTAINTY

by

AMIRHOSSEIN MANZOUROLAJDAD

BS, The University of Guilan, Rasht, Iran, 2004

MS, Isfahan University of Technology, Isfahan, Iran, 2007

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

Ab initio IDENTIFICATION OF REGULATORY RNAS USING INFORMATION-THEORETIC UNCERTAINTY

by

AMIRHOSSEIN MANZOUROLAJDAD

Approved:

Major Professor:	Jonathan Arnold
Committee:	Sidney Kushner
	Russell L. Malmberg
	Jan Mrazek

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

I dedicate this work to my mother, *Roya*

ACKNOWLEGEMENTS

I will be eternally indebted to my major advisor Dr. Jonathan Arnold and also to my committee—Dr. Sidney Kushner, Dr. Russell L. Malmberg, and Dr. Jan Mrazek—without whom this work would not have been possible. I would like to thank Dr. Jamal Golestani whose information theory course at Isfahan Univ. of Technology first inspired me to explore the field. I also would like to thank my first algorithm coach Abolfazl Javan.

I would like to thank all my fellow colleagues and the professors at the University of Georgia whom I met and learned a great deal from in various labs. Finally, I would like to acknowledge my family and friends for standing beside me throughout this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

With the exponential growth in genomic data, there is a fascinating opportunity to understand the blueprints of life. The interactions taking place between various molecules within living organisms, such as deoxyribonucleic acids (DNA) and ribonucleic acids (RNA), and proteins have informed us about the biological functions and pathways in the ever evolving kingdom of life.

Traditionally, protein-coding genes have been at the center of evolutionary biology studies, as molecular units of heredity. In fact, metabolic pathways can be comprised of complex systems of gene expression, which in turn are triggered by transcription of the DNA and subsequent translation of the messenger RNA (mRNA) into protein. A major class of molecules in the regulation of transcription and translation of genes are the non-protein-coding RNAs (ncRNA) (Morris, 2008; Barrandon et al., 2008; Repoila and Darfeuille, 2009; Morris, 2012). Non-protein-coding RNAs have been left on the *invisible* side of genomic research for some time (Eddy, 2001). They have certain functional similarities with proteins in that some ncRNAs can carry out catalytic activities. Also, the structure of such regulatory RNAs as well as their folding dynamics is essential to their function (Hall et al., 1982; Lu et al., 1996; Simmonds et al., 2008).

Comparative genomics in prokaryotes, which have simpler and a smaller genomes than that of eukaryotes, have led to the discovery of many ncRNAs that take part in the regulation of their downstream coding sequences [reviewed by Grundy and Henkin (2006)], a process known as *cis*-regulation. One of the earliest discovered regulatory RNA elements was the upstream region of the tryptophan operon in *Escherichia coli* (Oppenheim and Yanofsky, 1980), which is involved in the regulation of tryptophan biosynthesis.

1.1 Riboswitches

Cis-Regulatory RNA elements play an important role in the activation or termination of transcription and translation by altering their conformation. In this way these RNA regulatory elements can block or sequester translation start sites of downstream operons. They can potentially bind to a variety of protein factors, tRNAs, metabolites, amino acids, and other small molecules with high affinity and specificity to allow a specific response to signals in the cell or environment. They can also respond to changes in environmental factors such as pH, temperature, and ion concentration.

Recent sequence homology searches upstream of bacterial coding regions have led to the discovery of many riboswitches (Mironov et al., 2002; Nahvi et al., 2002; Winkler et al., 2002a). Riboswitches are defined as regulatory elements that do not require protein factors for their function, although the term riboswitch has had varying uses. Riboswitches, are usually located in the non-coding regions of the mRNA (Breaker, 2012) and are capable of regulating genes through both activation and attenuation of either transcription or translation [reviewed by Henkin (2008)].

1.2 Structural Homology

In many cases, elements that belong to the same class of riboswitch but reside in different organisms are observed to have similar conformations. Serganov et al. (2004) discuss this similarity for the purine riboswitches. Structural homology searches based on the RNA secondary structure upstream of prokaryotic untranslated regions have been very rewarding in discovering novel *cis*-regulatory elements over the past twenty years (Weinberg et al., 2007, 2010). Serganov and Nudler (2013) review the structural and functional complexities of already discovered riboswitches. Indeed, the secondary structure of the RNA plays a critical role in scaffolding the tertiary structure (Cech et al., 1994; Batey et al., 1999; Tinoco and Bustamante, 1999; Westhof et al., 2011; Bernauer et al., 2011).

However, structural homology methods have not always been successful in identification of all structural variants of riboswitches that bind to the same ligand across prokaryotes. Weinberg et al. (2008) describe the failure of detecting SAM-IV riboswitches in similarity searches based on structural profiles built from SAM-I riboswitches. They further hypothesized a far greater structural diversity for undiscovered riboswitches and suggested a possible lack of connection between structures of riboswitches and the nature of their cognate metabolites. Alternative approaches in riboswitch identification, are hence, more desirable than ever. Breaker (2012) raises the possibility of at least 100 more undiscovered riboswitches in the available bacterial genomes.

1.3 *Ab initio* Identification of Riboswitches

Experimental methods such as liquid-state nuclear magnetic resonance (NMR) [A review done by Scott and Hennig (2008)] can be effectively used to determine RNA structure. Structural alignment has also been used to trace conservation across homologous RNA genes (A collection are available in Rfam database [Griffiths-Jones et al. (2005); Gardner et al. (2009)]). On the computational side, however, there are two main methods to infer the secondary structure of a given RNA sequence: covariance models using stochastic context-free grammars (SCFG) (Chomsky, 1959; Dowell and Eddy, 2004; Nawrocki and Eddy, 2013) and minimization of folding energy on the RNA secondary structural level (MFE) (Zuker and Stiegler, 1981; McCaskill, 1990). Other predictions based on the Boltzmann ensemble such as the centroid-based approaches have also been calculated (Sato et al., 2009). The prediction of the final folding state of the riboswitch can be very informative and can assist us in inferring their biological functions.

Among other intriguing features of riboswitches are their ability to fold into two mutually exclusive secondary structures required by their biological function, hence the term *riboswitch*. In fact, the folding dynamics of riboswitches have been of great interest. Quarta et al. (2009) presented a case study of the TPP riboswitch by examining its energy landscape. They sampled the energy landscape of the TPP riboswitch and clustered the sampled structures into two groups based on their pair-wise base-pair distances. After repeating this process for various choices of length of the TPP riboswitch, they showed that for certain ranges of length, the each cluster corresponds to one of the two structures of the riboswitch (see Figure 1.1). However, to date there has not been a computational method that can identify the diverse and structurally complex riboswitches with high confidence.

In this work we have attempted to devise an *ab initio* method aimed at characterizing the folding space of the riboswitch which has applications to RNAs with the potential to have alternative fold(s). We de-

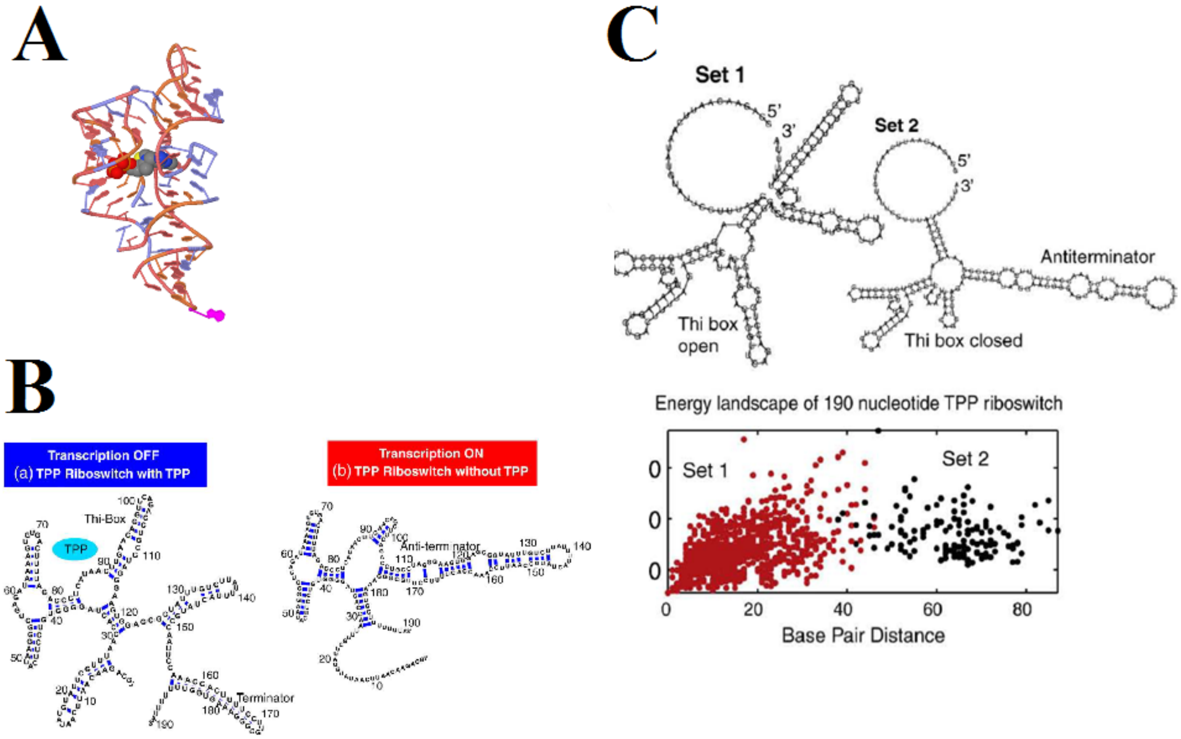

Figure 1.1: Energy Landscape of The TPP Riboswitch. **A:** Tertiary structure of an *E. coli* TPP (or *thi*-box) riboswitch bound to thiamine pyrophosphate (Edwards and Ferre-D'Amare, 2006). The image was generated by the Jmol from the PDB:2hoj structure taken from the Rfam website (Griffiths-Jones et al., 2005). **B:** Ligand-bound and unbound secondary structures of a TPP riboswitch in *B. subtilis*, taken from Quarta et al. (2009). **C:** Energy landscape of the *B. subtilis* riboswitch taken from Quarta et al. (2009). Set-1 and Set-2 clusters correspond to the two mutually exclusive secondary structures of the TPP riboswitch. Pairwise Base-pairing distance used as a measure of distance between two structures. Please refer to Quarta et al. (2009) for detailed information about the figure and clustering details.

ploy information-theoretic uncertainty or Shannon's entropy (Shannon, 1948) as a quantitative method to measure the diversity of the *complete* folding space of the RNA sequence under various SCFGs. Being a measure of entropy of a given probabilistic distribution[1], Shannon's entropy has been shown to be a very useful measure across various fields of science. In Chapter 2, we offer the derivations for calculating the Information-theoretic uncertainty of the secondary-structural folding space of any RNA sequence under a given SCFG folding model as a measure of structural entropy. We then investigate the significance of structural entropy of various RNA families not limited to riboswitches. This was done under various SCFGs and randomization tests. The work presented in Chapter 2 has been published in the journal of theoretical biology `http://www.sciencedirect.com/science/article/pii/S0022519312005620`.

After evaluating the structural entropy of various SCFG and their relationship to sequence and structural features of RNA structure, in Chapter 3 we focus on riboswitches, devising an *ab initio* approach for riboswitch identification based on structural entropy. The significance of structural entropy of riboswitches with respect to other biological sequences was then studied along with other measures of structural diversity. Unlike Chapter 2, in Chapter 3 we use to real biological sequences such as the antisense sequence of the riboswitch and intergenic regions of prokaryotes, avoiding the use of computer-generated random sequences for statistical analyses. We then report our results for various riboswitch identifiers tested against the *Bacillus subtilis* intergenic regions. Finally, the conclusions of our work is presented in chapter 4.

[1] The specific formulation of Shannon's entropy makes it possible to account for all structures in an SCFG-modeled RNA secondary structure space in polynomial time. The formulation has been used here mainly due to its computational convenience. Other formulations of entropy that do not use the $\log$ term in their definitions may not lead to polynomial time calculations. Shannon entropy, however, is not necessarily the best way to infer the entropy of a given distribution in itself. In fact, Christiansen et al. (2013) reject the validity of the underlying assumptions of Shannon entropy in the discipline of secure systems.

Chapter 2

Information-Theoretic Uncertainty of SCFG-Modeled Folding Space of The Non-coding RNA

Amirhossein Manzourolajdad, Yingfeng Wang, Timothy I. Shaw, and Russell L. Malmberg. Journal of Theoretical Biology. Year: 2013. Volume 318, 7 February 2013, Pages 140-163. Reprinted with permission of publisher.

2.1 Abstract

RNA secondary structure ensembles define probability distributions for alternative equilibrium secondary structures of an RNA sequence. Shannon's Entropy is a measure for the amount of diversity present in any ensemble. In this work, Shannon's entropy of the SCFG ensemble on an RNA sequence is derived and implemented in polynomial time for both structurally ambiguous and unambiguous grammars. Micro RNA sequences generally have low folding entropy, as previously discovered. Surprisingly, signs of significantly high folding entropy were observed in certain ncRNA families. More effective models coupled with targeted randomization tests can lead to a better insight into folding features of these families. Availability: `http://www.plantbio.uga.edu/~russell/index.php?s=1&n=5&r=0`.

2.2 Introduction

Non-protein-coding RNAs (ncRNA) have a critical role in gene regulation (Morris, 2008, 2012; Barrandon et al., 2008; Repoila and Darfeuille, 2009). They act as transcriptional and post-transcriptional regulators and are guides of chromatin-modifying complexes. Like protein-coding genes, small RNAs can also function either as activators or inhibitors of various genetic diseases (Taft et al., 2010). The function of a ncRNA is highly associated with its folding conformation (Hall et al., 1982; Lu et al., 1996; Simmonds et al., 2008).

Non-coding RNA sequences have different folding characteristics. Certain families of ncRNAs such as micro RNAs (miRNA) are believed to have a stable conformation definable by their secondary structure, while the folding conformation of transfer RNAs (tRNAs) is more complex and involves tertiary interactions (Scarabino et al., 1999; Du and Wang, 2003). Furthermore, depending on their regulatory roles, ncRNAs might possess more than one single conformation *in vivo*. Riboswitches are a group of regulatory ncRNAs that are generally required to have two alternative folds to perform their biological functions (Vitreschak et al., 2004; Quarta et al., 2009; Bocobza et al., 2007; Barash et al., 2006; Gilbert et al., 2007). Ribonuclease P (RNase P) sequences, another group of ncRNAs, are more complex in that they have an RNA component which directly binds to its protein component (Kazantsev and Pace, 2006; Niranjanakumari et al., 1998). The function of RNase P is to cleave off an extra, or precursor, sequence of RNA on tRNA molecules (Guerrier-Takada and Altman, 1993; Brannvall et al., 1998; Kazantsev and Pace, 2006; Niranjanakumari et al., 1998). RNase P sequences are generally longer than miRNAs and riboswitches and possess several pseudoknots in their conformation.

The secondary structure of RNA plays a critical role by scaffolding the tertiary structure (Cech et al., 1994; Batey et al., 1999; Tinoco and Bustamante, 1999; Westhof et al., 2011; Bernauer et al., 2011) making the RNA secondary structure modeling critical to ncRNA-related studies. The secondary structure consists of single strand loops enclosed by double-stranded helices formed by stacked canonical (here, Watson-Crick and Wobble) base-pairs of nucleotides. RNA secondary structure is mainly modeled based on the Minimum Free Energy (MFE) criterion; the structural conformation with the least free energy from amongst all possible conformations in a thermodynamic model is predicted as the secondary structure of the sequence. The MFE structure of the RNA sequence can be found through a dynamic programming optimization in polynomial time $O(n^3)$ (Zuker and Stiegler, 1981; McCaskill, 1990). Secondary structure prediction programs can achieve up to 70% accuracy by minimizing the global energy sum via dynamic programming (Zuker, 2003; Knudsen and Hein, 2003; Hofacker, 2003).

In addition to the thermodynamic models, Stochastic Context-free Grammars (SCFGs) have also been used for RNA secondary structure prediction and searches. A context-free grammar (CFG) is a formal descriptive system consisting of symbolic rewriting rules to generate languages of strings. For the alphabet of the four nucleotides, such grammars describe languages of RNA sequences, whose generations (called derivations) are by a series of rewriting rule applications. Context-free rules of form $X \to aYb$ model pairing between (possibly distant) nucleotides a and b. These pairings are either the result of hydrogen-bonds between complementary base pairs `A-U` and `C-G`, or the wobble pair `G-U`. A derivation process of a sequence thus yields an associated secondary structure. A stochastic CFG, with rules associated with probabilities, defines alternative structures of different probabilities for the same sequence, rendering an RNA secondary structure ensemble. In addition, a SCFG is reconfigurable for defining ncRNAs of specific secondary structures, e.g., in a structural homology search.

Accurate modeling of RNA folding is essential to ncRNA *ab initio* gene finding and structure prediction. Proper estimation of the probabilities associated with RNA structures is essential to developing an effective SCFG model. Maximum likelihood (ML) approaches such as the Cocke-Younger-Kasami (CYK)-based methods have demonstrated their merits in both SCFG-modeled RNA structure detection and prediction studies (Dowell and Eddy, 2004).

While ML approaches enable prediction of RNA structure under a probabilistic model, other targeted statistics may also lead to characterization of various ncRNA sequences. For instance, sampling of the folding space of certain ncRNA sequences under the Boltzmann thermodynamic model has proven useful in investigating alternative structures as well as distinguishing RNA sequences from random sequences (Ding and Lawrence, 2003; Chan and Ding, 2008; Miklos et al., 2005).

Our goal in this work was to define an application of Shannon's entropy to RNA structures and their structural variability to identify riboswitches. Our theoretical approach used stochastic context free grammars (SCFG) as folding models. We examined the properties of this measure by investigating the entropy of RNA sequences of various families under several well-established SCFG models. Additional tests are designed to investigate the possibility of significance of this measure on RNA sequences and various factors associated with it.

Information-theoretic uncertainty or Shannon's entropy $H(S)$ (Shannon, 1948) is a quantitative measure for the amount of (un)certainty about a random variable S, $H(S) = \sum_{s \in S} p(s) \log p(s)$. It can also be interpreted as the measure of diversity within a given distribution of values or probabilities. The ability of Shannon's entropy to capture the diversity in a given ensemble has made it useful for various disciplines of research such as genetic/biological evolutionary studies (Adami et al., 2000; Yockey, 2005) as well as characterization of DNA sequence motifs (Schneider and Stephens, 1990; D'Haeseleer, 2006). Shannon's entropy has also been deployed in RNA structural studies. A formulation for the base-pairing certainty has been introduced in (Huynen et al., 1997) and was shown to be able to capture effectively the structural stability of certain ncRNAs under both the Boltzmann and SCFG ensembles (Mathews, 2004; Wang et al., 2012; Shaw et al., 2011). Various formulations of base-pairing certainty, however, are only approximations of the secondary structural (un)certainty of a sequence. Here, we present a direct derivation of the information-theoretic uncertainty of the SCFG-modeled folding space of the RNA sequence, computable in polynomial time.

Sections 2 and 3 present our formulation of structural entropy using SCFG. Sections 4 and 5 consider the application of SCFG models to several RNA families with known structure. Results from applying the structural entropy to these RNAs are presented in section 6. Finally, sections 7 and 8 contain our discussion and overall conclusion.

2.3 Stochastic Context-Free Grammar (SCFG) ensemble of RNA secondary structure

The SCFG-modeled secondary structure space of a given sequence assigns posterior probabilities to all possible secondary structures definable over the sequence. The resulting ensemble of structures is pseudoknot-free conformations, due to the context-free nature of the grammar. That is, for any four nucleotides of positions, $i < j < k < l$, base pairs between positions i and k and between j and l cannot occur at the same time[1]. This constraint greatly reduces the structural space and ensures computational efficiency of structure prediction algorithms. Both RNA sequences and their secondary structures can be described with SCFGs. Since a CFG defines a language of strings using generating rules, a collection of RNA sequences can be defined by CFG using the alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$. Formally, let $y = y_1 y_2 \dots y_n$ be a given sequence, where $y_k \in \Sigma$, where $k = 1, 2, \dots, n$. A derivation π of y by the grammar is:

$$S_0 = \gamma_0 \Rightarrow \gamma_1 \Rightarrow \gamma_2 \Rightarrow \cdots \Rightarrow \gamma_{n_\pi} = y \tag{2.1}$$

where $\gamma_i \in (N \cup \Sigma)^*$, and $\gamma_i \Rightarrow \gamma_{i+1}$ satisfies that $\gamma_i = \omega X \beta$ and $\gamma_{i+1} = \omega \alpha \beta$ for some $\omega \in \Sigma^*$, $\beta \in (N \cup \Sigma)^*$ and rule $X \to \alpha$ in the grammar, i.e., the occurrence of nonterminal X is rewritten with string α. The derivation is also called a *left-most derivation* because the nonterminal X, chosen to be replaced by α, is the left-most nonterminal on the string γ_i (note ω is a string of all terminals).

We denote the derivation (2.1) by $S_0 \Rightarrow_\pi^* y$. Left-most derivations have one-to-one correspondence with *derivation trees* (or parsing trees). Each such derivation (and the corresponding parsing tree) contains all the information of the corresponding secondary structure folded by the sequence. Equation (2.2) illustrates the correspondence between derivations and secondary structures with CFG, where an example grammar with only four types of generic rules is used:

$$X \to aYbZ, \;\; X \to aYb, \;\; X \to aY, \;\; X \to a \tag{2.2}$$

where X, Y and Z are non-terminals and a and b are terminals for nucleotides in Σ. The first two rules define base pairs between two nucleotides represented by a and b, the last two define unpaired nucleotides represented by a. The first rule also allows assembly of parallel substructures.

[1] Structures with pseudoknots are of much higher computational complexity and are not considered in this paper.

Since sequences and derivations are completely defined by the grammar G, it also defines the space of structures for all sequences that it can derive. The probability distribution of the structures in the space is the probability distribution associated with derivations of all derivable sequences by G. In particular, the probability $P(S_0 \Rightarrow_\pi^* y)$ associated with the derivation π of sequence y in (2.1) under a given SCFG Model (G, Θ) is defined as

$$p(\pi, y|G, \Theta) \equiv p(\pi, y) = P(S_0 \Rightarrow_\pi^* y) = \prod_{i=1}^{n_\pi} P(R_\pi^i) \tag{2.3}$$

where R_π^i is the grammar rule associated with the one-step derivation $\gamma_{i-1} \Rightarrow \gamma_i$ in (2.1).

2.4 Structural entropy over SCFG ensembles

As noted previously, Shannon's entropy measures the (un)certainty associated with a random event. When the secondary structure folding of a given RNA sequence y is considered as such an event, it refers to the entropy of the probability distribution of the folding space of the given sequence. Denoted as $H(\Pi|y, G, \Theta)$, the folding entropy is both function of sequence y and folding model (G, Θ). In this section, we derive a closed form for the structural entropy of sequence y, which is computable in polynomial time. Let's use $H(\Pi|y)$ rather than $H(\Pi|y, G, \Theta)$ for simplicity of notation. Substituting $P(\pi, y) = P(S_0 \Rightarrow_\pi^* y)$ and $P(y) = P(S_0 \Rightarrow^* y)$ yields the structural entropy of a sequence y, $H(\Pi|y)$ to be equal to:

$$\log P(S_0 \Rightarrow^* y) - \frac{1}{P(S_0 \Rightarrow^* y)} \sum_{\pi \in \Pi(y)} P(S_0 \Rightarrow_\pi^* y) \log P(S_0 \Rightarrow_\pi^* y) \tag{2.4}$$

where $\Pi(y)$ is the space of secondary structures into which y can fold, defined by the underlying RNA secondary structure ensemble. The nonterminal $S_0 \in N$ is the start nonterminal symbol of the given SCFG and N is the set of nonterminals. We now show that the structure entropy can be directly derived over any given SCFG ensemble.

The total probability of y $P(S_0 \Rightarrow^* y)$ can be computed as $\alpha(S_0, 1, n_y, y)$. The inclusive definition of the *Inside* probability function α is used, here (Durbin, 1998). (See A.1)

We introduce some notations for the convenience of discussion. As used earlier, let π be a specific structure for y, defined by a specific left-most derivation $S_0 \Rightarrow_\pi^* y$. We use $\langle X \to \gamma, i, j\rangle_\pi$ to denote the instance of rule $X \to \gamma$ applied in π such that X derives $y_i \dots y_j$ in the left-most derivation $S_0 \Rightarrow_\pi^* y$, i.e.,

$$S_0 \Rightarrow_\pi^* y_1 \dots y_{i-1} X \lambda \Rightarrow_\pi^* y_1 \dots y_{i-1} \gamma \lambda$$

$$\Rightarrow_\pi^* y_1 \dots y_{i-1} y_i \dots y_j \lambda \Rightarrow_\pi^* y_1 \dots y_j y_{j+1} \dots y_{n_y} = y$$

for some $\lambda \in (N \cup \Sigma)^*$, where $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$.

Likewise, we denote an instance of rule $X \to \gamma$ applied in some structure by $\langle X \to \gamma, i, j\rangle$. Note that the applications of rule $X \to \gamma$ in $\langle X \to \gamma, i, j\rangle_\pi$ and in $\langle X \to \gamma, i, j\rangle$ have the same probability, which is the probability $\mathcal{F}(X \to \gamma)$ for rule $X \to \gamma$ given in the SCFG. The term $\sum\limits_{\pi \in \Pi(y)} P(S_0 \Rightarrow_\pi^* y) \log P(S_0 \Rightarrow_\pi^* y)$ in (2.4) becomes

$$\sum_{\pi \in \Pi(y)} \prod_{\langle Y \to \delta, k, l\rangle_\pi} \mathcal{F}(Y \to \delta) \log \prod_{\langle X \to \gamma, i, j\rangle_\pi} \mathcal{F}(X \to \gamma)$$

$$= \sum_{\pi \in \Pi(y)} \prod_{\langle Y \to \delta, k, l\rangle_\pi} \mathcal{F}(Y \to \delta) \sum_{\langle X \to \gamma, i, j\rangle_\pi} \log \mathcal{F}(X \to \gamma)$$

$$= \sum_{\langle X \to \gamma, i, j\rangle} \log \mathcal{F}(X \to \gamma) \sum_{\pi \ni \langle X \to \gamma, i, j\rangle} \prod_{\langle Y \to \delta, k, l\rangle_\pi} \mathcal{F}(Y \to \delta) \tag{2.5}$$

where term

$$\sum_{\pi \ni \langle X \to \gamma, i, j\rangle} \prod_{\langle Y \to \delta, k, l\rangle_\pi} \mathcal{F}(Y \to \delta)$$

is actually the total probability of all the left-most derivations $S_0 \Rightarrow_\pi^* y$ for all π that contain $\langle X \to \gamma, i, j\rangle$. That is, for $y = y_1 \dots y_{n_y}$

$$\sum_{\pi \ni \langle X \to \gamma, i, j\rangle} \prod_{\langle Y \to \delta, k, l\rangle_\pi} \mathcal{F}(Y \to \delta) =$$

$$\textstyle\sum_{\omega, \lambda} P(S_0 \Rightarrow^* \omega X \lambda, \omega \Rightarrow^* y_1 \dots y_i, \lambda \Rightarrow^* y_j \dots y_{n_y})$$

$$\times P(X \to \gamma, \gamma \Rightarrow^* y_{i+1} \dots y_{j-1})$$

$$= \beta(X, i, j, y) \mathcal{F}(X \to \gamma) P(\gamma \Rightarrow^* y_{i+1} \dots y_{j-1})$$

where the *Outside* probability function β is the inclusive definition of the outside probability function (Durbin, 1998). (See A.1) Replacing the corresponding terms in formulae (2.4) and (2.5) with the above derivations, the structural entropy of given sequence y is computed as

$$\log \alpha_0 - \frac{1}{\alpha_0} \sum_{i \le j} \sum_{X \to \gamma} \beta(X,i,j,y) \mathcal{F}(X \to \gamma) \log \mathcal{F}(X \to \gamma) P(\gamma \Rightarrow^* y_{i+1} \dots y_{j-1}) \tag{2.6}$$

where $\alpha_0 = \alpha(S_0, 1, n_y, y)$. Further derivations from $P(\gamma \Rightarrow^* y_i \dots y_j)$ will be γ-specific, though the technique is general and applicable to any SCFG. We will use the grammar rules of the four types given in (2.2), for simplicity of presentation (See A.1 for a derivation generalized to all types of non-stacking[2] grammar rules). Based on them, term $\sum_{X \to \gamma}$ in (2.6) can be computed as

$$\beta(X,i,j,y) \, [\, f(X, aYbZ) \sum_{i+2<k<j-1} \alpha(Y, i+2, k-1, y) \alpha(Z, k+1, j-1, y)$$

$$+ f(X, aYb) \alpha(Y, i+2, j-2, y) + f(X, aY) \alpha(Y, i+2, j-1) + \delta(i+2=j) f(X,a)) \,] \tag{2.7}$$

where $a = y_{i+1}, b = y_{j-1}$, $\delta(i = j)$ is the characteristic function, and the shorthand $f(X, \gamma)$ is used for

$$f(X,\gamma) = \mathcal{F}(X \to \gamma) \log \mathcal{F}(X \to \gamma).$$

Equation (2.7) is valid for all non-stacking structurally unambiguous grammars. In the case of structurally ambiguous grammars, however, the inside and outside probability functions in (2.7) must be modified according to the left-most derivation criterion to avoid redundant enumeration of derivation trees. A.1 contains the algorithmic details of a derivation generalized to ambiguous grammars. In dealing with structurally ambiguous grammars, we refer to (2.7) as the redundant derivation and its modified version in A.1 as the left-most derivation. The computational complexity of the left-most derivation entropy is the same as that of redundant derivation with the memory allocation being twice as high.

The uncertainty about the occurrence of a substructure s' can be computed via Shannon entropy $H(I_{s'}) = -p(s') \log p(s') - (1 - p(s')) \log(1 - p(s'))$. The most fundamental substructure of the secondary structure space of a sequence is the occurrence of pairing between two nucleotides. By summing up all the individual

[2] Non-stacking grammar rules here include: $X \to a$, $X \to aY$, $X \to Ya$, $X \to aYbZ$, $X \to YaZb$, and $X \to aYb$, but not $X \to \epsilon$.

base-pairing uncertainties, we can formulate a measure for the total pairing uncertainties for sequence y. We will refer to this figure as Total pairing (TP) entropy:

$$\texttt{TP Entropy}(y) = \sum_{i<j} H(I_{i,j}|y) \tag{2.8}$$

Where $I_{i,j}$ is a binary random variable representing one for pairing and zero for non-pairing events between two nucleotides i and j, and $H(I_{i,j}|y)$ is the uncertainty for the pairing of nucleotides of positions i and j in the sequence; $H(I_{i,j}|y) = -p(i,j)\log p(i,j) - (1-p(i,j))\log(1-p(i,j))$, where $p(i,j)$ refers to pairing probability of i and j. The total pairing entropy, however, is not a valid entropy of all base-pairs, since it neglects the interdependencies across base-pairs introduced by the corresponding SCFG. It can be easily shown that the total pairing entropy is an upper bound for the structural entropy for any given sequence and under any given SCFG model so long as the grammar model is structurally unambiguous (See A.9.1).

$$H(\Pi|y) \leq \texttt{TP Entropy}(y) \tag{2.9}$$

2.5 Evaluating the Structural Entropy of ncRNAs

The structural[3] entropy of a sequence can depend on various factors ranging from sequence and model features to folding characteristics of the family that the sequence belongs to. Primary structural variants such as sequence length and nucleotide composition can all affect the structural entropy of the RNA sequence. For instance, higher length is expected to increase the number of folding scenarios valid on the sequence, which in turn may increase folding entropy. High `GC`- or `AU`-composition can also affect the folding entropy of the sequence, since all folding models favor canonical base-pairs. Also, various modeling factors such as grammar rule design, grammar rule probability assignment, model accuracy to annotated secondary structure of the RNA sequence, inclusion of base-pair stacking in the model, and structural ambiguity of the grammar could all have an impact on the folding space and entropy of the sequence. Finally, *in vivo* conformational dynamics such as folding (in)stability, multiple folds, and formation of pseudoknots could all affect the structural entropy, since they are directly related to the folding distribution. There is no reason to believe that the entropy of sequences with such different structural features will have a similar behavior.

[3] Terms `folding entropy` and `structural entropy` both refer to the secondary structural entropy here and are used interchangeably in the text.

The impact of the above factors on the structural entropy makes the investigation of folding entropy of ncRNAs a challenging task. The degree of sensitivity of the structural entropy to a factor and how it might vary in the light of other factors is not known. Also, limitations of the secondary-structure modeling in capturing the tertiary conformation and its dynamics further complicates comparisons and biological interpretations based on the entropy, making co-evaluation of modeling and ncRNA conformational features inevitable. A thorough and comprehensive study is needed to effectively explore and compare the folding entropy of various classes of ncRNAs. Here, we only intend to offer a preliminary insight into this comparison with the specific goal of evaluating the significance of the mentioned factors on the folding entropy of the ncRNA. Tests are devised to study as many factors as possible given the time and complexity limitations of this work.

2.5.1 Prior Assumptions about the Micro RNA

A sequence of a single stable fold should have low entropy under a reasonably accurate folding model, since folding alternatives will be unlikely to occur for that sequence under the given model. Micro RNA sequences are known for a single and a stable secondary structure, having distinguishably low base-pairing entropy. We also expect the structural entropy of the miRNA to be low. This assumption, however, is only an intuition and is different from the null-hypotheses of various statistical tests performed here. As shall be described later in more detail, the null-hypothesis is that folding entropy values of classes of ncRNAs are neither significantly different from one another nor are they from that of a random sequence. Speculations about the entropy of miRNA shall be verified throughout the test. Should this assumption be confirmed, it may assist us in better investigating the impact of various modeling factors on the structural entropy.

2.6 Methodology

2.6.1 Choice of Folding Model

Two types of non-stacking grammars are considered here as representative folding models. The first type contains well-established CFG designs along with their corresponding parameter sets trained to imitate

RNA secondary structure. Three grammar models were arbitrarily selected from the four structurally unambiguous non-stacking grammars presented in (Dowell and Eddy, 2004) along with their trained parameters. Grammars G4 (RUN), originally developed jointly by (Dowell and Eddy, 2004) and Graeme Mitchison, G5 (IVO), developed jointly by (Dowell and Eddy, 2004) and Ivo Hofacker, and G6 (BJK), by Knudsen/Hein originally used in the Pfold package (Knudsen and Hein, 1999, 2003), were chosen. We used the conus software (Dowell and Eddy, 2004) to train each individual model based on the CYK-based training method described in (Dowell and Eddy, 2004). Three training sets: benchmark, mixed80, and rfamv5 (Dowell and Eddy, 2004) were deployed for training purposes. We use the notation `grammar (data set)` to refer to a particular grammar and its choice of parameter set. For instance `RUN (benchmark)` refers to deployment of the `RUN` grammar design whose corresponding parameter sets are obtained by training the model on the `benchmark` data set. Please refer to (Dowell and Eddy, 2004) for details about grammar rules and training data sets.

All the above three models structurally unambiguous with rule probabilities estimated by CYK-based training approaches. In order to avoid potential bias of results towards factors such as structural unambiguity and/or CYK-based model-training algorithms, we chose the second type of grammars to be structurally ambiguous with symmetrical rules and arbitrary rule probabilities:

$$S \rightarrow a \ \ (p_t), S \rightarrow aS \ \ (p_n), S \rightarrow Sa \ \ (p_n)$$

$$S \rightarrow aSbS \ \ (p_n), S \rightarrow SaSb \ \ (p_n), S \rightarrow aSb \ \ (p_n)$$

Values p_t and p_n are probabilities for the terminal rule and nonterminal rules, respectively. We examined two variations of the above model. In model denoted as RND1, we set $p_t = p_n$ and in the model denoted as RND10 we set $p_t = 0.1p_n$. Rule probabilities were then normalized for both models. Single nucleotide generation probabilities are set to 0.25 for all four nucleotides in both RND1 and RND10 models. For both models, the probability distribution of $\{0.25, 0.25, 0.17, 0.17, 0.08, 0.08\}$ is given to six canonical base-pairs G-C, C-G, A-U, U-A,G-U, and U-G, respectively. All non-canonical base pairs probabilities are set to zero. Both redundant and left-most derivation structural entropy calculation was implemented for RND1 and RND10 models, due to their structural ambiguity.

Folding models considered here are limited to non-stacking models, due to implementation constraints. In non-stacking grammars, base-pair probabilities are a priori independent of surrounding base-pairs. This is a great approximation of secondary structural folding compared to stacking models. Stacking SCFG models are much better imitations of the state-of-the-are thermodynamic folding models.

2.6.2 Data Collection

Having a reliable annotated secondary structure for the RNA sequence is essential to evaluating its folding entropy. Bralibase annotated secondary structures (Gardner et al., 2005) were carefully selected by the authors to be highly reliable. Bralibase secondary structures, however, are only available for a few classes of RNA sequences, namely tRNA, g2intron, U5, and rRNA. Rfam alignments, on the other hand, include more diverse classes and sequences than Bralibase. Rfam contains consensus secondary structure based on published literature or predicted using automated covariance-based methods (Gardner et al., 2009; Griffiths-Jones et al., 2005); however, the predicted structures are generally less reliable than Bralibase.

Finally, Rfam contains both SEED and FULL alignments. Unlike the SEED alignments, FULL alignments are not manually curated and often contain computationally predicted sequences, while they contain greater number and diversity of sequences. We chose the data of the Rfam SEED alignment for this study as a compromise between sequence diversity and reliability of annotated secondary structure. Conclusions about the relationship between model accuracy and structural entropy of the ncRNA, however, will then have to be made with great caution.

We downloaded 45 Rfam sequences from Rfam 10.0 SEED alignments. Our data set includes sequences of one stable secondary structure, such as miRNAs, as well as sequences known to have higher tertiary interactions such as tRNAs. Various sub-families of riboswitches are also included. Riboswitch sequences are known for alternative folds *in vivo*. The data also contain sequences known to have pseudoknots, such as RNAse P. Various other sequences are also included to have a more diverse general picture.

A total of 4116 sequences were downloaded. RNA families that were included in this work were: miRNA (12 families: 170 sequences), riboswitch (15 families: 1334 sequences, snRNA: 31 sequences), RNase MRP (1 family: 67 sequences), RNase P (4 families: 537 sequences) rRNA (3 families, 927 sequences), tRNA (1 family: 967 sequences), and snoRNAs (83 sequences). The detailed selection of Rfam accession number and additional information about the sequences regarding their average length and average identity is included in A.2.

Model Accuracy to Annotated Secondary Structure of ncRNAs

In order to evaluate the accuracy of the above folding models in predicting the RNA secondary structure, their CYK-based predictions were compared to the Bralibase annotated structures (Gardner et al., 2005) for the available sequences. Sensitivity and specificity of each model was calculated according to (Do et al., 2006), which is based on matched base-pairs and uses the Predictive Positive Value (PPV) as a means of model specificity.

The sensitivity and specificity of selected models for both Bralibase and Rfam SEED secondary structures is calculated to gain better insight into the overall accuracy of each model. Tables A.2 and A.3 contain average sensitivity and specificity (PPV) of various models to annotated secondary structure of classes of ncRNAs available in Bralibase and Rfam databases. The sensitivity and specificity of RUN and BJK grammars are significantly higher than those for the IVO grammar in both Bralibase and Rfam secondary structure annotations.

2.6.3 Measuring the Significance of Folding Entropy

Various primary and secondary structure randomizations have been performed to evaluate both folding significance of ncRNAs and investigate its relationship to various model-specific and sequence-specific factors. The p-value of folding entropy of various ncRNA sequences have been calculated against random background of sequences having similar length and nucleotide composition, since a priori we know that such primary structure features can each affect the structural entropy of the sequence. Hence, the general null-hypothesis is that structural entropy of classes of ncRNAs are neither significantly different from each other nor are they significantly different than that of a random sequence of similar length and nucleotide composition under any folding model. Various randomization techniques applied here will have a more specific null-hypothesis corresponding to the nature of the generated random sequence. Should the null-hypotheses be rejected, various sequence and model specific factors associated with significant folding entropy values are of interest.

In the first randomization, the significance of entropy values was calculated against a background of random sequences with the same length and nucleotide composition. Single-nucleotide composition preserved random sequences were generated using GenRGenS for each sequence, separately. P-values were then empirically obtained by comparing the entropy of the sequence to its corresponding distribution of structural entropy of random sequences. The null-hypothesis here is that the structural entropy of classes of ncRNAs are neither significantly different from each other nor are they significantly different than that of a random sequence of the same length and single-nucleotide composition under any folding model.

In the second randomization, the significance of entropy values was calculated against a background of random sequences with the same length and di-nucleotide composition. Random sequences with their di-nucleotide composition preserved were generated using the Altschul-Erikson algorithm for each sequence, separately. The algorithm was originally described in Altschul and Erickson (1985) and subsequently implemented and used in Clote et al. (2005) for comparison of structural RNA folding energy to random RNA. P-values were obtained by comparing the entropy of the sequence to its corresponding distribution of structural entropy of random sequences. The null-hypothesis here is that the structural entropy of classes of ncRNAs are neither significantly different from each other nor are they significantly different than that of a random sequence of the same length and di-nucleotide composition under any folding model.

A stability test of p-values on miRNAs and tRNAs with length 100 nucleotides under the BJK (mixed80) model was performed. The test shows that a random ensemble of size 100 will result in highly stable p-values (See A.7 for details of the stability test). Although the stability test results cannot be generalized to all sequences and choices of model, we chose the random ensemble size to be linearly proportional to sequence length for calculating p-values of ncRNA sequences in both single-nucleotide and di-nucleotide randomization tests; in the first and second randomization tests. Results for the rli54 riboswitch (5 sequences) are not available for randomization tests, due to high computational time. Also, results for bacterial type A RNase P and nuclear RNase P sequences are not available for the first and second randomization tests, due to their high computational time. Finally, only partial result is available for bacterial type B RNase P sequences in the first and second tests. (34 sequences out of the total of 114 in the single-nucleotide randomization test and 60 sequences for the di-nucleotide randomization test)[4]

[4] RNase P sequences used in single- and di-nucleotide randomization tests were limited, due to high computational time of the tests. 34 sequences were arbitrarily chosen for the single nucleotide randomization and are included in

In the third randomization, the significance of entropy values was calculated against a background of random sequences with structures typical of the corresponding model. The procedure is as follows: We first clustered sequences according to their length and nucleotide composition. Three-group k-means clustering of sequence lengths, yields cluster centers of $\{93, 185, 366\}$. We limited our test in this step to shortest cluster of ncRNA sequences. We then performed a second phase of clustering on single-nucleotide composition. A three-group clustering was chosen to be reasonable based on the k-means clustering curve (data not shown). The three-group clustering resulted clustering of short sequence into high, low, and average GC−composition, which are denoted as Clusters 1, 2, and 3, respectively. The sequences were then filtered for having length 93 ± 5, due to the observed high sensitivity of entropy to sequence length. Details about clusters and their corresponding sequences are available in A.5. The GenRGenS software package (Ponty et al., 2006) was then used to generate random sequences with structure typical of the given folding model. GenRGenS has the ability to generate random sequences of desired length from a given primary or secondary structural model. A.4 contains details about generating random structures for each model. In this randomization, only one background of random sequences is generated for a given cluster of sequences and a choice of model, rather than for each sequence separately as done in the first and second randomization tests. The above clustering scheme was arbitrarily selected as a compromise between significance and accuracy of comparison; i.e., further clustering sequences based on higher order of nucleotide composition, such as di-nucleotide composition, would yield a more accurate comparison between folding entropy of sequences while the generation of corresponding random sequences with structure would be less typical of the model, making investigations the significance of entropy values very difficult. Other clustering schemes may result in more comprehensive comparisons between ncRNAs with similar length and nucleotide composition. The null-hypothesis here is that the structural entropies of classes of ncRNAs are neither significantly different from each other nor are they significantly different than that of a sequence having a random secondary-structure of similar length and single-nucleotide composition under any folding model.

the 60 sequences used for di-nucleotide shuffling test. Number of sequences with available results slightly vary for various models.

2.6.4 Comparing Structural and Base-pairing Entropies

In order to evaluate the statistical power of structural entropy in characterizing ncRNAs compared to the entropy of their base-pairs, we compared the structural entropy with a formulation of their base-pairing entropy introduced in (Huynen et al., 1997). We refer to this formulation as base-pairing (BP) entropy:

$$\texttt{BP Entropy}(y) = \sum_{i<j} -p(i,j)\log p(i,j)$$

Where $p(i,j)$ refers to the pairing probability of nucleotides positioned at i and j in the given sequence y.

Neither the total pairing entropy introduced in (2.8) nor base-pairing entropy are equal to the actual overall base-pairing entropy of the sequence, since they both ignore base-pair dependencies introduced under the corresponding SCFG model. They both, however, have significantly similar statistics. In this work, we selected base-pairing entropy for making various comparisons with structural entropy. The first and second randomization tests were applied for base-pairing entropy in the same manner as the structural entropy.

2.7 Results

Entropy values of collected sequences were calculated under all selected folding models. Impact of various factors such as sequence length and choice of grammar model were investigated. Also, the folding entropy p-values of sequences were calculated and organized according to their class in various randomization tests to investigate the significance of folding entropy of classes of ncRNAs and their relationship to various folding models.

Figure 2.1 shows the distribution of entropy values of all collected 4116 sequences with respect to sequence length, under each model. The left-most derivation entropy of RND1 and RND10 models greatly reduces folding diversity compared to redundant derivation of entropy, with their average value for collected sequences being reduced from 132 and 105 to 116 and 85, respectively. Structural entropy and sequence length have a linear relationship regardless of the choice of folding model for the range of tested sequences $\{60\texttt{nt} - 600\texttt{nt}\}$[5]. This is also true for redundant derivation of structural entropy for the structurally am-

[5] Simulation results suggest higher-order relationship between sequence length and its structural entropy for length values higher than 600$\texttt{nt}$ (data not shown).

biguous models RND1 and RND10 (data not shown). The structural entropy across all grammars shows that the BJK grammar has a significantly more deterministic folding space compared to other grammars. The RND1 grammar shows significantly higher folding diversity than other grammars including RND10. This indicates that the ratio of terminal to nonterminal probabilities has a major effect on the structural entropy. The top row of figure 2.1 shows consistently high entropy for rfamv5-trained parameter sets compared to those for benchmark and mixed80 regardless of choice of grammar.

2.7.1 Significance of Folding Entropy of non-coding RNAs

The folding entropy significance of every ncRNA sequence was empirically calculated by comparing it to entropy values corresponding to ensemble of random sequences of same length and single-nucleotide composition (first randomization). The significance of folding entropy of each ncRNA was also calculated by comparing it to an ensemble of random sequences with same length and same di-nucleotide composition (second randomization). The above tests were applied under each model, separately. Figure 2.2 shows the structural entropy p-value distribution of classes of ncRNAs under the RUN (benchmark) model. We used this model for further examination, since the p-values obtained for all ncRNAs have a fairly uniform distribution [6] (see A.4). As we can see, various classes of ncRNAs have different p-value distribution. For instance, miRNA has a left-tilted distribution with 70% of its sequences having significantly low folding entropy (p-value less than or equal to 0.05). Bacterial RNase P type B sequences, on the other hand, have a right-tilted p-value distribution with 33% of sequences with significantly high structural entropy (p-value greater than or equal to 0.95). The distributions of p-values of other classes of ncRNAs are also different from one another.

As mentioned before, p-values obtained from steps 1 and 2 randomization tests are expected to be independent of sequence length making possible comparisons of folding entropy between sequences of different length. A qualitative inspection of p-values of various classes of ncRNAs with respect to their length also confirms this assumption (See A.5); i.e., p-values of some sequences belonging to a ncRNA class do not seem to be a function of length.

[6] Even though p-values are observed to be uniformly distributed across sequences, the inter-family p-value distribution distance is not maximal for RUN (benchmark) (See tables A.7 and A.8 for sum of pair-wise Kolmogorov-Smirnov (KS) distance corresponding to RNA families for each model)

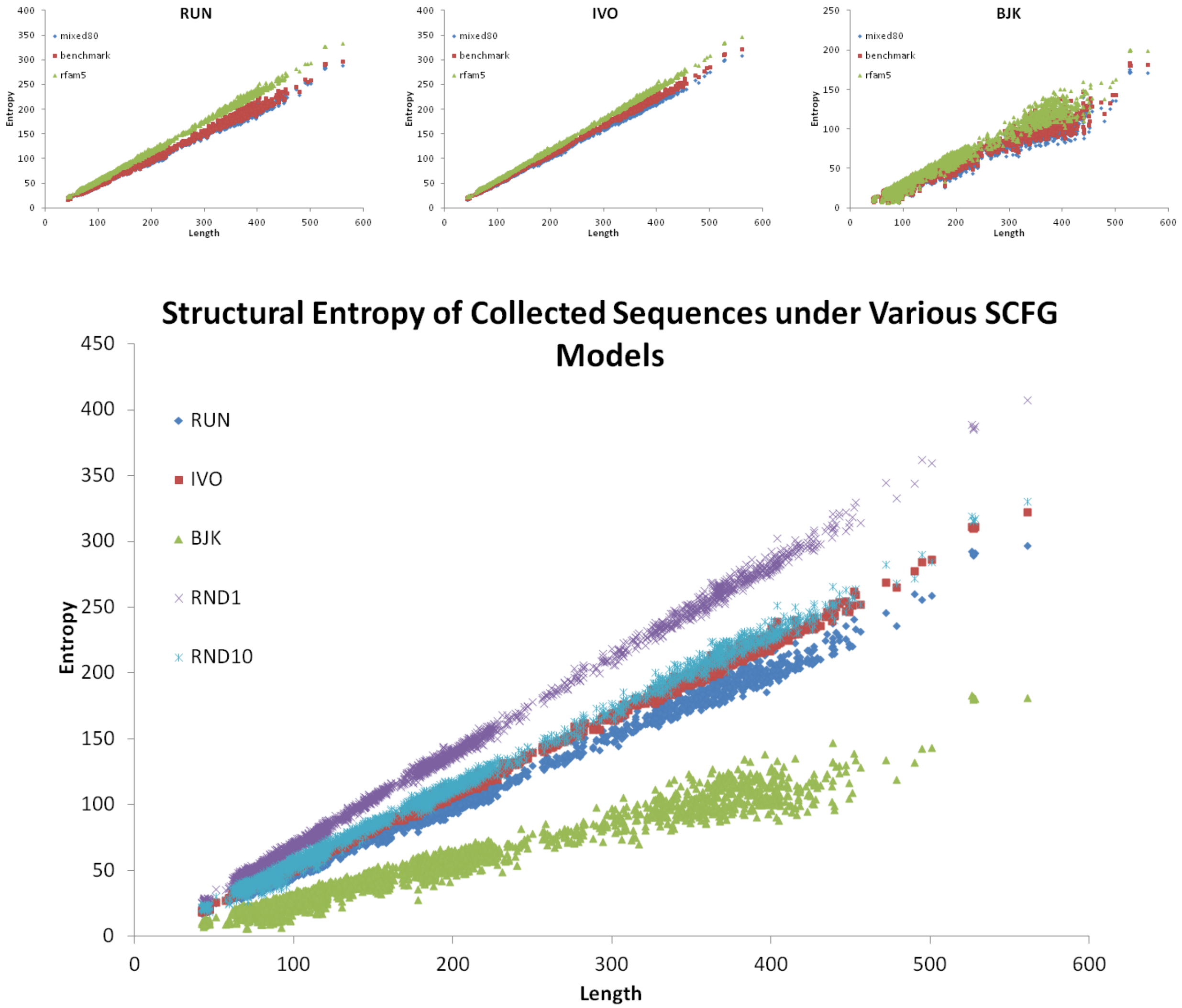

Figure 2.1: Entropy vs. Length. Structural entropy of all 4116 collected sequences with respect to their lengths. The top row corresponds to RUN, IVO, and BJK grammars, respectively. Values obtained from various model parameter sets are plotted for each grammar. Parameters are according to (Dowell and Eddy, 2004). The bottom graph plots structural entropy of sequences under various models. Benchmark-trained model parameters were selected for RUN, IVO, and BJK grammars. Values for RND1 and RND10 models correspond to left-most derivation of entropy.

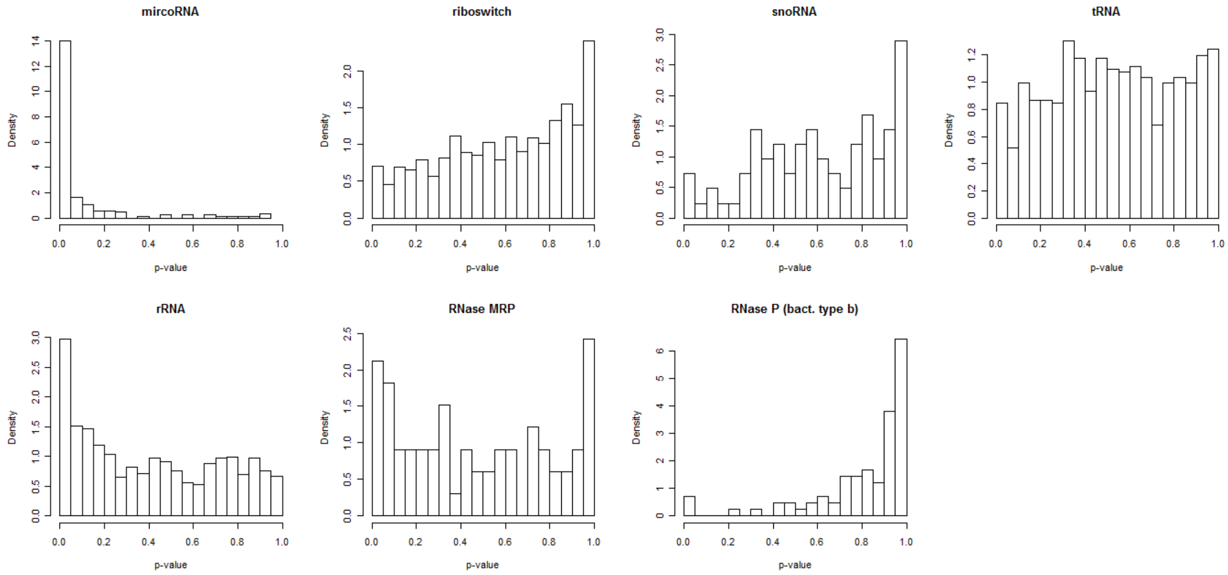

Figure 2.2: Structural Entropy p-Value Distribution. Structural entropy empirical p-values of ncRNAs families against dinucleotide-preserved random shuffles using the RUN (benchmark) model: Dinucleotide shuffling algorithm originally described in (Altschul and Erickson, 1985) and subsequently implemented in (Clote et al., 2005) was used to generate shuffled random sequence ensembles for each individual sequence, separately. Random sequences were of the same length and dinucleotide distribution as the original sequence. The size of the random ensemble was proportional to the original sequence length.

The p-values obtained from first and second randomization tests are also expected to be independent of single- and di-nucleotide compositions, respectively. However, this is not the case. Table A.9 contains correlation values between di-nucleotide composition and structural entropy p-values of two ncRNAs having most distant average p-values, i.e., miRNA and bacterial type B RNase P. The correlation values of miRNA and bacterial type B RNase P have opposite signs for most nucleotide compositions with values corresponding to `UA`−dinucleotide compositions having most correlation difference. Figure A.6 is a plot between entropy p-values obtained from di-nucleotide shuffling with respect to `UA`−dinucleotide compositions for miRNA and bacterial type B RNase P sequences under the RUN (benchmark) model. Micro RNA sequences of higher `UA`−dinucleotide composition tend to have lower entropy under the di-nucleotide shuffling test while the opposite is true for bacterial type B RNase P sequences. Hence, entropy p-values obtained from di-nucleotide randomization are *not* independent of di-nucleotide composition of that sequence. Furthermore, this dependence can have varying behavior depending on the class of the ncRNA sequence.

The p-value distribution of folding entropy of sequences belonging to the same class of ncRNA has been observed to be more linear than normal in most cases and under most models (data only shown for the RUN (benchmark) model, figure 2.2). The average p-value can be a useful representation of the folding entropy in making comparisons across various classes of ncRNAs. Average p-values of various ncRNA families from first and second randomization tests were observed to have fairly consistent ranking regardless of choice of parameter set under a given the grammar model (Table A.5 contains average p-values of various models for different ncRNA families). Hence, in order to have an overall view of the behavior of each model with respect to folding entropy of various classes of ncRNAs, we first averaged the p-values of the sequences across parameter sets of a given model. We then applied a second level of averaging across all sequences belonging to the same class of ncRNA for each model. Figure 2.3 contains average p-values for all classes of ncRNAs and under different grammars. The left bar-plots contain values corresponding to di-nucleotide shuffling test and the right bar-plots contains values corresponding to single-nucleotide composition randomization. The choice of grammar design has a high impact on structural entropy p-values obtained under various randomization tests. BJK and RND10 models yield lower p-values while the IVO grammar shows high average p-values. The ordering of average p-values of ncRNA families is fairly consistent between single- and di-nucleotide composition randomization tests under a given folding model. Furthermore, average folding entropy p-values of classes of ncRNAs are different from one another. The ranking of the average structural entropy p-values of ncRNA families, however, are consistent across most models. Micro RNAs have the lowest average p-values under all models and randomizations excluding results for di-nucleotide shuffling test under the IVO grammar model. Folding entropy of the miRNA is also significantly lower than that of a random sequence with same length and nucleotide composition under most models. On the other hand, results of RND10 and RUN models assign highest structural entropy average p-values to bacterial type B RNase P sequences under both single- and di-nucleotide randomization tests. SnoRNAs and riboswitches also show high average p-values than tRNA, rRNA, and RNase MRP under the same models. The BJK grammar shows slightly different results in this regard with snoRNA having highest p-value average than other ncRNAs.

The same statistical test was performed to assess the significance of base-pairing entropy of ncRNAs and make comparisons with structural entropy. Figure 2.4 is a bar plot of the corresponding base-pairing entropy p-values for each grammar design. Similar to results of the structural entropy statistics, the base-

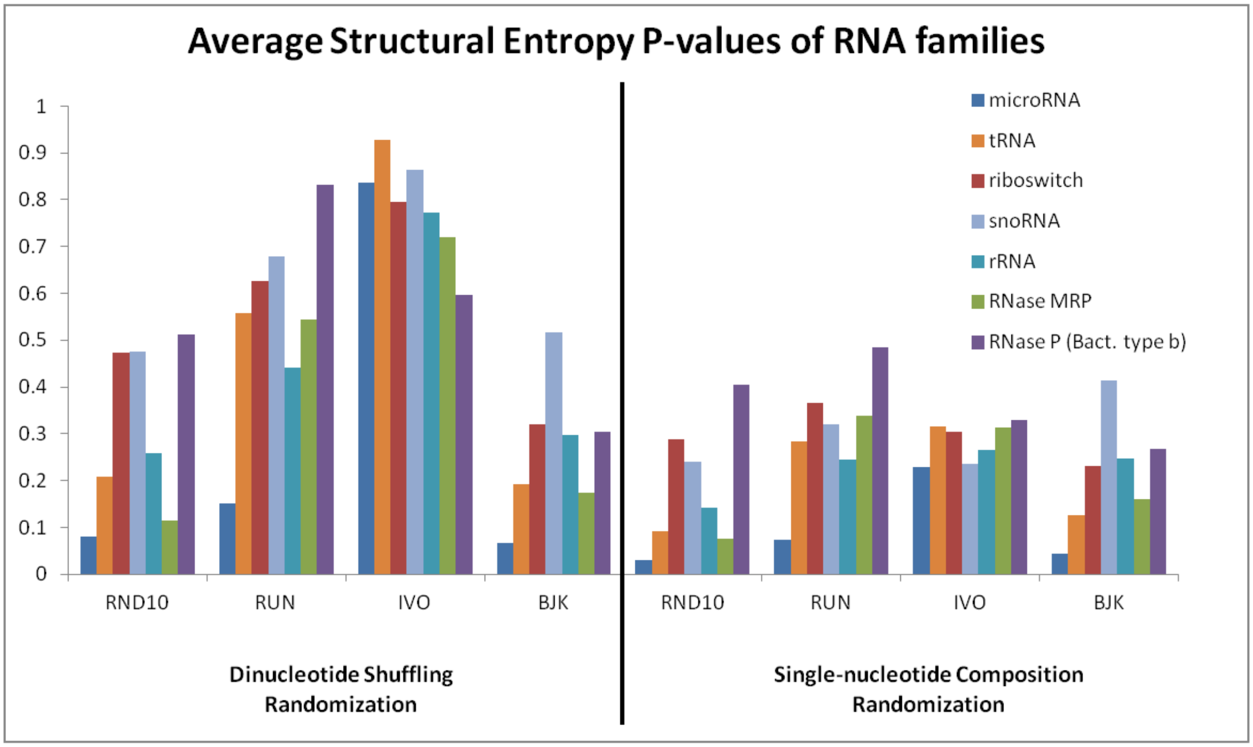

Figure 2.3: Structural Entropy p-Value Average. Average structural entropy p-values of ncRNA families under different models and randomization tests. P-values were averaged across results from three different training sets (benchmark, mixed80, and rfam5) for each individual model. The four sets of plots on the left are average p-values against dinucleotide-preserved random shuffles (Altschul and Erickson, 1985). The four sets of plots on the right are average p-values against single nucleotide-preserved random sequences using GenRGenS Software (Ponty et al., 2006). Independent random ensembles were generated for each individual sequence, separately. The random ensemble size was chosen proportional to sequence length. Sequences containing ambiguous nucleotides were eliminated. Sequences used are miRNA (163 sequences), riboswitch (1359 sequences), RNase (54 sequences), rRNAs (926 sequences), tRNA (966 sequences), snoRNA (82 sequences), and bacterial type B RNase P (34 sequences).

pairing entropy p-values of miRNAs are significantly lower than that of other classes of ncRNAs. The p-values of miRNAs are also significantly lower than random sequences of same length and nucleotide compositions under most models. Also, bacterial type B RNase P sequences have higher folding entropy average p-values than other classes of ncRNAs. SnoRNAs and riboswitches have similar results to the structural entropy statistics. Slight difference of ranking of classes of ncRNAs between structural entropy and base-pairing entropy is observed. Furthermore, p-values obtained from structural entropy tend to be generally higher than those for base-pair entropy especially under the di-nucleotide shuffling randomization test. For instance, average structural entropy p-value of the bacterial type B RNase P under the RUN model is 0.83 while its average base-pairing entropy p-value is 0.62 under the same model.

In order to further investigate the statistically high structural entropy of bacterial type B RNase P, we performed the dinucleotide randomization test on all 114 bacterial type B sequences along with all 117 nuclear and all 306 bacterial type A RNase P sequences. The RUN (benchmark) folding model was used. Figure 2.5 shows that the distribution of structural entropy p-values of bacterial type B RNase P is significantly tilted to the right (31% of sequences with p-value higher than or equal to 0.95) compared to those for bacterial type A RNase P and nuclear RNase P.

The collected riboswitches (1365 sequences) contain several sub-families. We calculated the average p-values for each sub-family under the RND10, RUN, and BJK models to evaluate sub-family specific folding entropy behavior of riboswitch sequences under the above models. Benchmark-trained parameter sets were used arbitrarily for the BJK and RUN models. Values for other parameter sets vary slightly. Figure 2.6 contains structural entropy average p-values of riboswitch sub-families under various folding models. Results suggest that modeling has a great impact on folding entropy significance of riboswitch sequences. The RUN (benchmark) model generally assigns higher p-values to sequences while the BJK (benchmark) model assigns lower p-values. This is true for most cases of riboswitch sub-families. The sub-class of the riboswitch also has an impact on structural entropy p-values. Average p-values of various folding models are closer for certain classes of riboswitch, such as Hammerhead, while they may drastically vary for certain other riboswitches, for example rli62. Figure 2.7 shows the structural entropy p-values of ncRNAs empirically calculated against random sequences with structure (third randomization). The top row corresponds to cluster of high GC−composition sequences (cluster 1), the middle row corresponds to low GC−composition sequences (cluster 2), and the bottom row corresponds to average GC−composition sequences (cluster 3)

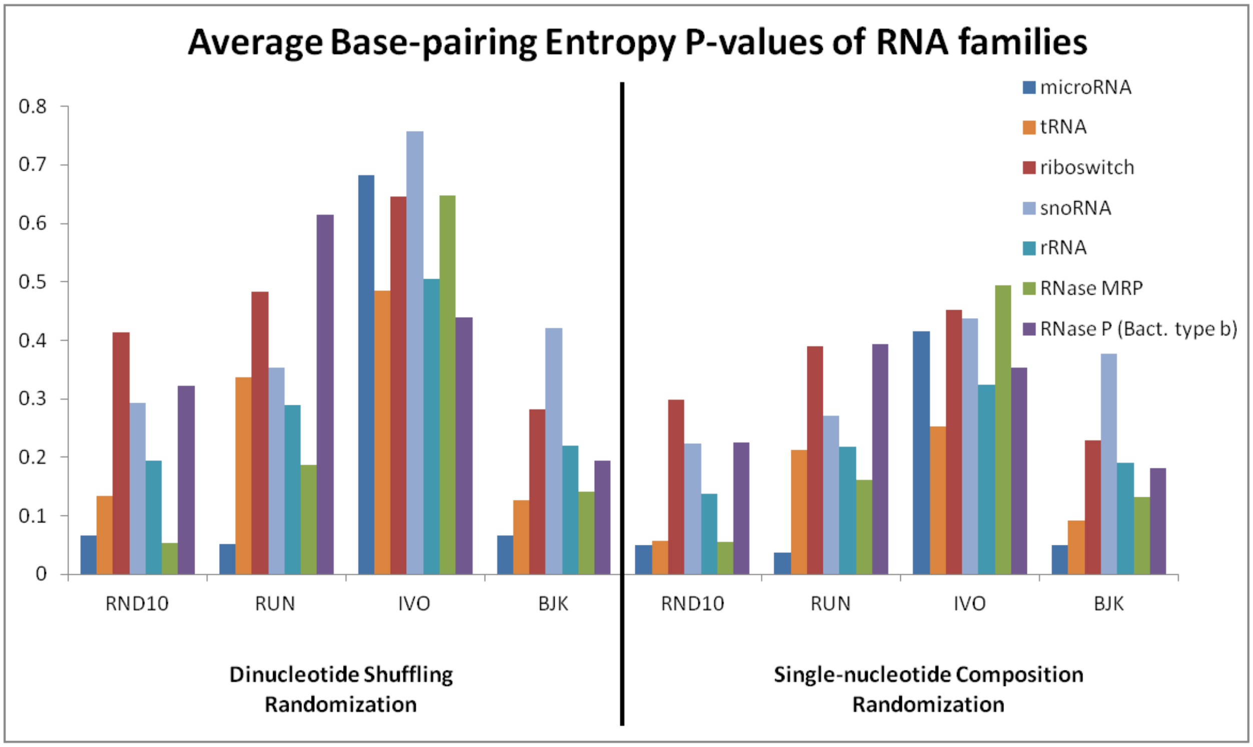

Figure 2.4: Base-pairing Entropy p-Value Average. Average base-pairing entropy p-values of ncRNA families under different models and randomization tests. P-values were averaged across results from three different training sets (benchmark, mixed80, and rfam5) for each individual model. The four sets of plots on the left are average p-values against dinucleotide-preserved random shuffles (Altschul and Erickson, 1985). The four sets of plots on the right are average p-values against single nucleotide-preserved random sequences using GenRGenS Software (Ponty et al., 2006). Independent random ensembles were generated for each individual sequence, separately. The random ensemble size was chosen proportional to sequence length. Sequences containing ambiguous nucleotides were eliminated. Sequences used are miRNA (163 sequences), riboswitch (1359 sequences), RNase (54 sequences), rRNAs (926 sequences), tRNA (966 sequences), snoRNA (82 sequences), and bacterial type B RNase P (34 sequences).

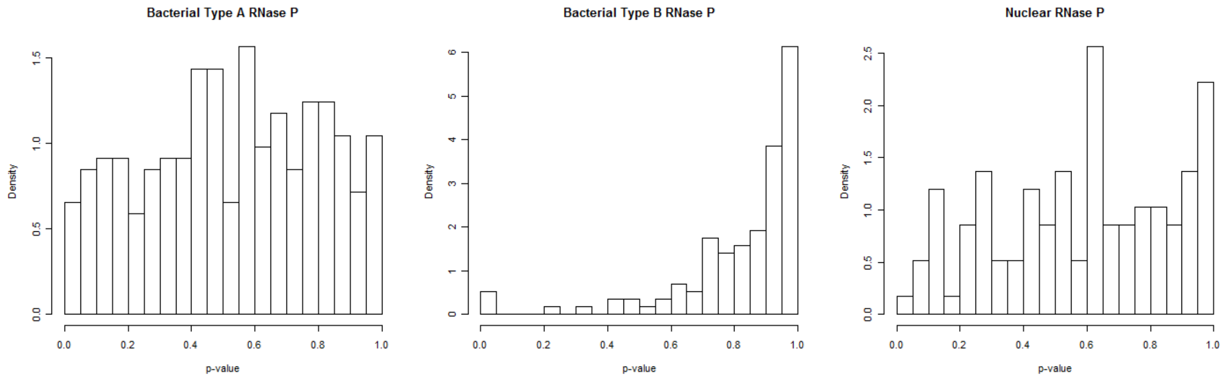

Figure 2.5: Structural Entropy p-Values of RNase P. Structural entropy p-values of bacterial type A RNase P (306 sequences), bacterial type B RNase P (114 sequences), and nuclear RNase P (117 sequences). P-values were obtained from dinucleotide shuffling of sequences and under the RUN (benchmark) folding space model.

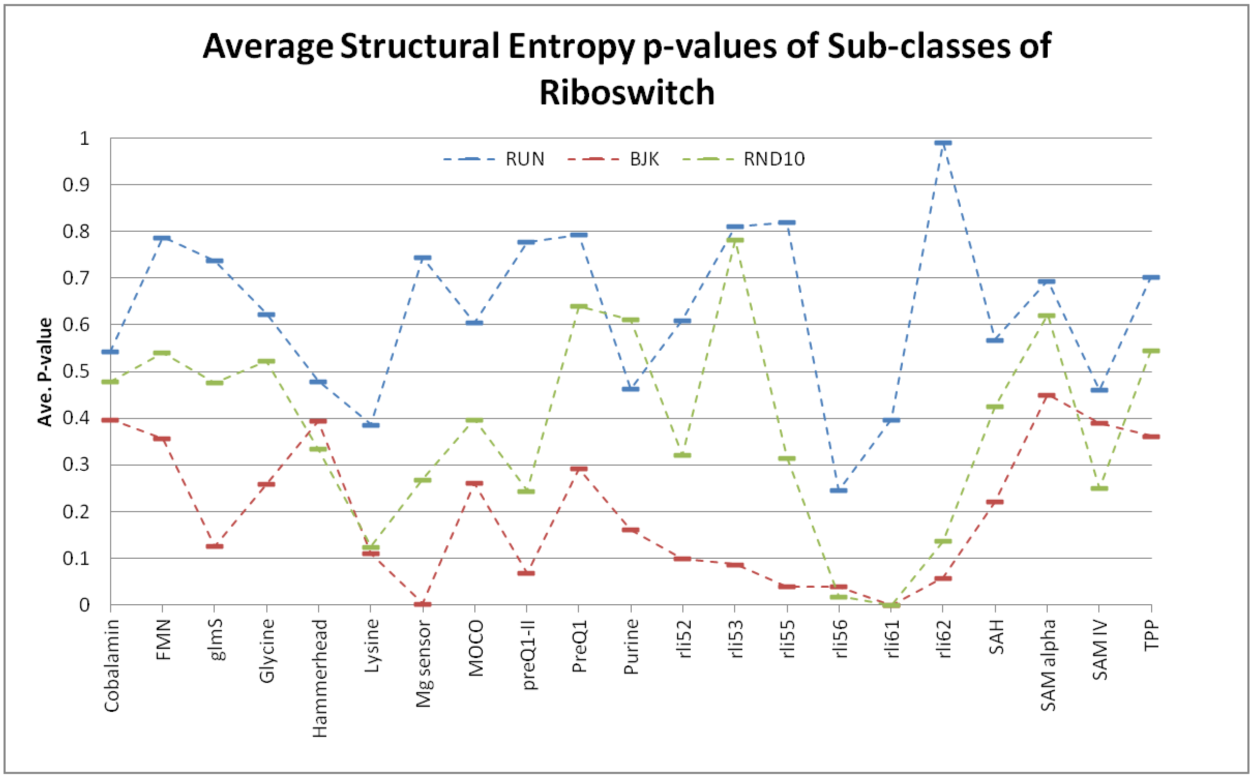

Figure 2.6: Structural Entropy p-Values of Riboswitches. Average structural entropy p-values of riboswitch sub-families under RUN (benchmark), BJK (benchmark), and RND10 models using di-nucleotide shuffling randomization test (Altschul and Erickson, 1985). The numbers of sequences in each sub-family are: Cobalamin (431), FMN (146), glmS (17), Glycine (44), Hammerhead 1 (30), Lysine (47), Mg sensor (4), MOCO RNA motif (179), PreQ1 (42), preQ1-II (14), Purine (133), rli52 (6), rli53 (5), rli55 (3), rli56 (6), rli61 (4), rli62 (2), SAH riboswitch (52), SAM alpha (40), SAM IV (40), TPP (115). Results for rli54 (5) is not available. Please refer to table A.1 for information about sequences.

(See A.5 for details about clusters and sequences). Results for the IVO (benchmark) grammar are moved to A.1. A qualitative inspection suggests that entropy values of ncRNAs and their significance vary across models and depending on their GC−composition. By looking at the columns of figure 2.7, we can see that the significance of entropy p-values of classes of ncRNAs varies across models. For instance, the tested tRNAs have significantly lower entropy under the RND10 model compared to other models. Also, by inspecting the rows of figure 2.7, we can see that under certain GC−compositions, various classes of ncRNA are relatively more distinguishable. For instance, riboswitch sequences have higher entropy than micro RNAs where GC−composition is either high or low. Average GC−composition sequences, however, do not show such distinction. The above observation about riboswitch entropy p-values being relatively higher than those for miRNAs is examined further. Table A.4 is a quantitative comparison of the percentage of sequences with significantly high structural entropy from both classes of miRNA and riboswitch across all models. The tested riboswitch sequences have higher folding entropy than miRNAs for most nucleotide compositions and under most of the models. Furthermore, p-values corresponding to the riboswitch sequences are overall significantly higher than expected from a random sequence with structure. Finally, results from the IVO grammar do not suggest any distinction between entropy of various classes of ncRNAs. (See figure A.1 and table A.4)

2.7.2 Model Accuracy and Structural Entropy

In order to gain insight into the impact of folding model accuracy on the structural entropy of a sequence, we selected three sets of results where significant value of folding entropy is observed:

1. Significantly low folding entropy of micro RNA sequences in di-nucleotide shuffling test; Observation of Figure 2.3.

2. Significantly high folding entropy of bacterial type B RNase P sequences in di-nucleotide shuffling test under the RUN folding model; Observation of Figure 2.3.

3. Significantly high folding entropy of the high GC−composition riboswitch sequences of length 93 ± 5 in the random sequence with structure test; Observations of Figure 2.7 and Table A.4.

The folding entropy p-values of miRNA sequences obtained from di-nucleotide shuffling were plotted against sensitivity and specificity folding model to their corresponding annotated secondary structures.

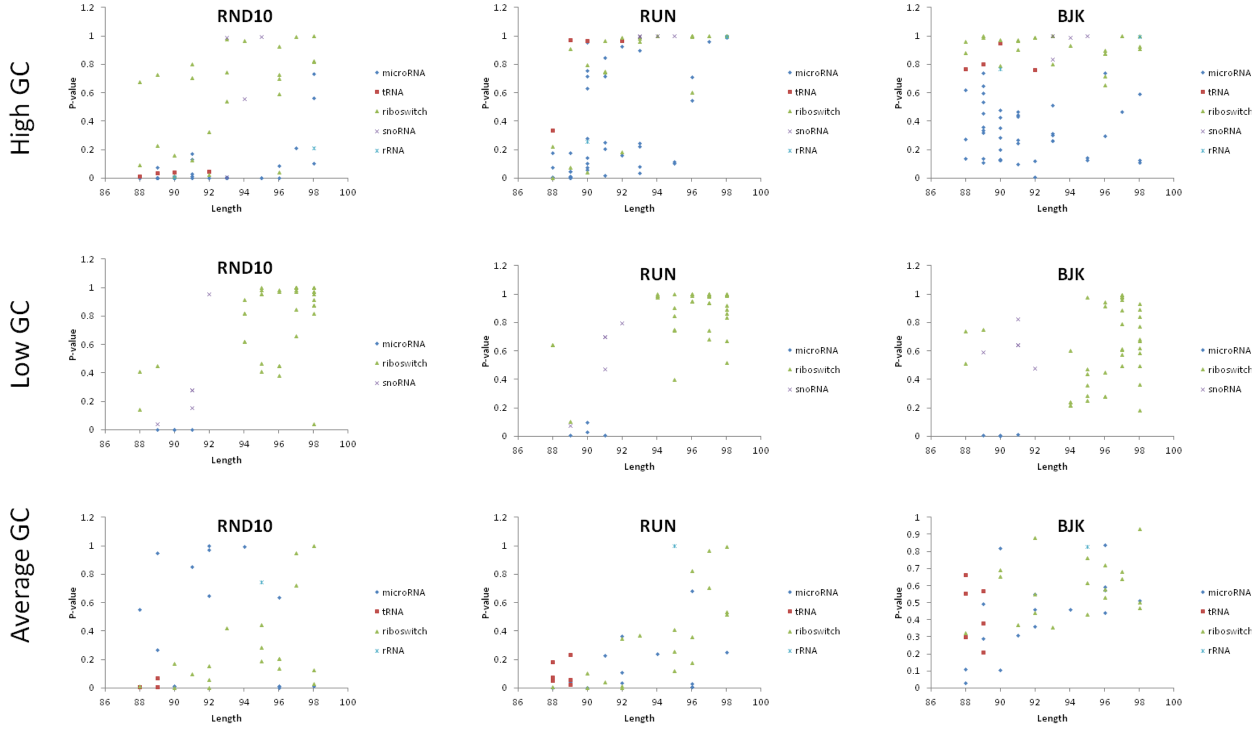

Figure 2.7: Structural Entropy p-Values of Structural Randomization. Structural entropy p-values of short ncRNAs sequences against random sequences with structure (See A.5 for details about clusters and sequences). Benchmark-trained parameter sets were arbitrary selected for RUN and BJK grammars. Other parameter sets yield similar results (See table A.4).

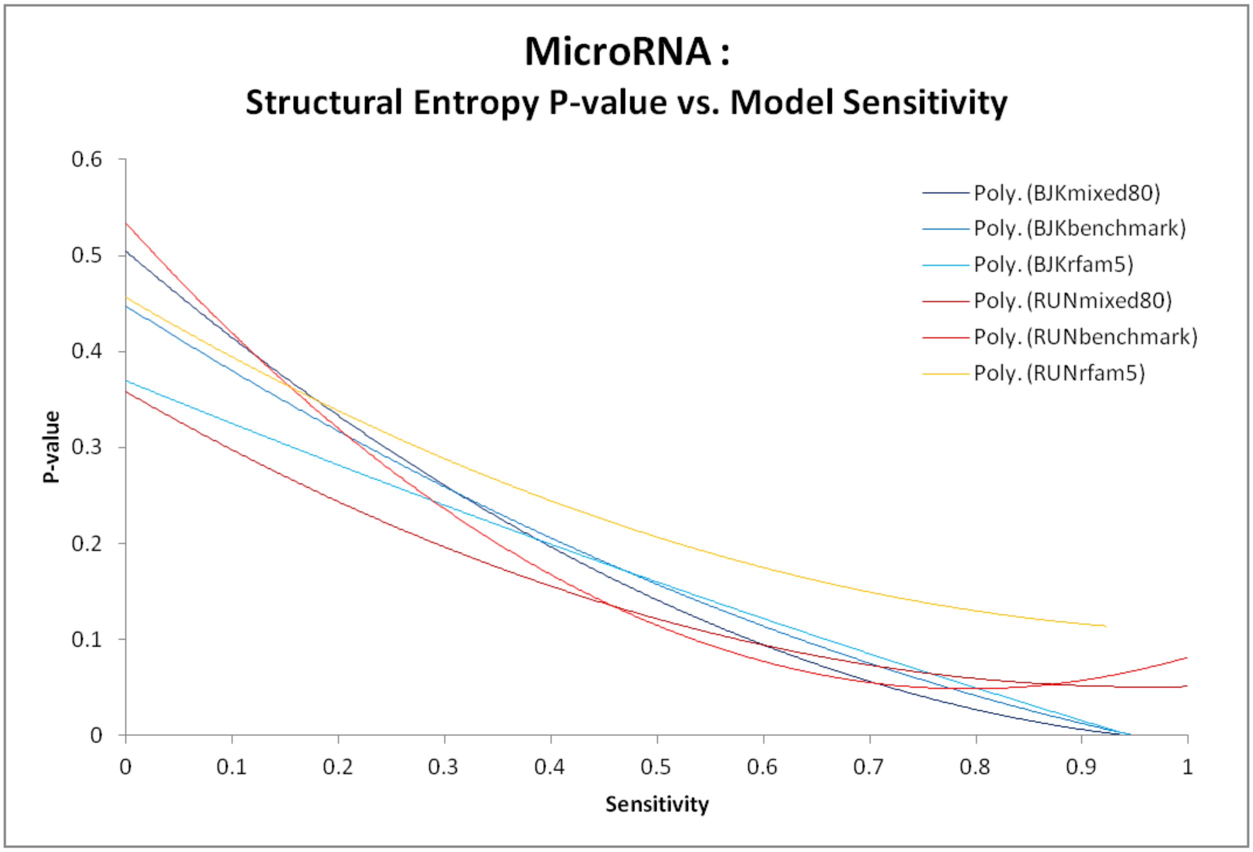

Figure 2.8: Structural Entropy p-Values of miRNAs vs. Model Sensitivity. Structural entropy p-values of miRNA sequences against folding model sensitivity to their secondary structure. Di-nucleotide shuffling was used to calculate p-values. 2-order polynomial trendline of p-values are shown for each grammar model.

Figures 2.8 and 2.9 show the relationship of structural entropy p-values of miRNA sequences to model sensitivity and specificity to annotated secondary structure, under various models. Results from IVO grammar is excluded due to the small range of available sensitivity values. Figure 2.8 shows that lower folding entropy p-value and model sensitivity to miRNA annotated structure are correlated with each other regardless of choice of model. Similar observation is made with respect to model specificity (See 2.9). The relationship between folding entropy p-value and model accuracy is slightly stronger under the BJK model than for the RUN model, regardless of choice of parameter set.

We also plotted the structural entropy p-values of the bacterial type B RNase P sequences obtained from di-nucleotide shuffling against model sensitivity and specificity to annotated secondary structure. Figure A.7

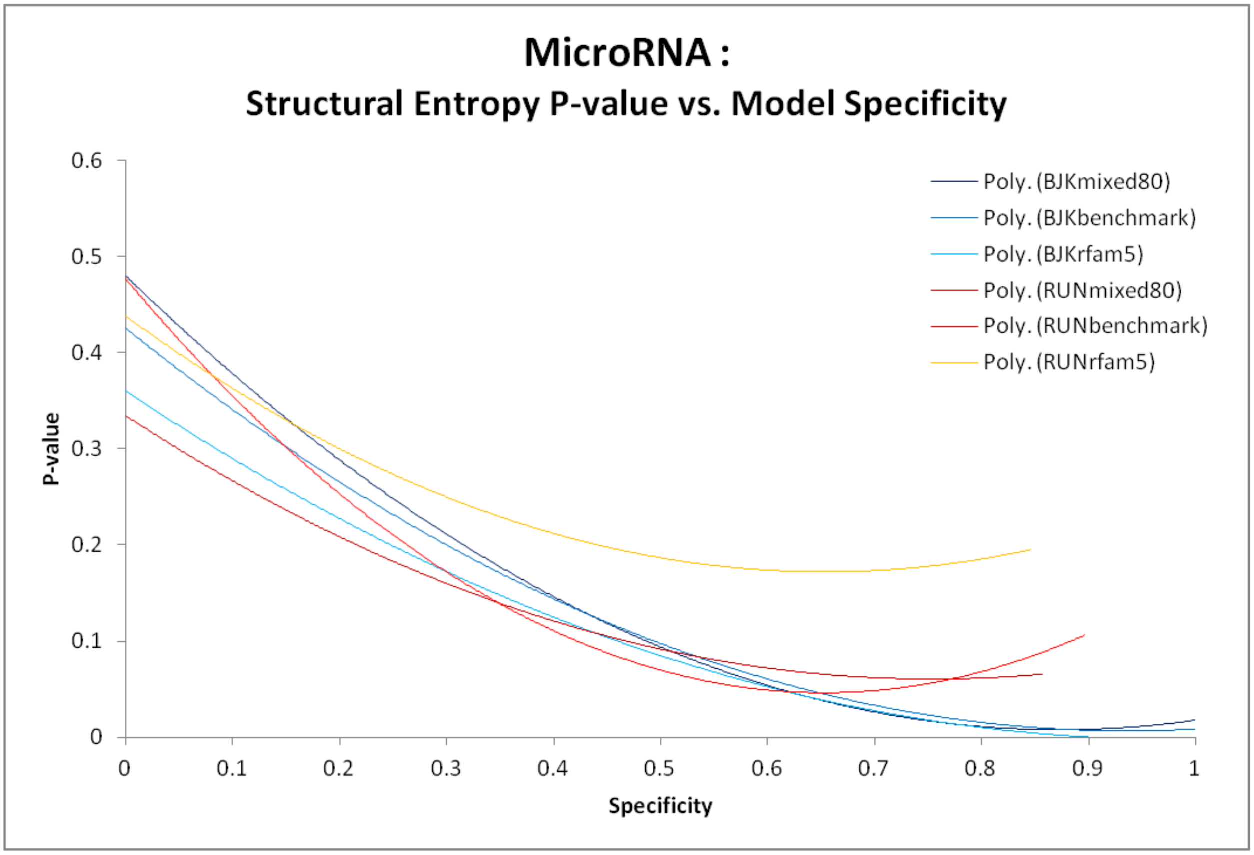

Figure 2.9: Structural Entropy p-Values of miRNAs vs. Model Specificity. Structural entropy p-values of miRNA sequences against folding model specificity to their secondary structure. Di-nucleotide shuffling was used to calculate p-values. 2-order polynomial trendline of p-values are shown for each grammar model.

plots p-values against model sensitivity. Similar to results from miRNA, the trendline suggests that higher entropy p-values are associated with lower sensitivity. This relationship is, however, less significant under the RUN model, due to lower range of available sensitivity values. Assessing relationship between p-values and model specificity was not possible, due to the unavailability of a reasonable range of specificity values.

Figures 2.10 and 2.11 plot structural entropy p-values of the riboswitch sequences against model sensitivity and specificity, respectively. Riboswitch sequences here have high GC−composition and belong to cluster 1. (For results of riboswitches in clusters 2 and 3, see A.8, A.9, A.10, and A.11). In contrary to results obtained from di-nucleotide shuffling about the miRNA, low folding entropy p-values do not seem to be associated with high model accuracy to annotated secondary structure. Results from the BJK model, suggest that significantly high structural entropy p-values of the riboswitch are *not* associated with lower model accuracy. Results from riboswitches in clusters 2 and 3 also suggest independence of folding entropy from model sensitivity to annotated secondary structure under the BJK grammar model. Results from the RUN grammar model are unclear, due to lower range of available sensitivity values.

2.8 Discussion

2.8.1 Discerning Structural Ambiguity

Shannon's entropy of SCFG ensembles on a given sequence was derived and implemented for both structurally ambiguous and unambiguous grammars. A.1 offers the modifications needed for generalization of structural entropy for structurally ambiguous grammars. The modifications insure non-redundant counts of derivation trees in calculating the structural entropy compared to the redundant derivation. Implementation of left-most derivation entropy can be used to calculate the entropy of any folding model, so long as it can be mirrored to a non-stacking SCFG, regardless of its ambiguity.

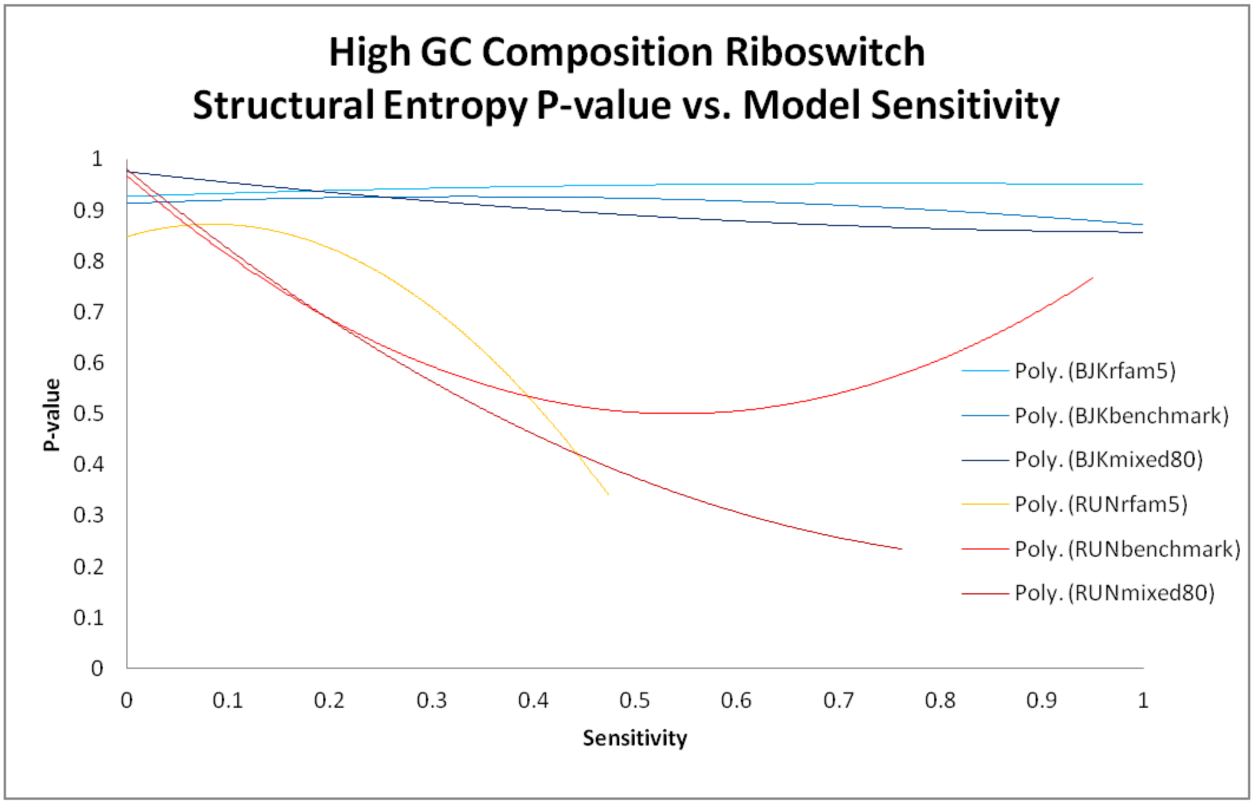

Figure 2.10: Structural Entropy p-Values of Riboswitches vs. Model Sensitivity. Structural entropy p-values of high-GC composition riboswitch sequences of length 93 ± 5 against folding model sensitivity to their secondary structure. Riboswitch sequences belong to cluster 1 (See A.5 for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See A.4 for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

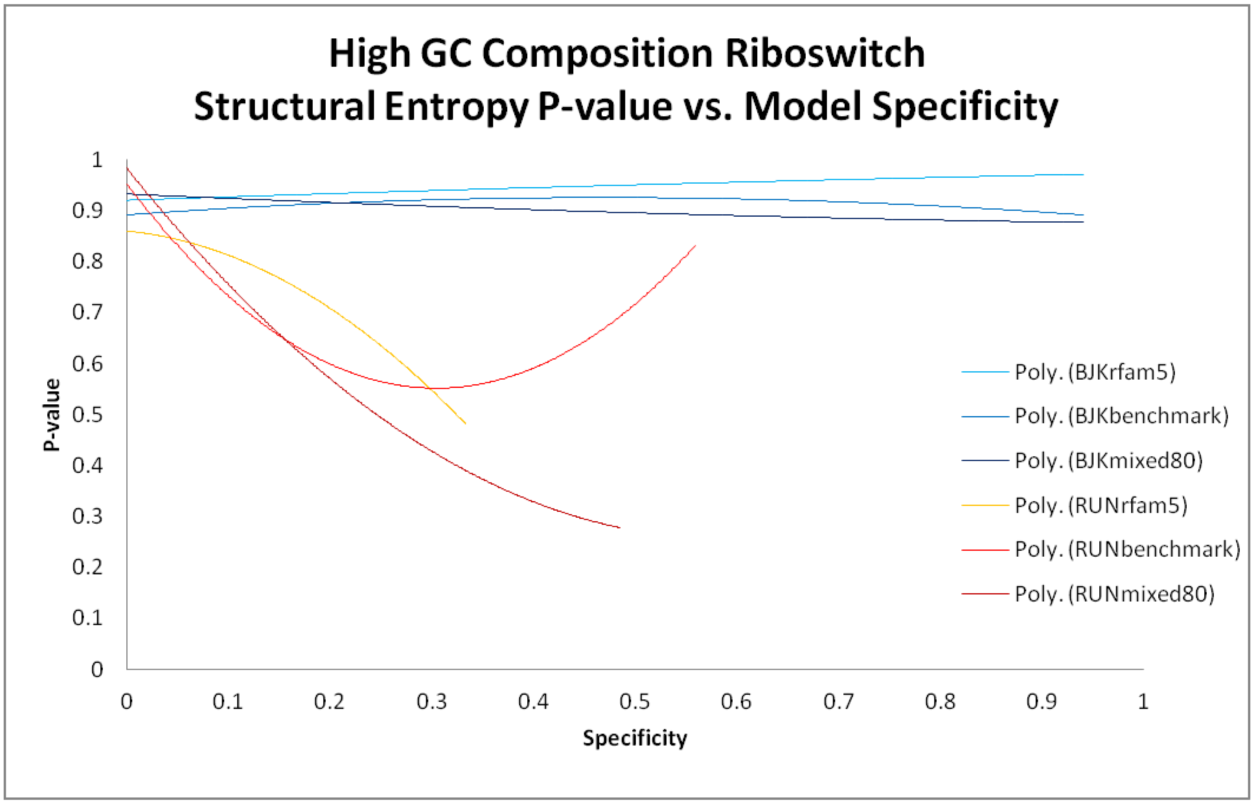

Figure 2.11: Structural Entropy p-Values of Riboswitches vs. Model Specificity. Structural entropy p-values of high-GC composition riboswitch sequences of length 93 ± 5 against folding model specificity to their secondary structure. Riboswitch sequences belong to cluster 1 (See A.5 for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See A.4 for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

Both the redundant derivation and the left-most derivation entropy must yield the same entropy value on any sequence as long as the folding model is structurally unambiguous. For a structurally ambiguous model, however, the redundant derivation entropy must be higher than or equal to that of the left-most derivation. Our results were in line with the above claims (data not shown). Hence, the two implementations can be used to discern the structural ambiguity of a model in a given set of RNA sequences.

2.8.2 Dominating Factors in Entropy Calculation

Factors such as length of the sequence, choice of CFG design, structural ambiguity of the model and terminal-to-nonterminal probability ratio have a much more significant impact on the entropy value than other factors such as CYK-trained model parameters, nucleotide composition, and class of the ncRNA sequence under investigation (See figure 2.1).

2.8.3 The Significance of Folding Entropy of ncRNAs

A qualitative inspection of figures 2.3 and 2.7 is sufficient to reject the null-hypotheses in all three randomization tests; both the choice of model and the class of the ncRNA can have an impact on the folding entropy causing it to be significantly low or high in certain cases. Co-association of model accuracy and entropy p-values, however, were observed under first and second randomization tests. Comparing average sensitivity and specificity values in A.3 to average entropy p-values in 2.3 shows that this co-association exists across classes of sequences and choices of models. The impacts of choice of model and class of the ncRNA on the structural entropy of the sequence are separately analyzed in the following:

Impact of Choice of Model

Various CFG designs yield different structural entropy p-value distributions for ncRNAs under both the single- and di-nucleotide composition randomization tests (See A.3 and A.4). ncRNA structural entropy p-values under the BJK and RND10 folding models are tilted to the left with average p-values less than 0.5 while p-values corresponding to the IVO grammar are tilted towards right with an average higher than 0.5, under the di-nucleotide shuffling test. Furthermore, comparison of the rankings of average structural entropy p-values of various classes of ncRNAs to their corresponding average model sensitivity and specificity shows consistency between the two (See tables A.3 and A.5). Hence, high model accuracy and low folding entropy are generally associated with each other under the single- and di-nucleotide randomization test.

A qualitative observation of results from the third randomization test, also suggests that modeling can affect the significance of folding entropy. The RND10 model shows a slight distinction between various classes of ncRNAs, namely miRNA and riboswitch. Results from the RUN and BJK grammar are also consistent with that of RND10. Entropy p-values corresponding to the IVO model, which is relatively less accurate than RUN and BJK grammar models, are not as suggestive.

Impact of Class of ncRNA

The average structural entropy p-values obtained from single- and di-nucleotide randomization tests show a somewhat consistent ranking across most models (See figure 2.3 and tables A.5 and A.6). The third randomization test also suggests association of entropy p-value with the class of the ncRNA. Nucleotide composition also has an impact on the significance of folding entropy of various classes of ncRNAs under the third randomization test, regardless of choice of model.

Structural entropy is not necessarily similar across sub-families of the same class of ncRNA. Figure 2.5 shows that entropy p-values obtained in di-nucleotide randomization test have different distributions across various sub-families of the RNase P sequences. Sub-families of the riboswitch sequences also have different entropy p-values in the same randomization test (See 2.6 for average p-values of riboswitch sub-families under various folding models).

2.8.4 Micro RNA Has Low Secondary-Structural Entropy

Micro RNA structural entropy p-values obtained from the third randomization are not significantly higher than random sequences with structure under most folding models (See figure 2.7 and table A.4). This implies that the entropy of miRNA sequences is very typical of that of a sequence with a single structure.

Furthermore, both single- and di-nucleotide composition randomizations yield lower entropy p-values for miRNA sequences than other classes of ncRNAs under all models, excluding the IVO grammar. The P-values of most miRNA sequences are also significantly lower than random sequences of same length and single- and di-nucleotide composition under various models. This significance reaches its maximum in the single-nucleotide composition randomization test and under the un-trained structurally ambiguous RND10 model, with its average being 0.029 (89% of sequences having values less than or equal to 0.05). This implies that perturbing the nucleotide arrangement of the miRNA sequence significantly increases the uncertainty about its secondary structural fold. A closer look at individual p-values of sequences and the accuracy of their corresponding model to annotated secondary structure shows that more accurate modeling of the secondary structural conformation leads lower folding entropy, further distinguishing the miRNA from a random background (See figures 2.8 and 2.9). The low folding entropy results of the miRNA are also in line with results of the base-pairing entropy test presented here, 2.4, and previous findings about the more deterministic folding behavior of miRNA. Our overall conclusion about the miRNA results is that not only do these sequences tend to have low secondary structural entropy, but also the annotated secondary structure seems sufficient to characterize this class of ncRNA.

2.8.5 Unexpectedly High Folding Entropy Observed

As mentioned before, entropy captures the (un)certainty of a probabilistic model. A low entropy value could imply more deterministic behavior, while a high entropy value describes more diversity in a given distribution, depending on the application. The observation of significantly high entropy, more than what is expected from a random event, is an unintuitive but theoretically possible observation. An example of significantly high structural entropy is offered in A.9.3 where the entropy p-value of a hypothetical miRNA can be as high as 1 in di-nucleotide randomization and under an arbitrary single stem-loop SCFG model. The interested reader in this regard is also referred to A.9.2 for a more mathematical justification.

The prior intention of this work, and the design of the methodology, was in favor of a one-tailed test; i.e., significantly low entropy could imply a more deterministic folding scenario, while anything else is random. Our results, however, show signs of significantly higher folding entropy than expected from a random sequence. The following three cases were observed:

1. High folding entropy of miRNA in di-nucleotide shuffling test under the IVO model: A re-evaluation of results from figure 2.3 with taking into account the knowledge about miRNA folding, shows that inaccurate modeling coupled with di-nucleotide shuffling randomization test could lead to contradictory results.
2. High folding entropy of bacterial type B RNase P in di-nucleotide shuffling test under the RUN model (Figure 2.3). 31% of sequences of all sequences in this class of ncRNA had p-values higher than or equal to 0.95). The RUN model is less accurate on the annotated structure of the RNase P compared to that of miRNA and hence, higher p-values of the RNase P are generally associated with lower model accuracy. At this point, the reason for high folding entropy of RNase P is unclear. The two following scenarios are possible:
a. *Scenario 1:* High entropy could be due to lower model accuracy, since high p-values of the di-nucleotide shuffling test are generally associated with low model accuracy. A plot of p-values of RNase P against model accuracy is also suggestive in this regard (See A.7). Furthermore, poor modeling of the RNA secondary structure can be very misleading in assessing the structural diversity especially under di-nucleotide shuffling, since miRNA sequences have higher p-values than other sequences under the inaccurate IVO model (See 2.3).
b. *Scenario 2:* High entropy could imply high folding diversity in this class of RNase P. None of the average p-values obtained the performed single- and di-shuffling randomization test, including the untrained models suggest low entropy for the bacterial type B RNase P. In other words, perturbing the nucleotide arrangement of this class of ncRNA does not significantly increase the uncertainty about its folding, within the limits of our models. Furthermore, the plot of A.7 shows that p-values of bacterial type B RNase P that correspond to sequences whose model sensitivity is close to zero, are higher than that of miRNAs under the same model (Comparing results of the RUN grammar of figure A.7 to 2.8). Model sensitivity may not be sufficient to explain high p-values of RNase P. Finally, the same di-nucleotide frequencies that are associated with low folding entropy p-values for the miRNA are associated with high p-values for the bacterial type B RNase P (See table A.9 and figure A.6). This can mean that various primary-structural motifs, possibly higher order nucleotide frequencies, residing in such RNase P sequences may be the cause of its high structural entropy p-values. We have not found independent evidence regarding high structural entropy for the RNase P.

Our overall conclusion about high p-values of folding entropy of the bacterial type B RNase P and other classes in the di-nucleotide shuffling test is that this test may not be suitable for investigating the folding

diversity of an RNA sequence segment. We believe that the di-nucleotide shuffling test and other primary-structure randomization/perturbation tests are more suitable for a one-tailed test; i.e., perturbing the primary structure is expected to disturb the fairly deterministic secondary structure of the sequence such as that of the miRNA, but its effect on the folding distribution of an RNA with a diverse secondary structural space is unclear if not confusing. This could be partly due to the fact that SCFG models are *non-linear* models, meaning that one-sided generation of sequences is not possible while di-nucleotide shuffling randomization and other primary-structure tests are usually linear procedures, meaning that statistics needed to perform randomization can be derived from one-sided observation of the sequences; for example counting di-nucleotide frequency can be done from one-side but counting base-pairs cannot.
3. High entropy p-values of certain riboswitch sequences in the third randomization; Random sequences with structure: Certain riboswitch sequences of length 93 ± 5, especially high GC−composition sequences were observed to have significantly higher folding entropy values than both miRNAs and random sequences with single structure. This was true for the three models RUN, BJK, and the untrained RND10 model. 71% of high GC−composition sequences had p-values higher than or equal to 0.95 under the BJK (benchmark) model. Results from IVO and RND1 models have higher significance but a qualitative inspection of figures 2.7 and A.1 suggests that high p-values of RND10 and IVO are more associated with the length of the sequence rather than what class it belongs to. Plots of p-values against model sensitivity and specificity, figures 2.10 and 2.11, show that high structural entropy p-value of the high GC−composition riboswitch sequences *are* independent of model accuracy to annotated secondary structure of these sequences under the BJK model. We consider this a significant observation, since clustering was performed regardless of which family the sequences belong to. On one hand, BJK model being relatively more accurate than other models, assigns higher folding entropy to selected riboswitches and better distinguishes them from miRNAs of the same cluster. On the other hand, the three BJK models, suggest that accuracy to annotated secondary structure and significantly high folding entropy are unrelated. This means that high p-values of selected riboswitches are not only unrelated to model inaccuracy, but they are also unrelated to annotated secondary structures of these sequences.

Our overall conclusion about the selected riboswitch sequences is that their high folding diversity is more due to nucleotide arrangements intrinsic to these sequences and less due to their annotated secondary structures. As we know, riboswitches can have alternative conformations *vivo*. At this point, it is not clear whether high folding entropy is specifically related to this folding dynamic feature of riboswitches or not. A thorough and comprehensive examination of riboswitches is needed in this regard.

2.8.6 Structural and Base-pairing Entropies

The overall statistical power of structural entropy and base-pairing entropy are similar in distinguishing miRNA sequences from random shuffles (See figures 2.3 and 2.4, and tables A.5 and A.6). Average structural entropy p-values of classes of ncRNAs obtained from di-nucleotide randomization test are slightly more distant from one another than that of the base-pairing entropy statistics, under various models. This is a confirmation on structural entropy being generally more *informative* than base-pairing entropy as expected (2.9). Conclusions derived from the di-nucleotide randomization test are, however, subject to flaws of this test, as previously discussed.

2.8.7 SCFG Modeling of Non-coding RNA Sequences

Grammar Design

In characterizing ncRNA sequences, the performance of structurally unambiguous models trained to predict annotated secondary structure is not significantly greater than the structurally ambiguous folding with whose rule probabilities arbitrary assigned. The RND10 model actually has a slightly higher performance than the relatively accurate BJK model in distinguishing miRNAs from random sequences. Also, results obtained from both RND1 and RND10 are very consistent with the BJK model in characterizing the selected riboswitch sequences and distinguishing them from either miRNAs or random structures. Performance of the RND10 model is surprising; Even the left-most derivation of folding entropy values of sequences under this model shows higher folding diversity of sequences than the BJK model (See 2.1). Our conclusion about grammar design is that structural ambiguity can play a major role in characterizing ncRNA sequences.

Model Training

Model accuracy to annotated secondary-structure is essential to but not sufficient for characterizing all classes of ncRNA sequences. Being relatively more accurate than the other CYK-trained models, the BJK grammar models can effectively characterize miRNA sequences against random sequences. Furthermore, the BJK model was observed to distinguish riboswitch sequences from both miRNAs and random sequences having single structure in some cases. Hence, an effective grammar design trained to best predict RNA annotated secondary structure is very essential in investigating folding features of various classes of ncRNAs. However, plots of p-value vs. sensitivity and specificity of BJK models to annotated secondary structure in the above test, 2.8 and 2.9, suggest that high entropy is not necessarily related to model accuracy to annotated secondary structure in these sequence. In other words, annotated secondary structure cannot be the *only* criterion in capturing the folding features of certain sequences belonging to this class of ncRNA; i.e., diversity of folding distribution (here, Shannon's entropy) can also contain information about a class of ncRNA, regardless of how well its most likely scenario predicts the annotated secondary structure. Our overall conclusion of comparisons across various CYK-trained models is that model accuracy to annotated secondary structure is necessary but not sufficient for an effective SCFG to capture folding features of certain ncRNAs.

2.9 Conclusion

We developed a method for applying Shannon's entropy to RNA secondary structures, using SCFG. The analysis of known RNA structures showed this method could be useful provided an appropriate grammar is chosen. Used as a quantitative method for capturing the folding diversity of an RNA sequence, Shannon's entropy was shown to successfully capture the deterministic folding behavior of biological sequences on the secondary structural level. Signs of distinctly high folding diversity were also observed in certain classes of ncRNA sequences. While predicted secondary structure is essential to understanding the functions of the many ncRNAs, the diversity of folding space distribution of the sequence should not be overlooked. In certain cases, this diversity can lead to further characterization of the ncRNA as well as exploration into the limits of secondary structural modeling in understanding the *in vivo* conformational behavior of ncRNAs.

Chapter 3

Ab initio Riboswitch Identification Based on The Secondary Structure Folding Space

3.1 Introduction

Non-protein-coding RNA (ncRNA) elements play an important role in biological pathways such as gene regulation across the kingdom of life (Morris, 2008, 2012; Barrandon et al., 2008; Repoila and Darfeuille, 2009). It has been shown that conformational features of many such RNA elements play a major part in their biological function (Hall et al., 1982; Simmonds et al., 2008). In bacteria, RNA structural rearrangements can have a major effect on the expression of their downstream coding sequences [reviewed by Grundy and Henkin (2006)], a process known as *cis*-regulation. A classic example and one of the earliest such elements discovered is the complex regulatory mechanism that takes place upstream of the tryptophan operon in *Escherichia coli* during its expression (Oppenheim and Yanofsky, 1980). Regulation of the tryptophan biosynthetic operon, however, is achieved via different mechanisms in other organisms, such as *B. subtilis* and *Lactobacillus lactis* [reviewed by Merino and Yanofsky (2005)]. With much attention given to protein-coding genes in the past, ncRNAs have been left on the *invisible* side of genomic research for some time (Eddy, 2001).

3.1.1 Riboswitches

An interesting group of RNA regulatory elements are riboswitches. Originally found through sequence homology upstream of bacterial coding regions dating back about ten years ago (Mironov et al., 2002; Nahvi et al., 2002; Winkler et al., 2002a), these regulatory elements have been shown to be more abundant than previously expected. Riboswitches are defined as regulatory elements that take part in biological pathways by selectively binding to a specific ligand or metabolite, or uncharged tRNAs, without the need for protein factors. Environmental factors such as pH, ion concentration, and temperature can also trigger RNA conformational changes affecting gene regulation. Furthermore, nearly all riboswitches are located in the non-coding regions of messenger RNAs (Breaker, 2012) and are capable of regulating genes through both activation and attenuation of either transcription or translation [reviewed by Henkin (2008)]. Finally, other factors such as the transcription rate of RNA polymerase and concentration rates of the ligand and metabolites to be bound to the riboswitches add other dimensions to categorizing riboswitches, mapping them to a spectrum of kinetic/thermodynamic-driven folding trajectories in which to function. Riboswitches have also been found in cooperative or tandem arrangements (Breaker, 2012). It is speculated that there are at least 100 more undiscovered riboswitches in already sequenced bacterial genomes (Breaker, 2012). Conformational factors are essential to ligand-binding specificity of riboswitches. Many riboswitches can discriminate between similar small molecules with the aid of their structural geometry. For instance, the thiamine pyrophosphate (TPP) and SAM riboswitches measure the length of the ligand that binds to them (Thore et al., 2006; Serganov et al., 2006; Montange and Batey, 2006).

3.1.2 RNA Secondary Structure

The secondary structural topology of the RNA is very effective in scaffolding the tertiary conformation. Secondary structure mainly consists of a two-dimensional schema that depicts the base-pairing interactions within the RNA structure and is dominated by Watson-Crick base-pairing. One major computational method to predict RNA secondary structure is minimization of its free energy (MFE) within a thermodynamic ensemble, such as the Boltzmann ensemble Minimum Free Energy (Zuker and Stiegler, 1981; McCaskill, 1990). State-of-the-art thermodynamic models have proven to be effective in RNA secondary structural predictions in the cases of most RNA elements (one exception being Hammerhead type I ribozyme where loop tertiary interactions have a dominating effect on the structural conformation [Canny et al. (2004)]), although most programs may give multiple predictions with similar energy levels.

Stochastic context-free grammars (SCFG) have also been shown to be effective in secondary structural prediction of various RNA regulatory elements. SCFGs have a similar logic to Markov models except that they are nonlinear. Under such models, a secondary structure is recursively constructed through base-pairing and loop prediction given grammar rules and corresponding probabilities. Markov models, on the other hand aim to predict nucleotide arrangements from one side of the sequence to the other. Nawrocki and Eddy (2013) have shown that more sophisticated grammars, designed to mirror the thermodynamic models can exhaust the limits of prediction accuracy of structures, once trained on known RNA structures based on maximum-likelihood criteria. Pseudoknots, another RNA structural feature, are a kind of base-pairings that resemble structural knots and cannot be predicted via context-free grammars. Predictions of pseudoknots based on minimum free energy and context-sensitive grammars are possible, though computationally expensive (Rivas and Eddy, 1999).

Most classes of riboswitches, such as the purine riboswitches, exhibit strong secondary structural conservation. The *add* adenine riboswitch from *V. vulnificus* and *xpt* guanine riboswitch *B. subtilis* have very similar tertiary as well as secondary conformations, despite different crystal packing interactions, pH, and Mg crystallization conditions (Serganov et al., 2004). In fact, investigation of secondary-structural homology upstream of genomic regions containing the same genes has led to the discovery of more *cis*-regulatory elements in bacteria (Weinberg et al., 2007, 2010), making them the major current approach for riboswitch identification.

In addition, efforts have also been made to discover novel regulatory elements based on combining structural motifs gathered from a variety of known ncRNA genes. Tran et al. (2009) apply a neural-network classifier to *Escherichia coli* and *Sulfolobus solfataricus* for genome-wide prediction of ncRNAs based on features derived from sequences and structures of discovered ncRNAs that are available. Most of the discovered RNA regulatory elements are located upstream of the genes they regulate, as *cis*-regulatory elements and exhibit strong secondary structural conservation. Some exceptions to *cis*-regulation of prokaryotic riboswitches are two *trans*-acting S-adenosylmethionine (SAM) riboswitches (Loh et al., 2009) and an antisense regulation of a vitamin B_{12}-binding riboswitch (Mellin et al., 2013) in *Listeria monocytogenes*.

Serganov and Nudler (2013) have offered insights into structural and functional complexity of riboswitches already discovered.

It is difficult, however, to assess just how much secondary structural conservation is expected to be prevalent in undiscovered regulatory elements, since the methodologies that led to the discovery of known regulatory elements have for the most part been based on homology methods. Structural homology is not always successful in riboswitch identification. Though both SAM-I and SAM-IV riboswitches bind to the same ligand, Weinberg et al. (2008) indicated that all their efforts failed to detect SAM-IV riboswitches, despite rigorous sequence and structural homology searches based on the SAM-I riboswitch. The authors further hypothesized that the structural diversity of riboswitches could be far greater than what has been already observed. Serganov and Nudler (2013) suggest that there may not even be an interconnection between the structures of riboswitches and the nature of their cognate metabolites and consequently, the biochemical and structural information gathered so far may not be as useful in riboswitch validation as expected. Hence, *de novo* riboswitch prediction approaches would be very useful to help with finding new classes of riboswitches.

3.1.3 Conformational Dynamics

While secondary-structure conformational features are very descriptive of many classes of riboswitches, their folding dynamics are also critical. One of the major computational tools to explore possible folding trajectories is the free energy landscape. Originally defined for protein folding (Bryngelson et al., 1995), the probability for each structure is associated with a free energy and a distance from other possibilities. In an effort to investigate the thermodynamic equilibrium of RNA folding, Quarta et al. (2009) presented a case study of the folding landscape of the TPP riboswitch where the base-pairing distances between the structural possibilities form two major clusters, each of which correspond to either a native or ligand-bound structural conformation. Quarta et al. (2012) investigated the dynamics of energy landscapes across elongation of various riboswitches and showed that such landscapes have different clustering dynamics across kinetically and thermodynamically driven riboswitches. In a more recent work, energy landscape analyses led to strong evidence of evolutionary co-variation of base-pairs that favor conserved alternative structure of the purine riboswitch (Ritz et al., 2013). In addition, Freyhult et al. (2007) examined the lowest free energy structural conformations having a certain base-pairing distance to the actual structure of the RNA to explore the structural neighbors of an intermediate, biologically active structure.

Investigation into the folding dynamics of the nascent RNA based on its free energy sampling and corresponding pair-wise structural distance was computationally expensive according to our computations. Sampling the very large folding space of the RNA element in such a way that reflects its overall behavior and examining pair-wise base-pairing distances between predicted structures can be difficult and prone to model parameter biases. Furthermore, even if optimized parameters and sufficient samples of folding scenarios were available, finding the statistics that can lead to an effective quantified comparative measure across various RNA sequences would be a formidable challenge. The latter is mainly due to the fact that the characteristics of folding distribution (here, free energy vs. structural distance within a given ensemble of secondary structures) are not well understood. For instance, in a typical case of a stable single structure, we expect the free energy of the structural neighbors to the minimum free structure to have a funnel-like shape, where predictions with higher free energy are more distant than the MFE prediction. As for RNA elements with two functional and mutually exclusive secondary structures, more than one hill in the energy landscape may be detectable in certain cases. Obtaining a universal criterion that reflects RNA folding dynamics and potential for structural alternative(s), however, could be a formidable task.

One statistic to evaluate the distribution characteristics of any probabilistic model is the Shannon entropy (Shannon, 1948). While the conformation with maximum-likelihood under a given SCFG is referred to as the optimum structure under that model, all of the other sub-optimal conformations can be associated with a probability. The information-theoretic uncertainty (or here structural entropy) of SCFG-modeled folding space of an RNA is computationally convenient and can be calculated in polynomial time (Manzourolajdad et al., 2013). In this work, we have investigated the significance of structural entropy in capturing the thermodynamic characteristics of RNA elements having potential for an alternative fold. We then made an attempt to develop a computational method for *ab initio* riboswitch identification via structural entropy. We then evaluated a diverse set of prokaryotic RNA elements validated to have such potential.

3.2 Results

3.2.1 Considerations of our Approach in *ab initio* Riboswitch Identification

The folding model for which the structural entropy of the RNA is computed is very critical. SCFG folding models can be very lightweight and consist of only few grammar rules and parameters, or they can be very sophisticated. Parameters of SCFG models are usually set by maximizing their prediction accuracy of the actual RNA secondary structures using maximum-likelihood approaches. There is no guarantee, however, that folding models optimized for such criteria also preserve information about folding dynamics of such RNAs. It could be that increasing the accuracy of folding models under current approaches is done at the expense of altering the folding space of possible structures under that model and hence losing the information about folding dynamics of the RNA. Hence, it is essential that we examine the significance of structural entropy of RNAs under models not trained to best predict secondary structure, as well.

It has been previously shown that both high and low entropy values of certain classes of ncRNAs can be potentially significant with respect to random sequences of similar nucleotide composition and under certain SCFGs. It was also shown that factors, such as sequence length and model structures, are dominant factors in entropy calculations. Finally, it was shown that structural entropy is sensitive to factors such as grammar parameters and nucleotide composition (Manzourolajdad et al., 2013). For instance, for certain riboswitches, GC-composition was co-associated with significantly high entropy. Comparisons between RNA sequences and computer-generated random ones could be problematic, since structural entropy is highly sensitive to primary structural features as well as folding model parameters and conclusions based on comparisons with computer-generated sequences may not apply. This argument is further strengthened by the fact that some ncRNA sequences had higher folding entropy than random ones, which is counter-intuitive. Structural entropy of random sequences, whether generated based on primary or secondary structural features, could be too distant from that of real biological RNA sequences. Hence, we refrained from using random sequences in our assessment of significance of structural entropy.

Although riboswitches are very diverse in sequence and structure, there is a significant amount of sequence and/or structural similarity within each class of riboswitch. This is due to the fact that these riboswitches have been discovered using homology as explained above. In order for our approach not to be dominated by structural homology, we avoided using homologous RNA sequences or sequences that belong to closely related organisms, where possible. We also resorted to only evaluating riboswitches that have been experimentally validated to be functional rather than computationally discovered ones. The data set gathered in this work (see Materials and Methods) is a compromise between the above considerations and the need to include a diverse set of riboswitches in our data set. In this work, we use the term riboswitch for all gathered sequences including ribo-regulators such as the Tryptophan RNA element.

3.2.2 Using The Antisense as Internal Negative Control

Riboswitches are under selective pressure to preserve their potential for alternative folds due to their biological role. In this work, we made the assumption that there is more selective pressure for alternative fold on the sense sequences vs. the antisense strand, on average. We make this assumption based on the fact that *cis*-regulatory elements typically undergo conformational rearrangement which affects downstream gene regulation while the antisense strand does not necessarily have such a regulatory role. The sensitivity of structural entropy to various sequence and structural features is either very high (Manzourolajdad et al., 2013) or unknown. Using the antisense sequence to a putative riboswitch as negative control has the additional statistical convenience of enabling us to evaluate the significance of an ensemble of real sequences having identical sequence and structural features, such as length, GC-composition, and complementary G-C base-pairings.

Apart from the antisense sequence, untranslated regions (UTR) shorter than 80 nt have been selected as another negative set, since they are unlikely to contain structures over such a short length. Some riboswitch sequence segments, however, were selected to be shorter than this length. The length of the corresponding UTR (from transcription binding site to the translation start codon) for riboswitches, however, were not shorter than 80 nt. UTRs corresponding to the σ-70 in *E. coli* with distance less than 80 nt from the translation start codon were used here as sequences that do not contain structure (see Materials and Methods). We also included various features that were correlated with entropy such as MFE, length, and GC-composition. The reason for inclusion of MFE as a feature was that higher structural stability has been previously shown to be related to structural entropy for sequences of a single stable structure, namely microRNAs (miRNA) (Manzourolajdad et al., 2013). We also examined the utility of energy features such as alternatives to MFE (Ding et al., 2005). We then examined the performance of classifiers optimized for antisense discrimination on the *Bacillus subtilis* and *Escherichia coli* genomes, since most of the riboswitches were from these two organisms. Finally, the significance of structural entropy of the riboswitch was also evaluated in comparison to mutants altered to have less potential for alternative fold. We used biologically tested mutants in this regard.

3.2.3 Classification Results

Classification of the RNA sequence into three categories of having potential for alternative structure, one structure, or no structure were evaluated. We used a multinomial logistic regression approach for classification of sequences. Sequence features, such as Length (L), MFE, GC-composition (GC), and secondary structural entropy of the SCFG-modeled folding distribution of the RNA sequence (Manzourolajdad et al., 2013) were used for classification. One new feature considered was the free energy of the centroid structure (Ding et al., 2005) calculated by CentroidFold©(Sato et al., 2009), denoted here as CFE. The two lightweight SCFG folding models used to calculate folding entropy are denoted here as BJK and RND models, which are taken from the literature (see Materials and Methods). Other features, such as the base-pairing entropy of the BJK model *BJKbp* as defined by Huynen et al. (1997) and two-cluster average silhouette index of the energy landscape of the RNA *Sil* as calculated by Quarta et al. (2012) were also included. RNA encoded sequence from Bacteria validated to have potential for two alternative folds were gathered from the literature (see Table 3.1) as representatives of RNAs having potential for alterative folding. This generally consisted of riboswitches and some other ribo-regulators, although we refer to all these sequences as riboswitches, here. A sub-set of such sequences were selected as the positive control set of sequences having two structures. The criterion for selecting such a sub-set was minimum length of the RNA that exhibits alternative folds for each sequence. This criterion is further explained in Materials and Methods. The resulting set of length variant sequences was then divided into training and test sets described in Tables 3.1and B.1. Sequence segments and their corresponding structures are included in B.5.1 and B.5.2. The antisense of each of the sequence segments was used as an internal control, which represents sequences having only one structure. Antisense sequences were assumed not to have potential for two alternative structures while they may have at most one structure, since they are complementary to the sense; a *cis*-regulator has an alternative fold, typically through conformational rearrangements of the expression platform to be able to regulate the expression of the downstream genes in the same mRNA, while the antisense is not under such evolutionary pressure. A shortcoming for sense/antisense comparisons, however, is possible co-association with other sequence features such as U-composition; G-C base-pairs also exist on the antisense, while G-U pairs may differ between sense and antisense structures, under simple Watson-Crick and Wobble-pair folding models. 30 sequence segments were selected from σ-70 *E. coli* UTRs shorter than 80 nt, since they were believed to not form structure (see Table B.3 for information on sequence locations). They were fairly divided into training and test sets based on their MFE, and GC-compositions (See B.4 for average and standard deviation of features L, MFE, and GC for the training and test sets). The section Materials and Methods extensively discusses the criteria for selecting the sub-set, dividing the riboswitches and *E. coli* UTRs into training/test sets, as well as information on data sets.

An initial investigation of the power of selected features in sense/antisense discrimination was done via cross-validation for all the 104 (52 riboswitches and 52 antisense) sequences consisting of both the training and test sets. *E. coli* UTRs were excluded at this phase. Binomial logistic regression classification probabilities were assigned to each sequence based on the other 104 sense and antisense sequences. It is shown in Table B.2 that Features {*L,GC,GCU,Sil*} result in the highest true positive rate, lowest false positive rate, and highest area under the receiver operating characteristic (ROC) curve. Computing the *Sil* feature was based on the energy landscape structural sampling and was computationally expensive (See Materials and Methods for details). From amongst features that incorporate various entropy values, the features sets {*L,MFE,GC,RND*} and {*L,MFE,GC,BJK*} had a fairly acceptable performance, which was higher than that of the {*L,MFE,GC*} classifier without the SCFG feature (See Table B.2). We conclude classification based on the SCFG feature is significant, since length and GC-composition between sense and antisense are equal for every riboswitch and its corresponding antisense. The performance of other classifiers that involved Uracil composition were more dependent on sequence features rather that structural. Furthermore, the performance of the feature set {*L,MFE,GC,U*} was lower than that of {L,MFE,GC}, and hence not included in results. We selected classifiers {*L,MFE,GC,RND*}, {*L,MFE,GC,BJK*}, and {*L,MFE,GC*} for further investigation and will refer to them as LMFEGCRND, LMFEGCBJK, and LMFEGC, respectively. The ROC curve corresponding to these classifiers is shown in Figure B.1.

The performance of the tri-state classifier was evaluated by estimating classifier parameters from multinomial logistic regression of the training sets and then calculating the correct classification of sequences having zero (*E. coli*), one (antisense), or two (riboswitch) structures that are in the test set. 3D-plot of the MFE, GC-composition, and Entropy values under the RND model for sequence of the training set are depicted in figure 3.1. This distinction, however, becomes more subtle in comparison to the antisense control. Top and bottom views of the grid-view of values normalized to sequence length roughly shows this distinction (see Table 3.1). Sense-antisense differential entropy ($\Delta\, Entropy = 100 \times (Entropy_{sense} - Entropy_{antisense})/Entropy_{antisense}$) against the minimum free energy between the sense and the antisense ($\Delta\, MFE\% = 100 \times (MFE_{sense} - MFE_{antisense})/abs(MFE_{antisense})$) under RND and BJK models have been shown in Figures B.2 and B.3). Classification performance values are denoted in Table 3.2 along with sensitivity of each classifier. Sensitivity of tri-state classifiers were defined here as total number of correctly classified sequences divided by total number of sequences classified. Model LMFEGCBJK resulted in both highest sensitivity (80.2%) and highest percentage of correctly classified riboswitches (91.3%). Model LMFEGCRND had a sensitivity 73.9% which is slightly lower than the LMFEGC Model that excludes structural entropy. Regression coefficients of the classifiers are shown in Table 3.4. Testing the classifiers on constant lengths of sequences (for all training and test sets) did not increase performance (see Table B.5), although the *RND* was significant for sequences of length 150 nt in the training set. Constant length selection was based on extending (or shortening) the original choice of length of sequences from both 5' and 3' directions such that the center of the sequence does not change. We refer to this original choice of length as the actual length, hereon. We chose this scheme for simplicity. Other sequence selection methods may be preferred, since the alternative fold may occur on varying part (5' or 3') of the riboswitch sequence, in general. Substitution of CFE feature instead of MFE feature resulted in lower performance of classifiers (comparing Table 3.3 to Table 3.2).

3.2.4 *Bacillus subtilis*

The performance of the three tri-state classifiers on the eleven riboswitches and all other intergenic regions of the gram-positive bacterium, *Bacillus subtilis* are shown in Tables 3.5 and B.6, for the actual variable lengths and constant lengths of the test set, respectively. Operon coordinates were taken from Taboada et al.

[1] Table 3.1: This sequence overlaps codons. pH also has a role in alteration of structure.

[2] Table 3.1: Downstream-peptide

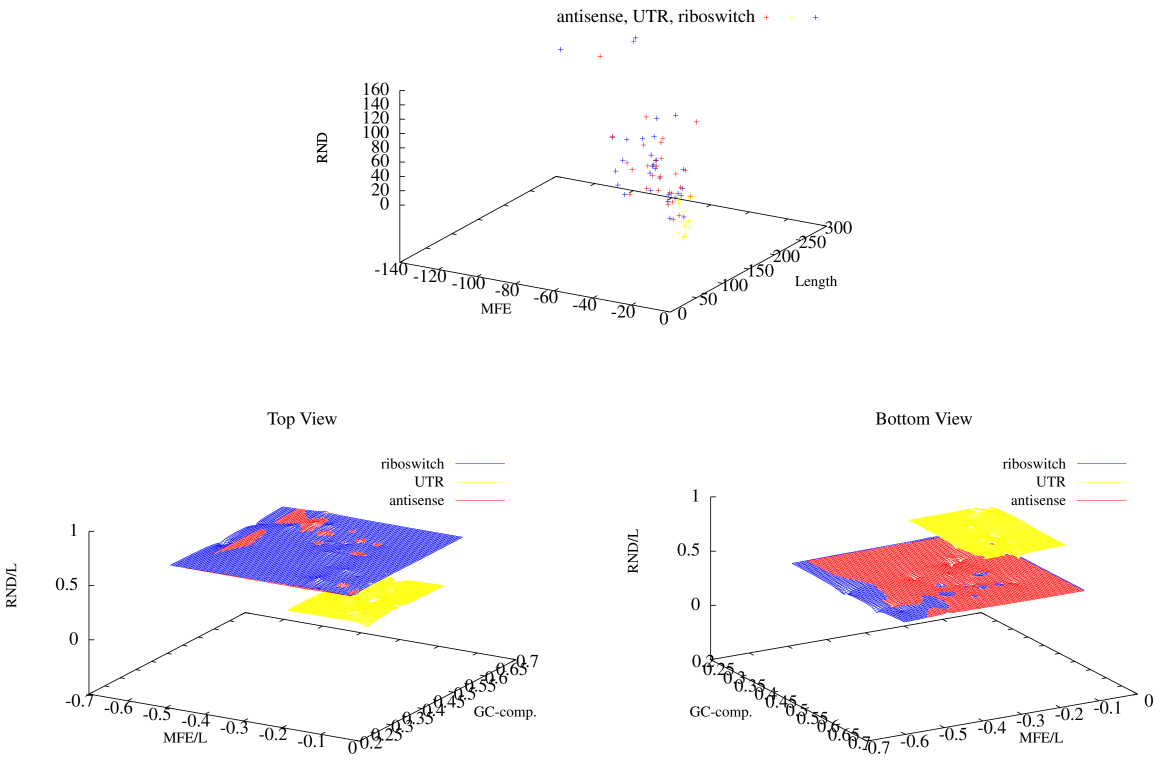

Figure 3.1: Structural Entropy vs. GC-comp. and MFE. 3D-plot of features MFE, Length and Structural Entropy of the training set sequences classifier under the RND model. Grid-view of different sets of sequences are shown in the top and Bottom views, riboswitches, *E. coli* UTRs, and antisense sequences. Axes RND/L and MFE/L show Structural Entropy and MFE normalized by the length of the sequence, respectively. Euclidean distance to actual values was used to generate the grids.

(2012). The details of the pipeline is described in Materials and Methods and Figure 3.3. Performance of classifiers was higher for length 157 nt rather than lengths 100 nt, 150 nt, or 200 nt. This was true even though overlapping sliding windows were used for those lengths (sequence segment with highest overlap was selected as positive hit). In addition, we can see from Table B.6 that as window size increased, the number of intergenic regions classified as riboswitches ($TP_2\%$) decreased. The classification performance of the LMFEGCRND model, however, was maximum at length 157 nt. (Length 157 was found using a rough optimization of various constant-length sequence selection and under the LMFEGCBJK model). We further examined the 157 nt length for two different sets of tests. In the first case, 157 nt-long segments were selected centered at the riboswitch (routine procedure) and in the second case, 157 nt extension of the 5' start of sequences were chosen. Classification performance is shown in Tables and 3.5 and 3.6. Performance was very sensitive to positioning of the sequence segment of constant length. For the case of 5' selections, the LMFEGCBJK model outperformed other models having $TP\% = 90.9$ while the centered-segment test had a performance even lower than choosing random positioning. Hence, the LMFEGCBJK is more suitable for high performance where computational complexity is not an issue. For faster genome-wide tests where examining all sequence positions is not possible the LMFEGCRND seemed more appropriate ($TP = 81.8\%$) and was based on selection of segments in a non-overlaping fashion, starting at the start codon for each operon. Selecting segments centered at riboswitches resulted in poor performance in *B. subtilis*.

The performance of classifiers on the eleven riboswitches were highly dependent on the length and positioning of sequence segments to be tested (see Tables 3.5, 3.6, and B.6). Furthermore, various riboswitches had different sensitivities to such features (data not shown). We found that sequence segments of length 157 nt resulted in higher performance compared to other lengths tested. Also, without knowledge of the exact location of the riboswitch, the LMFEGCRND model outperformed the LMFEGCBJK model, though the LMFEGCBJK model had a significantly higher performance if sequence segments were positioned at the right locations on the riboswitch. The likelihood for such desired positioning is very low; $1/WL$ for each riboswitch, where WL is the length of the non-overlapping sliding window.

The ranking of *B. subtilis* riboswitches using their actual length and constant length of 157 nt are shown in Tables 3.7 and 3.8, respectively. Classification probability of the LMFEGCRND model corresponding to the sequence segment overlapping with the TPP riboswitch (0.76) was higher than that of other riboswitches with ranking empirical p-value 0.0122. Results for the SAM-I riboswitch, however, were very poor. Interestingly, the actual length of the SAM-I riboswitch used in this study was also 157 nt.

Table B.7 contains the top 50 best hits from each strand of the *B. subtilis* intergenic regions and their corresponding probability values. Sequence segments having classification probabilities higher than or equal to 0.8 fall in the top 50. Plot of Entropy under the RND model and Uracil composition of the sequence segments form the *B. subtilis* showed that entropy values were correlated with higher Uracil composition (see Figure B.4). This may have been partly due to the fact that Uracil can bind with more nucleotides to form base-pairs under folding models. In order to suppress the effect of high Uracil composition, we sorted the top hits having Uracil compositions within the range of known riboswitches in *B. subtilis* in Table B.8. The location distribution of these hits can be seen in Figure 3.2. Sequence segments predicted to be riboswitches were not uniformly distributed across the genome. In order to investigate sequence location of segments having significantly high entropy values, regardless of their regression probabilities, we sorted segments having significantly low MFE (empirical p-value <0.05) while also having entropy values on the high 50 percentile. Hits with significant values that had GC and U compositions within the range of known riboswitches in *B. subtilis* are shown in Table 3.9. Interestingly, all of the hits also had significant Entropy p-values (<0.05). P-values are calculated empirically and separately for each choice of window size in the genome-wide scan. Finally, significantly high Entropy values of the 200 nt window scan that also have probability values higher than 0.8, along with other significant hits, are available in Tables B.9 and B.10 regardless of their MFE or nucleotide compositions.

Table 3.1: Data Collection. Collected sequences from literature observed to have more than one secondary structure. P corresponds to gram-positive and N corresponds to gram-negative. Genomic locations are available in Table B.1.

ID	Riboswitch	Organism (P/N)	Alteration	Grouping	References
ID01	Alpha Operon	*Escherichia coli* (N)	slow-fast	Train	(Gluick et al., 1997; Schlax et al., 2001)
ID02	ATP	*Bacillus subtilis* (P)	enzyme	Test	(Watson and Fedor, 2012)
	ATP[1]	*Salmonella* (N)	enzyme	None	(Lee and Groisman, 2012)
ID03	c-di-GMP	*Geobacter sulfurreducens* (N)	ligand	Train	(Weinberg et al., 2007)
ID04	c-di-GMP	*Candidatus Desulforudis* (P)	ligand	Test	(Smith et al., 2009)
ID05	Cobalamin	*Escherichia coli* (N)	ligand	Train	(Nahvi et al., 2002)
ID06	Cobalamin	*Bradyrhizobium japonicum* (N)	ligand	Train	(Vitreschak et al., 2003)
ID07	Cobalamin	*Salmonella* (N)	ligand	Test	(Ravnum and Andersson, 2001)
	D. peptide[2]	*Synechococcus sp. CC9902* (N)	motif	None	(Ames and Breaker, 2011)
ID08	Fluoride	*Pseudomonas syringae* (N)	ligand	Train	(Baker et al., 2012)
ID09	Fluoride	*Thermotoga petrophila* (N)	ligand	Train	(Ren et al., 2012)
ID10	Fluoride	*Bacillus cereus* (P)	ligand	Test	(Baker et al., 2012)
ID11	FMN	*Fusobacterium nucleatum* (N)	ligand	Train	(Serganov et al., 2009; Vicens et al., 2011)
ID12	FMN	*Escherichia coli* (N)	ligand	Train	(Winkler et al., 2002c; Hollands, 2012)
ID13	FMN	*Bacillus subtilis* (P)	ligand	Test	(Winkler et al., 2002c; Serganov et al., 2009; Vicens et al., 2011)
	glmS	*T. tengcongensis* (N)	None	None	(Barrick et al., 2004; Winkler et al., 2004; Cochrane et al., 2007; Klein and Ferre-D'Amare, 2009)
	glnA	*Synechococcus elongatus* (N)	motif	None	(Ames and Breaker, 2011)
ID14	Glycine	*Fusobacterium nucleatum* (N)	ligand	Train	(Mandal et al., 2004; Kwon and Strobel, 2008; Butler et al., 2011)
ID15	Glycine	*Bacillus subtilis* (P)	ligand	Test	(Mandal et al., 2004)
	Hammerhead I	*Schistosoma Mansoni* (-)	None	None	(Canny et al., 2004; Martick and Scott, 2006)
	Hammerhead II	*Marine metagenome* (-)	None	None	(Perreault et al., 2011)
ID16	Lysine	*Thermotoga maritima* (N)	ligand	Train	(Serganov et al., 2008; Garst et al., 2008)
ID17	Lysine	*Bacillus subtilis* (P)	ligand	Test	(Garst et al., 2008)
ID18	Magnesium	*Salmonella enterica* (N)	$Mg2^{+}$	Train	(Cromie et al., 2006; Hollands, 2012)
ID19	Magnesium	*Escherichia coli* (N)	$Mg2^{+}$	Train	(Cromie et al., 2006)
ID20	Magnesium	*Bacillus subtilis* (P)	$Mg2^{+}$	Test	(Dann et al., 2007)
ID21	Moco	*Escherichia coli* (N)	ligand	Train	(Regulski et al., 2008)
ID22	pH-responsive	*Escherichia coli* (N)	pH	Train	(Nechooshtan, 2009)
ID23	pH-responsive	*Serratia marcescens* (N)	pH	Test	(Nechooshtan, 2009)
ID24	preQ1 II	*Streptococcus pneumoniae* (P)	ligand	Train	(Weinberg et al., 2007; Meyer et al., 2008)
ID25	preQ1 I	*Bacillus subtilis* (P)	ligand	Test	(Klein et al., 2009)
ID26	Purine (Adenine)	*Vibrio vulnificus* (N)	ligand	Train	(Serganov et al., 2004)
ID27	Purine (Adenine)	*Bacillus subtilis* (P)	ligand	Test	(Serganov et al., 2004)
ID28	Purine (Guanine)	*Bacillus subtilis* (P)	ligand	Test	(Serganov et al., 2004; Batey et al., 2004)
ID29	ROSE-1	*Bradyrhizobium japonicum* (N)	Heat	Train	(Nocker et al., 2001; Chowdhury et al., 2006)
ID30	ROSE-2	*Escherichia coli* (N)	Heat	Train	(Nocker et al., 2001)
ID31	ROSE-2387	*Mesorhizobium loti* (N)	Heat	Test	(Nocker et al., 2001)
ID32	ROSE-N1	*Rhizobium* (N)	Heat	Test	(Nocker et al., 2001)
ID33	ROSE-P2	*Bradyrhizobium* (N)	Heat	Train	(Nocker et al., 2001)
ID34	SAH	*Ralstonia solanacearum* (N)	ligand	Train	(Weinberg et al., 2007; Edwards et al., 2010)
ID35	SAM-I	*T. tengcongensis* (N)	ligand	Train	(Montange and Batey, 2006)
ID36	SAM-I	*Bacillus subtilis* (P)	ligand	Test	(Grundy and Henkin, 1998; Tomsic et al., 2008; Lu et al., 2010; Boyapati et al., 2012)
ID37	SAM-II	*Agrobacterium tumefaciens* (N)	ligand	Train	(Corbino et al., 2005)
ID38	SAM-III (SMK)	*Streptococcus gordonii* (P)	ligand	Train	(Fuchs et al., 2006)
ID39	SAM-III (SMK)	*Enterococcus faecalis* (P)	ligand	Test	(Fuchs et al., 2006; Lu et al., 2008; Wilson et al., 2011)
ID40	SAM-IV	*Streptomyces coelicolor* (P)	ligand	Train	(Weinberg et al., 2008)
ID41	SAM-IV	*Mycobacterium tuberculosis* (P)	ligand	Test	(Weinberg et al., 2008)
ID42	SAM-SAH	*Roseobacter* (N)	ligand	Train	(Weinberg et al., 2010)
ID43	SAM-SAH	*Oceanibulbus indolifex* (N)	ligand	Test	(Weinberg et al., 2010)
ID44	SAM-V	*Cand. P. ubique* (N)	ligand	Train	(Poiata et al., 2009)
ID45	SAM-V	*Cand. P. ubique* (N)	ligand	Test	(Meyer et al., 2009)
ID46	THF	*Eubacterium siraeum* (P)	ligand	Train	(Ames et al., 2010; Huang et al., 2011)
ID47	THF	*Clostridium kluyveri* (P)	ligand	Test	(Ames et al., 2010)
ID48	TPP	*Escherichia coli* (N)	ligand	Train	(Winkler et al., 2002b; Serganov et al., 2006; Nudler, 2006; Haller et al., 2013)
ID49	TPP	*Bacillus subtilis* (P)	ligand	Test	(Mironov et al., 2002; Nudler, 2006)
ID50	Tryptophan	*Escherichia coli* (N)	complex	Train	(Oppenheim and Yanofsky, 1980; Neidhardt, 1996)
ID51	Tryptophan	*Bacillus subtilis* (P)	complex	Test	(Babitzke and Gollnick, 2001; Babitzke et al., 2003)
ID52	Tuco	*Geobacter metallireducens* (N)	ligand	Test	(Regulski et al., 2008)
	yxkD	*Bacillus subtilis* (P)	motif	None	(Barrick et al., 2004)

Table 3.2: Classification Performance. Classifier Performance. Actual length of sequences used. Column `Classifier` denotes features used from the training set. $TP\%$ denotes percentage of true positives. $FP_1\%$ and $FP_2\%$ represent the percentages of antisense sequences and *E. coli* UTRs that are misclassified as riboswitches, respectively. Sensitivity denotes overall percentage of correctly classified sequences. Sig. denotes significant (less than 0.05 in the training set) features of the multinomial classifier.

Classifier	$TP\%$	$FP_1\%$	$FP_2\%$	Sensitivity	Sig.
LMFEGCBJK	91.3	43.5	15.4	72.9	MFE
LMFEGC	82.6	30.4	23.1	71.2	MFE
LMFEGCRND	73.9	30.4	38.5	64.4	L,MFE

Table 3.3: Classification Performance Using Centroid Free Energy. Classifier Performance. Actual length of sequences used. Feature CFE denotes centroid free energy as calculated by CentroidFold©(Sato et al., 2009). Column `Classifier` denotes features used from the training set. $TP\%$ denotes percentage of true positives. $FP_1\%$ and $FP_2\%$ represent the percentages of antisense sequences and *E. coli* UTRs that are misclassified as riboswitches, respectively. Sensitivity denotes overall percentage of correctly classified sequences. Sig. denotes significant (less than 0.05 in the training set) features of the multinomial classifier.

Classifier	$TP\%$	$FP_1\%$	$FP_2\%$	Sensitivity	Sig.
LCFEGCRND	65.2	30.4	15.4	66.1	CFE
LCFEGC	78.3	56.5	15.4	61	L,CFE
LCFEGCBJK	82.6	65.2	15.4	59.3	GC

Table 3.4: Logistic Regression Coefficients of Classifiers. Regression coefficients (exponents) of the multinomial logistic regression classifier: intercept, Length, MFE, GC-composition, Entropy. Parameter vectors β_1 and β_2 denote coefficients for *E. coli* UTRs and ribswitch sense sequences for the riboswitches of the training set, respectively. Coefficients normalized with respect to those for riboswitch antisenses. i. e. antisense coeficients being 0.

Classifier	β_1	β_2
LMFEGCRND	3.191,.336,.683,-.723,-.465	5.052,-.161,-.089,-7.454,.220
LMFEGCBJK	10.597,-.203,.367,-10.856,.651	5.524,-.082,-.132,-9.247,.120
LMFEGC	3.869,.052,.525,-1.419	3.373,-.025,-.068,-6.234

Table 3.5: Classification Performance in *B. subtilis*. Classifier Performance on the eleven *B. subtilis* riboswitches. Actual length of sequences used. Column `Features` denotes features used from the training set. $TP\%$ denotes percentage of true positives. $FP_1\%$ represent the percentages of antisense sequences that are misclassified as riboswitches.

Classifier	$TP\%$	$FP_1\%$
LMFEGCBJK	91.1	54.5
LMFEGC	81.2	36.4
LMFEGCRND	72.7	36.4

Table 3.6: Classification Performance in *B. subtilis* Under Constant Length. Classifier performance on the eleven riboswitches in *B. subtilis*. Constant length of 157 nt from 5' of riboswitch downstream is used for riboswitches (first two columns). Constant length of 157 nt centered at the center of riboswitches used (last two columns). $TP\%$ denotes percentages of correctly classified riboswitches. $FP_1\%$ denotes percentage of misclassified antisense.

Segment	5'		center	
Classifier	$TP\%$	$FP_1\%$	$TP\%$	$FP_1\%$
L,MFE,GC,BJK	90.9	9.1	63.4	18.2
L,MFE,GC,RND	54.5	0	63.4	36.4
L,MFE,GC	72.7	9.1	63.4	18.2

Table 3.7: *B. subtilis* Riboswitches Ranking Under Actual-Length Test. Ranking probabilities of the eleven *B. subtilis* riboswitches of *B. subtilis* under the LMFEGCBJK classifier. Actual sequence length used as test. Column `Probability` is the classification probability that the sequence is a riboswitch.

Name	Probability
Adenine	0.82
FMN	0.70
TPP	0.68
Tryptophan	0.67
Glycine	0.63
Lysine	0.63
Guanine	0.60
ATP	0.58
Magnesium	0.54
SAM-I	0.53
preQ1	0.451

Table 3.8: *B. subtilis* Riboswitches Ranking Under Constant-Length Test. Ranking probabilities of the eleven *B. subtilis* riboswitches within the 157 nt non-overlaping window scan of the intergenic regions of *B. subtilis* under the LMFEGCRND classifier. Total of 28340 sequence segments belonging to intergenic regions longer than 150 nt were analyzed. Operon coordinates: (Taboada et al., 2012). Overlap denotes the percentage of overlap of the sequence segment with the riboswitch. Column p-Value is the ranking divided by 28340. Column `Probability` is the classification probability that the sequence is a riboswitch.

Name	Overlap	Rank	p-Value	Probability
TPP	82.9	347	0.0122	0.76
Guanine	90.1	535	0.0189	0.735
ATP	85.6	1159	0.0409	0.676
Lysine	83.5	2278	0.0804	0.612
Adenine	100	2459	0.0868	0.604
FMN	51.7	3880	0.1369	0.547
preQ1	80	4051	0.1429	0.541
Magnesium	62.3	4212	0.1486	0.536
Glycine	91.7	5200	0.1835	0.508
Tryptophan	100	6074	0.2143	0.484
SAM-I	68.8	12330	0.4351	0.356

Table 3.9: Top Entropy Hits in *B. subtilis* Filtered for GC-comp. and Uracil-comp. Significant hits of the forward and reverse strands of the *B. subtilis* intergenic regions having significantly high RND entropy (p-Val.<0.0500), significantly low (p.Val.<0.050), GC and Uracil compositions within the range of those for known riboswitches Threshold values and their corresponding p-values have been calculated separately for each genome-wide test. No overlap used for 157 nt scan (28340 segments). 175nt overlap used for 150 nt scan (60204 segments). 100 nt overlap used for 200 nt scan (44847 segments). Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the multinomial regression likelihood of being a riboswitch under the LMFEGCRND model. Negative values indicate distance to upstream operon. Columns `Upsream/Downstream Operon` show gene ID within the operon.

B. subtilis	Start	End	Strand	Upstream Operon	Upstream Gene	Dist. to Upstream	MFE	MFE p. Val.	GC	RND	RND p. Val.	Uracil	Dist. to Downstream	Downstream Gene	Downstream Operon	Probability
157nt	191850	192006	forward	BSU01590	ybaS	-12186	-54.16	0.01	0.4904	94.7470016	0.0359	0.3057	2277	trnSL-Glu2	BSU_tRNA_75	0.8561704159
157nt	749147	749303	forward	BSU06780	yeeC	-3069	-49.19	-	0.4458	94.8936005	0.0310	0.3630	550	yeeG	BSU06820	0.8463344574
157nt	665425	665581	forward	BSU06130	ydjC	-677	-51.50	-	0.4968	95.6813965	0.0169	0.3439	1963	gutB	BSU06150	0.8462108970
157nt	1017114	1017270	forward	BSU09400	spoVR	18	-53.10	-	0.4968	94.5084991	0.0412	0.3376	1806	lytE	BSU09420	0.8305525184
157nt	823013	823169	forward	BSU07480	yfmG	-604	-48.60	-	0.5032	94.2139969	0.0507	0.2866	4161	yfmA	BSU07540	0.7458834648
157nt	3421066	3421222	reverse	BSU33340	sspJ	-320	-49.40	-	0.4713	97.1984024	0.0049	0.3312	79	lysP	BSU33330	0.8851321340
157nt	3158851	3159007	reverse	BSU30890	ytxO	-328	-48.50	-	0.4395	95.1657028	0.0250	0.3630	3376	ytdA	BSU30850	0.8522043228
157nt	736435	736591	reverse	BSU06740	yefB	-2481	-50.51	-	0.4904	95.0205994	0.0279	0.3376	3690	yerO	BSU06700	0.8204180002
157nt	201248	201404	reverse	BSU01800	alkA	-1220	-49.46	-	0.4968	95.0683975	0.0269	0.2930	7301	ybbK	BSU01720	0.8003951907
157nt	4129689	4129845	reverse	BSU40200	yydD	-810	-48.40	-	0.4904	94.8125000	0.0332	0.3567	2120	yydF	BSU40180	0.7834032774
150nt	4134601	4134750	forward	BSU40190	fbp	-4483	-50.91	-	0.4733	91.2214966	-	0.3800	677	yycS	BSU40240	0.8779885173
150nt	3359819	3359968	forward	BSU32650	yurS	-5258	-46.80	-	0.4600	92.3918991	-	0.3600	12677	yuzL	BSU32849	0.8770275712
150nt	749175	749324	forward	BSU06780	yeeC	-3102	-46.50	-	0.4600	92.0363998	-	0.3867	527	yeeG	BSU06820	0.8652582169
150nt	1958237	1958386	forward	BSU18190	yngC	-9899	-48.30	-	0.4733	90.9682007	-	0.3533	44327	iseA	BSU18380	0.8436317444
150nt	1540761	1540910	forward	BSU14680	ykzC	-1995	-46.42	-	0.4333	90.3455963	-	0.3400	1052	ylaA	BSU14710	0.8428211212
150nt	3199841	3199990	forward	BSU31170	yulF	-1875	-47.20	-	0.4467	89.9530029	-	0.3733	12677	tgl	BSU31270	0.8267914653
150nt	3421066	3421215	reverse	BSU33340	sspJ	-325	-49.40	-	0.4800	93.4540024	-	0.3333	74	lysP	BSU33330	0.9072541595
150nt	933665	933814	reverse	BSU08620	yfhP	-718	-49.54	-	0.4600	89.9813995	-	0.3733	5474	sspK	BSU08550	0.8426564932
200nt	3359769	3359968	forward	BSU32650	yurS	-5183	-66.40	-	0.4600	123.4530029	-	0.3450	12702	yuzL	BSU32849	0.9236087203
200nt	339225	339424	forward	BSU03130	nadE	-20	-63.60	-	0.4700	121.1809998	-	0.3250	702	aroK	BSU03150	0.8414211273
200nt	1678852	1679051	reverse	BSU17060	ymzD	-101667	-62.81	-	0.4750	122.0299988	-	0.3300	7299	ylqB	BSU15960	0.8517054319
200nt	3717398	3717597	reverse	BSU36100	ywrD	-1637	-51.30	-	0.3650	130.8540039	-	0.3950	399	cotH	BSU36060	0.9702541828
200nt	198226	198425	reverse	BSU01800	alkA	-4222	-30.81	-	0.3750	130.7449951	-	0.5150	4299	ybbK	BSU01720	0.8267450333
157nt	235800	235956	reverse	BSU02180	ybfE	-2285	-54.99	-	0.3312	66.4815979	0[7]	0.3439	550	glpT	BSU02140	0.0401644297
200nt	3236257	3236456	forward	BSU31500	yuxK	61	-82.70	-	0.4200	93.3933029	0[8]	0.2650	802	yufL	BSU31520	0.0853443071

3.2.5 *Escherichia coli*

Nine out of the 29 riboswitches in the training set are from the *E. coli* genome. As a test of the generality of the results on *B. subtilis*, we evaluated the performance of the three classifiers on various constant-length riboswitches, 100 nt, 150 nt, 157 nt, and 200 nt on *E. coli*. The performance of the LMFEGCRND classifier for the 100 nt-constant length was slightly higher than other tests (data not shown). Hence, the 100 nt constant-length window scan of 50 nt overlap was used to examine the intergenic regions of *E. coli*. The operon coordinates were taken from RegulonDB website (Salgado et al., 2013). Top 50 hits on each strands are available in Table B.11. Top 50 hits having Uracil compositions within the range of known riboswitches are organized in Table B.12. The genomic distribution of the latter set is shown in Figure 3.2. Sequence segments having significant MFE and high Entropy values are sorted in Tables 3.10, and B.13 for significant and insignificant entropy values, respectively.

[7] Table 3.9: The entropy of this sequence is the lowest within the test. The significance of this value is also shown in B.4 as the lowest blue point on the graph.

[8] Table 3.9: The entropy of this sequence is the lowest within the test.

Table 3.10: Top Entropy Hits of *E. coli* Filtered for GC- and Uracil-comp. Significant hits of the forward and reverse strands of the *E. coli* intergenic regions having significantly high RND entropy (p-Val.<0.0500), significantly low (p.Val.<0.050), GC and Uracil compositions within the range of those for known riboswitches Threshold values and their corresponding p-values have been calculated separately for each genome-wide test. 50 nt overlap used for 100 nt scan (100090 segments). 175 nt overlap used for 150 nt scan (66414 segments). Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the multinomial regression likelihood of being a riboswitch under the LMFEGCRND model. Positions are according to gb|U00096.2 version of *E. coli* and not gb|U00096.3 version. Negative values indicate distance to upstream operon. Columns `Upsream/Downstream Operon` show gene ID within the operon.

E. coli	Start	End	Strand	Upstream Operon	Dist. to Upstream	MFE	MFE p. Val.	GC	RND	RND p. Val.	Uracil	Dist. to Downstream	Downstream Operon	Probability
100nt	4083889	4083988	forward	yiiF	-5848	-38.4	0.0267	0.53	58.6367989	0.0365	0.29	102	fdhD	0.789
100nt	187962	188061	forward	cdaR	-4293	-36.4	0.0466	0.53	59.0985985	0.0229	0.32	1702	rpsB,tff,tsf	0.776
100nt	952485	952584	forward	ycaK	-2955	-36.8	0.0419	0.52	58.3203011	0.0494	0.27	3452	ycaP	0.765
100nt	4115038	4115137	forward	uspD,yiiS	-3245	-37	0.0396	0.53	58.3563995	0.0477	0.33	1452	zapB	0.756

E. coli	Start	End	Strand	Upstream Operon	Dist. to Upstream	MFE	MFE p. Val.	GC	RND	RND p. Val.	Uracil	Dist. to Downstream	Downstream Operon	Probability
150nt	2686923	2687072	forward	hmp	-1802	-56.00	-	0.5333	90.7522964	0.0077	0.32000	6827	mltF	0.8671584129
150nt	2887386	2887535	forward	iap	-11672	-56.40	-	0.5333	89.1240005	-	0.0294	2777	queD	0.8254097700
150nt	3467187	3467336	forward	gspO[9]	-2871	-56.10	-	0.5200	88.5419006	0.0450	0.29333	8402	slyX	0.8172816634
150nt	3576825	3576974	reverse	yhhW	-74	-55.60	-	0.4800	88.6371994	0.0419	0.30666	149	gntK,gntR,gntU	0.8547886610
150nt	2195866	2196015	reverse	yehS	-13808	-58.00	-	0.5333	88.6897964	0.0405	0.27333	3749	mrp	0.8320623040

[9] Table 3.10: Complete list of genes in this operon is gspC,gspD,gspE,gspF,gspG,gspH,gspI,gspJ,gspK, gspL,gspM,gspO.

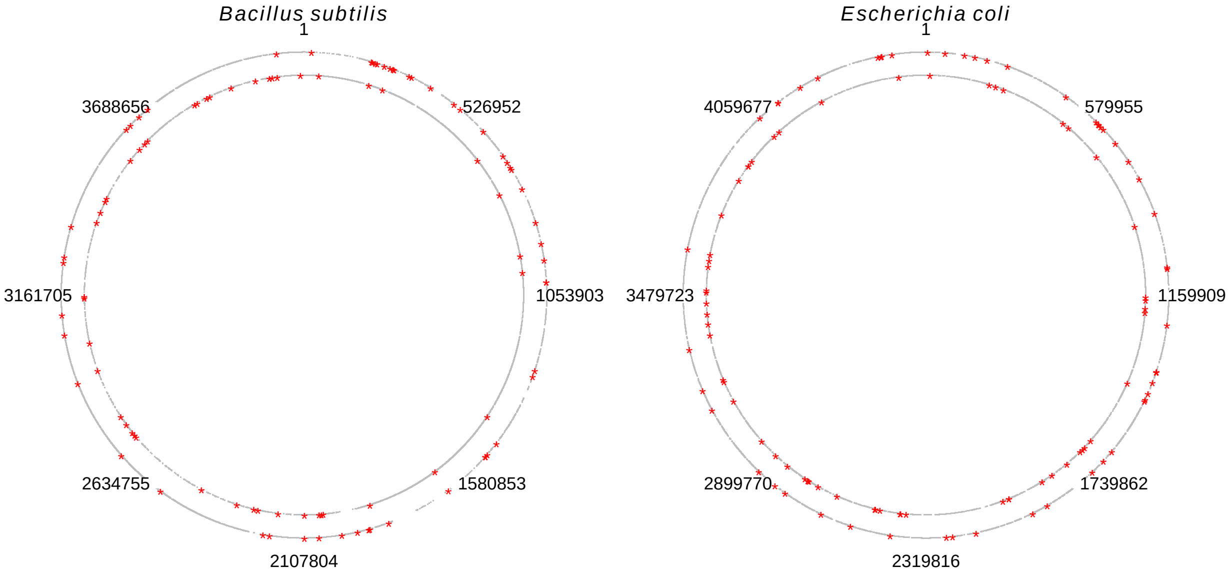

Figure 3.2: Structural Entropy Genomic Distribution. **Left:** *Bacillus subtilis***.** Distribution of locations of sequence segments of the non-overlapping 157 nt window scan of the *B. subtilis* intergenic regions. Location of all segments tested is depicted as grey. Location of segments with Uracil composition between 0.2484 and 0.40127 and probabilities higher than 0.8 under the classifier LMFEGCRND are shown in red. (Please see Table B.8 for more information on the hits). Outer circle represents the direct strand while the inner circle represents the complementary strand. 72.2% (39 out of 54) of hits on the forward strand are located in the first half of the genome. 69.6% (32 out of 46) of the hits on the reverse strand are located in the second half of the genome. **Right:** *Escherichia coli***.** Distribution of locations of sequence segments of the 50 nt-overlapping 100 nt window scan of the *E. coli* non-annotated intergenic regions. Location of all segments tested is depicted as grey. Location of segments with Uracil composition between 0.23 and 0.34 and probabilities higher than 0.768 under the classifier LMFEGCRND are shown in red (Please see Table B.12 for more information on the hits). Outer circle represents the direct strand while the inner circle represents the complementary strand. 64.2% (34 out of 53) of hits on the forward strand are located in the first half of the genome. 60% (36 out of 60) of the hits on the reverse strand are located in the second half of the genome.

3.2.6 Mutagenesis

In order to investigate the sensitivity of various structural features to the folding space of the riboswitches, we compared their wild-type value to that of *structural* mutants. By structural mutants, we mean those mutant sequences that were designed to disrupt either of the two biologically functional conformations of the riboswitch. Such structural mutants whose regulatory functions had been experimentally investigated were gathered from the literature. The percentage of change in entropy values for mutants relative to the wild type is shown in Table 3.12. These mutant sequences may not have been naturally occurring biological sequences, but they had very similar sequence features to their wild type, enabling us to evaluate the variations of structural features with respect to loss of functionality.

The criterion for the performance of each feature was as follows: For each riboswitch, we compared entropy values of mutants with structural alteration (denoted as YES) with those of the wild type and non-structural alterations (denoted as NO). We then counted the structural mutants that have *lower* values than those for both the wild type and non-structural mutants and divided it by the total number of structural mutants. We did this for every features and riboswitch. We then averaged the computed percentages across all riboswitches, since each riboswitch can be looked at as an independent test. Performance of each feature is shown in Table 3.12. The performance of the base-pairing entropy *BJKbp* is higher than other features on average. This suggests that structural mutants are expected to have lower base-pairing entropy than non-structural mutants and wild type 77.8 percent of the time, regardless of the riboswitch tested. For certain riboswitches or riboswitch segments, such as the *B. subtilis* Magnesium and the expression platform of the *Salmonella* Cobalamin riboswitch, however, various structural entropy values have higher entropy than the wild type, which means that our hypothesis of higher entropy and alternative fold does not always hold. The average silhouette index of energy landscapes (*Sil*) has a much better performance for the mentioned riboswitches.

Table 3.11: Mutagenesis. Percentage of change in entropy values of mutants compared to wild type. Mutation names are according to the literature. Type of disruption to wild type activity/conformation is denoted in column `function` (Please see references for more detail on mutation information). Mutants have same length as the wild type, except for the ROSE-P2 thermosensor. Wild-type segments are the same as gathered data, except for the SAM-I riboswitch where a homologue has been used. ΔRND% and ΔBJK%, refer to structural entropy values for the RND and BJK models, respectively. ΔBJKbp% refers to the base-pairing entropy of the BJK model as defined by (Huynen et al., 1997). ΔSil% refers to the two-cluster average silhouette index of the energy landscape of the RNA as calculated by (Quarta et al., 2012). Sensitivity% and specificity% refer to BJK model accuracy to the secondary structural conformation, with disregard to pseudoknots.

Wild-type	Riboswitch (Length)	Organism		Sensitivity %	Specificity %	
ID49	TPP (158)	*B. subtilis*		56.9	51.8	
Mutants (Mironov et al., 2002)	Function	Disruption of only *one* structure	ΔRND %	ΔBJK %	ΔBJKbp %	ΔSil %
+30	Disrupts anti-antiterminator	Yes	0.7	-2.6	-3.9	-55.2
+118	Disrupts anti-terminator	Yes	-0.4	5.3	-0.7	-50.3
+80	Disrupts thi-box	No	0.8	3.3	0.8	-38.2
+97	Disrupts thi-box	No	-0.8	1.9	1.6	-63.2
Wild-type	Riboswitch (Length)	Organism		Sensitivity %	Specificity %	
ID13	FMN (236)	*B. subtilis*		81.8[1]	64.3	
Mutants (Mironov et al., 2002)	Function	Disruption of only *one* structure	ΔRND %	ΔBJK %	ΔBJKbp %	ΔSil %
G34C/G35C	Disrupts anti-terminator	Yes	-1.6	-5.5	-2.4	15.4
C86T	Disrupts rfn-box	No	0.2	-0.1	0.6	11.8
C49T	Disrupts rfn-box	No	0.3	0.5	0	-14.3
G157A/G160A	Disrupts anti-antiterminator	Yes	0	-0.7	-0.9	66.7
Wild-type	Riboswitch (Length)	Organism		Sensitivity %	Specificity %	
ID36.1[2]	SAM-I (159)	*B. subtilis*		94	88.7	
Mutants (Winkler et al., 2003)	Function	Disruption of only *one* structure	ΔRND %	ΔBJK %	ΔBJKbp %	ΔSil %
Ma	Disturbs both structures	No	2.3	15.8	10.7	-48.8
Mab	Disrupts anti-terminator	Yes	-2.3	-0.29	-0.4	4.1
Mc	Disrupts anti-terminator	Yes	0.3	-0.31	-0.8	-0.3
Mabc	Compensates mutations to wild type	No	-1.1	-0.32	-0.7	-3.2
Wild-type	Riboswitch (Length)	Organism	Reference	Sensitivity %	Specificity %	
ID18	Magnesium (172)	*Salmonella enterica*		64.5[3]	43.5	
Mutant (Hollands, 2012)	Function	Disruption of only *one* structure	ΔRND %	ΔBJK %	ΔBJKbp %	ΔSil %
C145G	Favors high Mg2$^+$ conformation	Yes	1.7	-1.8	-4.7	-10.1
Wild-type	Riboswitch (Length)	Organism		Sensitivity %	Specificity %	
ID12	FMN (264)	*E. coli*		38.9	32.3	
Mutants (Hollands, 2012)	Function	Disruption of only *one* structure	ΔRND %	ΔBJK %	ΔBJKbp %	ΔSil %
M1	Favors +FMN conformation	Yes	0.4	-3.8	-5.8	-43
M2	Favors -FMN conformation	Yes	-1.4	-1.9	-1.2	-5.7
Wild-type	Riboswitch (Length)	Organism		Sensitivity %	Specificity %	
						Continued on next page

						– continued from previous page
ID20	Magnesium (204)	*B. subtilis*		78	65	
Mutants (Dann et al., 2007)	Function	Disruption of only *one* structure	ΔRND %	ΔBJK %	ΔBJKbp %	ΔSil %
M5	Disrupts termination	Yes	2.7	0.9	0.7	-12.3
M6	Distrupts anti-terminator	Yes	3.9	12.4	8	-14.8
Wild-type	Riboswitch (Length)	Organism		Sensitivity %	Specificity %	
ID07	Cobalamin (95)	*Salmonella*		59.4	61.3	
Mutants (Ravnum and Andersson, 2001)	Function	Disruption of only *one* structure	ΔRND %	ΔBJK %	ΔBJKbp %	ΔSil %
G373 →C	Distrupts alteration[4]	Yes	1.9	7.4	5.2	31.3
G375 →C	Distrupts alteration	Yes	0.6	10.7	5.1	13.2
G376 →C	Distrupts alteration	Yes	1.3	1.4	0.2	5.1
C440 →G	Distrupts alteration	Yes	5.8	11.8	5.5	-22.8
C441 →G	Distrupts alteration	Yes	3.7	10.1	5.4	-11.1
C443G460 →GC	Distrupts alteration	Yes	1.4	-3.2	-2.1	-1.9
G373C443G460 →CGC	Compensates mutations to wild type	No	-0.2	6.5	4.3	8.6
Wild-type	Riboswitch (Length)	Organism		Sensitivity %	Specificity %	
ID33	ROSE-P2 (135)	*Bradyrhizobium*		22.7[5]	22.2	
Mutant (Chowdhury et al., 2006)	Function	Disruption of only *one* structure	ΔRND %	ΔBJK %	ΔBJKbp %	ΔSil %
ΔG83[6]	Deletion of a critical nucleotide	Yes	-2.6	-8.1	-4.7	8.6

We also used the Ribex (Abreu-Goodger and Merino, 2005) tool in order to have an overall view of the identification power of riboswitches in the test set based on their sequence annotation. This tool uses similarity measures and annotation information from sequenced bacterial genomes to identify the type and position of a given RNA sequence based on other riboswitches. Performance of the riboswitch identification tool is denoted in Table 3.13. The tool was set to search sequences predicted to be ribo-regulators, as well. Twelve out of the 23 riboswitches (52.17%) in the test set are correctly identified as riboswitches using the search tool. However, if we exclude the annotation of corresponding family of the riboswitch from the

[1] Table 3.11: Two out of the 55 base-pairings of the *B. subtilis* FMN sequence are G-A pairs.

[2] Table 3.11: ID36.1 is the *metI* SAM-I riboswitch in *B. subtilis* and has sequence identity of 76% with ID36 *yitJ B. subtilis* SAM-I riboswitch using BLAST©. Sequence location on Location on the *B. subtilis* str. 168 strain emb|AL009126.3 (1258304-1258462), forward strand.

[3] Table 3.11: CYK structural prediction under the BJK model and that of the MFE model via vienna©detect different alteration of the Magnesium riboswitch in *Salmonella enterica* serovar Typhimurium. Structural distance of the MFE prediction to the high Mg2^{+} and low Mg2^{+} structures are 28 and 120, respectively while they are 114 and 74, under CYK-based structural prediction of the BJK model. Sensitivity and specificity values for the BJK model prediction of the low Mg2^{+} conformation are 22% and 22%.

[4] Table 3.11: All mutations in expression platform of the *Salmonella* Cobalamin riboswitch tested here, disrupt pseudoknot formation in the encompassing structure. Results may not apply to our hypothesis.

[5] Table 3.11: One out of the 44 base-pairings of the *Bradyrhizobium* ROSE-P2 sequence is a G-G pair.

[6] Table 3.11: The ΔG83 mutant is one nucleotide shorter than the ROSE-P2 135nt-long wild type.

Table 3.12: Mutagenesis Results. Percentages (%) of mutants having lower value than both wild-type and non structural mutations. Base-pairing entropy (BJKbp), Structural entropy under BJK and RND models, (BJK) and (RND), and two-cluster average silhouette index of energy landscape (Sil) were investigated. Percentages were calculated as follows: For each riboswitch, the percentage of structural mutants (annotated by YES in Table 3.11) having lower values than both the wild type and non-structural mutations (annotated by NO in Table 3.11) were calculated. Then, the average of percentage is taken accross the six riboswitches of Table 3.11. For the case of the *Bradyrhizobium* ROSE-P2, entropy values were compared with -0.74 rather than zero for wild type, since the length of the 135nt-long riboswitch was decreased by 1 and this decrease in length is expected to have linear effect on structural entropy values.

Feature	(%)
BJKbp	77.8
BJK	61.1
Sil	58.3
RND	41.6

search, this number drops down to 2 (8.7%). Only Fluoride and preQ1 were identified correctly. It is worth noting that the aforementioned tool involves searching the intergenic regions of different organisms and the comparison to *ab initio* riboswitch identification may not be applicable.

Table 3.13: Ribex Performance. Performance of Ribex (Abreu-Goodger and Merino, 2005) to identify the riboswitches. riboswitch known/unknown denotes if the family of the riboswitch is known to the tool or not, if Yes, it implies, the search tool will search genomic regions corresponding to that riboswitch. Ribo. % denotes percentage of the test-set riboswitches correctly identified as a riboswitch.

Sequence	Num.	Ribo. %	Ribo.	Not Ribo.
riboswitch known	23	52.17	12	11
riboswitch unknown	23	8.7	2[1]	21
antisense	23	0	0	23

3.3 Discussion

Riboswitches are comprised of a diversity of biological functionality, as well as having different conformational dynamics. In this work, we made an attempt to characterize the potential for an alternative fold ubiquitous in various regulatory elements, regardless of their annotation and structural complexities. Secondary Structural entropy of the SCFG-modeled folding space of the RNA was used as one of the main features in this regard, based on the assumption that there should to be a relationship between theoretical diversity of possible folding scenarios (here, folding entropy) and the potential for the RNA to have an alternative stable secondary structure. We purposely refrained from including homologous sequences in our data set to avoid bias to a specific family of riboswitches. Regression approaches to estimate the structural entropy of the riboswitch with respect to various sequence and structural features, such as MFE, lead to higher classification performance in discriminating riboswitches from their antisense control, compared to classfiers that do not incorporate the structural entropy measure. We believe this increase in performance is significant, since it is least likely to be caused by primary structural biases; both sense and antisense sequences have the same GC composition and length. Other primary structural features were not included in our classifiers, since they may cause bias towards certain sequences. In fact, the inclusion of Uracil composition did not yield better results in most cases. We hypothesize that the folding entropy value of riboswitches may be a significant factor within the context of their length, GC-composition, and folding stability (here, MFE). Multinomial logistic regression based classifiers based on structural entropy were successfully designed as *ab initio* riboswitch identifiers.

Some of the challenges in our approach to develop *ab initio* riboswitch identifiers were choices of sequence segment and folding model. We found it very difficult to find a subset of sequence segments from riboswitches for our training set that had the highest structural entropy. These difficulties included but were not limited to high sensitivity of structural entropy to sequence length and location and the possible varying lengths of riboswitches that have alternative structures. We arbitrarily included varying lengths of riboswitches in our training set rather than constant length, since the performance of classifiers with constant length was either lower or similar to those with varying length.

The optimum length of a sequence segment that leads to identifying riboswitches can vary from one organism to another; Constant length of 100 nt segments for *E. coli* are more suitable, while 157 nt segments lead to higher performance for *B. subtilis* riboswitches. Results about sequence segments, however, had low significance due to low number of riboswitches tested in each case. We only propose that it may be possible that riboswitches from different organisms may have different ranges of sequence lengths over which alternative structure prediction becomes significant. Optimizing search parameters on a new organism sequence is potentially a difficult task. One alternative may be evaluating the behavior of entropy based classifiers on data sets that are peculiar to that organism. We have not explored this approach.

3.3.1 Choice of Model

Classification performance of sense/antisense, genome-wide sliding window tests, and mutagenesis all suggest that the BJK folding model is more sensitive to changes in the folding space than the structurally ambiguous RND model. The classification performance of the LMFEGCBJK model both on the test set and on the *B. subtilis* riboswitches is high given the right sequence segment is chosen. Also, the RND model does very poorly in distinguishing the folding space of riboswitch mutants from that of their wild types. On the other hand, binomial logistic regression based classification of sense and antisense of all riboswitches assigns slightly higher ROC area to the classifier that deploys the RND model (see Figure B.1). Furthermore, riboswitch identifiers based on the RND model are more robust in terms of sequence positioning than their BJK counterparts. The RND model only enforces Watson-Crick and G-U base-pairing and is fairly a simplistic structural model. The acceptable performance of the RND model in genome-wide approaches may be due to having less structural constraints than BJK. It may be possible that training secondary structural folding models to predict RNA secondary structures comes at the cost of loss in folding information.

3.3.2 Genome-wide Analysis

Sequence segments predicted to have potential for alternative fold for the two *B. subtilis* and *E. coli* intergenic regions are presented in the Results section. Many such hits fell immediately upstream of operons, which could be indicative of *cis*-regulation. A rigorous analysis of all predictions under various hypotheses, however, falls outside the scope of this work. Here we discuss only a few significant hits.

The *cotH* Gene

The top two sequence segments predicted to be riboswitches are both upstream of *cotH* gene and in close proximity of one another (see top two rows of Table B.7). In fact, a 628 nt long segment is classified to be a riboswitch (four consecutive sequence segments). the 5' half of this segment, {3717412 nt - 3717725 nt}, contains the top two hits which are also predicted to be riboswitches by the model LMFEGCBJK in position {3717098 nt - 3717725 nt} on the complementary strand of *B. subtilis*. Naclerio et al. (1996) discuss possible regulation in the vicinity of *cotH* gene. They also stated that no homology to this gene was revealed in the sequences presented in the data bank at the time. They hypothesized that this gene plays an important role in the formation of the spore coat. A more recent paper (Giglio et al., 2011) reports about the *cotH* promoter mapping 812 bp upstream of beginning of its coding region. This region covers the top two hits we have. In fact, 200 nt scan reveals that many consecutive segments belonging to this region have significant RND entropy values (<0.05). Most interestingly, the segment with highest RND entropy value on a genome-wide level and under the 200 nt window occurs 399 nt upstream of the *cotH* gene at location {3717398-3717597 +}. The authors also talk about *cotG* and *cotH* genes and that they are both divergently transcribed by σ-K and a potential for extensive secondary RNA structures in this unusually long untranslated region. The *cotG* is located in the forward strand. There are also many hits around 2000 nt upstream of the *cotG*-containing operon under various sliding window tests. An interesting observation about the nucleotide composition of the top hit reveals that it uniquely contains periodic runs of 6 consecutive thymines with periodicity of 12 and 15 interchangeably. A search for similar runs of thymines was done on both strands of *B. subtilis* by relaxing perdiodicity to 10 nt to 18 nt and constraining it to having at least six consecutive runs of 6-thymines using the pattern locator software (Mrazek and Xie, 2006). The only two hits were found both on the reverse strand and overlapping with the top hit:{3717502-3717606} and {3717367-3717468}.

The most significant structural entropy value for the longest window size (200 nt) on the *B. subtilis* genome occurred in an unusually extensive secondary structure within that genome. It may be possible that RNA structures contain segments having significantly high secondary structural space (here shannon's entropy) on a genome-wide scale. This implies that long ncRNAs potentially have a uniquely high number of secondary structural conformations. This unusually high secondary structural diversity may be related to their regulatory role. We have not yet examined the secondary structural space of other long secondary structures in various organisms. The significantly high secondary-structural-entropy feature, however, may be typical of other longer secondary strucures. In a recent study on the newly discovered class of RNAs known as long ncRNAs (lncRNA), Cloutier et al. (2013) show that yeast lncRNAs are involved in the timing of gene expression. Hence, it may be possible that the proposed lncRNA-dependent *quick shift* of gene expression be related to their luxury of having a potential for diverse secondary structural conformations.

The BSU tRNA 75 Operon

The sequence segment with highest classification probability that also has signficant MFE and entropy values is located about 2277 nt of the upstream region of the BSU tRNA 75 Operon. The antisense control of this segment is located in a putative transcriptional regulator. It is interesting, however, that this hit occurs upstream of a tRNA operon. A 200 nt scan reveals more hits upstream of this operon that have significant entropy values some of which are closer to the tRNA operon (around 2000 nt upstream). From locations of tRNA operon (Chan and Lowe, 2009), it can be seen that out of the five consecutive tRNA genes with isotypes Glu, Val, Thr, Tyr, and Gln, The Thr operon has attenuation (Kolter and Yanofsky, 1982). Although the long distance from the downstream translation start codon does not make this a reliable riboswitch prediction, the significance of hits in the intergenic region upstream of the Thr gene and the fact that the other top hit in our classification approach was located in a long RNA, suggest the possibility that there may be a long regulatory RNA residing upstream of the mentioned tRNA operon, raising the interesting possibility of a putative riboswitch regulating an attenuation mechanism.

lysP

One of the most significant hits in our classification under the 157 nt scan occurs immediately upstream of the *lysP* gene. The segment corresponding to this location also has the most significant (highest) RND entropy value while having significantly low MFE (p-Val. <0.05) on a genome-wide level. This is also true for the 150 nt window scan. Furthermore, the 200 nt scan assigns significantly high structural entropy (RND p-Val. <0.05) as well as classification probability of higher than 0.8 for this location. The 150 nt-long segment is located at {3421066-3421215 -} between the lysine permease and BSU MISC RNA 54. Other adjacent hits that overlap BSU MISC RNA 54 do not have such high entropy or classification probability. It may be possible that this segment plays a crucial role in regulating the downstream gene.

3.4 conclusion

In this work, structural entropy was investigated for characterization of RNA potential for alternative folds. Classifiers based on structural entropy optimized via sequence and structural features were devised to discriminate between the putative riboswitch and the antisense control. They were also used as *ab initio* riboswitch identifiers in *B. subtilis* and *E. coli*. It was shown that secondary structural entropy is an effective feature for capturing folding characteristics of riboswitches as a whole and could be a potential alternative method to homology searches. In addition, although structural entropy is very sensitive to model parameters and sequence features, when it comes to longer sequences (>150 nt), simplistic folding models tend to have a very consistent and robust result in distinguishing extensive secondary structures from other intergenic regions on genome-wide scale, regardless of test parameters. Application of structural entropy in finding RNA genes that regulate, especially for longer sequences, may be very rewarding.

3.5 Materials and Methods

3.5.1 Data Collection

Sequences with concrete evidence of alternative structures were gathered from the literature (See Table 3.1). Prokaryotic sequences believed not to have structure were selected from *E. coli* and are listed in Table B.3 as negative set. 30 genome locations corresponding to σ-70 transcription factor binding sites that are less than 80 nt upstream of their corresponding start codon were randomly chosen from *E. coli* such that they are fairly evenly distributed across the genome. Data was manually gathered from the EcoCyc website (http://ecocyc.org/).

3.5.2 Classification

Preparing the Positive Control set: The criterion for building the positive control set was taking the minimum-length sub-sequence for the corresponding riboswitch with evidence for alternative structures. Comprehensive structure information was not available for certain sequences. We decided to include them to increase our data set size. The structures of most sequences were experimentally validated, although a few structures of the riboswitches were inferred in combination with structural homology approaches. Only the expression platform components for the Cobalamin riboswitches were used, since they contain alterations; a typical riboswitch has an aptamer and an expression platform component, where the aptamer binds to the ligand, triggering allosteric rearrangement of the conformation of the expression platform component of the riboswitch which in turn regulates the expression of the downstream gene. Cobalamin riboswitches are also significantly longer than other sequences, e.g. *Salmonella enterica serovar Typhimurium*'s Cobalamin riboswitch was over 300 nt long. Including such long sequences could have been problematic, both for sensitivity of structural entropy on sequence length and the fact that RNA structures longer than 200 nt are usually predicted with low confidence under SCFG models as well as computational constraints. Also, certain sequences were excluded from the test. In the column `Grouping` of Table 3.1 we denote `None` for such sequences. Excluded sequences are as follows: *Salmonella* ATP regulatory element, located in the *mgtM* gene before the *mgtCBR* operon, was excluded since it was the only RNA in our set that had complete overlap with codons (Lee and Groisman, 2012). *Synechococcus sp. CC9902* Downstream-peptide motif was excluded, since evidence for alteration was not available. *T. tengcongensis* glmS ribozyme-riboswitch was excluded, since the glmS ribozyme does not undergo "large conformational changes concomitant with ligand binding" (Ferre-D'Amare, 2010) and is the only RNA element in our gathered data that functions as a self-cleaving ribozyme upon binding to glucosamine-6-phosphate (GlcN6P) ligand (Winkler et al., 2004). *Synechococcus elongatus* glnA motif was excluded, since no evidence of alteration was available. *Schistosoma Mansoni* Hammerhead type I ribozyme was excluded, since its structure does not alter. The pseudoknotted *marine metagenome* Hammerhead type II ribozyme was also excluded, since there was no evidence of alteration of the secondary structure. Finally, *Bacillus subtilis* yxkD motif was excluded, since there was no concrete biological evidence for it being a functional riboswitch or ribo-regulator, although it is predicted partially to have an alternative structure (Barrick et al., 2004).

Training and Test Sets: The positive control set was divided into the training and test sets. Distributions of training and test sets were similar in differential entropy vs. differential MFE (see Figures B.2 and B.3). While most gathered sequences were in the two organisms, *B. subtilis* and *E. coli*, they cover a variety of biological functions and structures. We were interested in an *ab initio* method that can identify potential for the RNA to have an alternative secondary structure from a thermodynamic perspective regardless of a specific function or a secondary structural conformation. Hence, the categorization was done such that each of the training and test sets would contain as diverse sequences and structures as possible. Furthermore, the training sequences contain those from *E. coli* while the test set contains those of *B. subtilis*. For those riboswitches that did not exist in both gram-positive and gram-negatives, they were evenly distributed between the two tests. Division of data into training and test sets was a compromise between having as diverse riboswitches as possible while being able to assess significance of classification on riboswitches from phylogenetically distant organisms, namely the gram-positive *B. subtilis* and the gram-negative *E. coli*. In the column `Grouping` of Table 3.1, the categorization of each sequence is shown. There are a total of 29 sequences in the training set and 23 sequences, in the test set. The 30 *E. coli* UTRs were divided into sets of 17 and 13 for training and test sets, respectively. The categorization was selected such that for an extension of 100 nt UTRs upstream of their corresponding start codons, `GC`-composition, and the minimum free energy having similar distribution in both sets. A further internal control was the antisense sequence of the riboswitch, adding additional sets of sequences of size 29 and 23 sequence to the training and test sets, respectively. Various classifications in this work always use antisense sequences of riboswitches of identical length for training and test sets, unless indicated otherwise.

Classification Criterion: Classification probabilities of having an alternative fold (riboswitch), possibly only one fold (antisense), or no structure (*E. coli* UTR) were assigned to each sequence based on multinomial logistic regression of sequences in the training. SPSS 16.0©software was used to estimate the corresponding parameter vectors. Entropy calculations were done according to Manzourolajdad et al. (2013). Two different lightweight context-free secondary structural models were used as folding distribution models. The first model, denoted here as BJK, was developed by Knudsen/Hein and originally used in the Pfold package (Knudsen and Hein, 1999, 2003). The structurally unambiguous grammar was subsequently used in Dowell and Eddy (2004) under the name G6 to predict RNA secondary structure using different training sets for RNA secondary structures. Model parameters used here correspond to the `benchmark`-trained version of this grammar (Dowell and Eddy, 2004) and will be referred to as the BJK model. Average sensitivity and specificity values for the BJK model on the test set of riboswitches are 75.6 and 76.3, respectively. The second model, denoted here as RND was introduced in Manzourolajdad et al. (2013) under the name RND10. This model consists of a structurally ambiguous simple grammar with symmetric rules and probabilities set according to Manzourolajdad et al. (2013). Also, an effort was done to convert non-stacking heavyweight grammars from Nawrocki and Eddy (2013). Such grammars aim at mirroring the state-of-the-art thermodynamic folding models and are extremely sophisticated, requiring their specific software implementation. The translation of these models into our simple implementation eliminated many of its features. What was left did not yield original performance of the model to predict RNA secondary structure, nor was its entropy showing any significant performance in the classifier (data not shown). Minimum-free-energy calculation was done by Vienna©Software Package (Hofacker, 2003) using default parameters. Base-pairing entropy for the BJK model, denoted here as BJKbp, was calculated as defined in Huynen et al. (1997) [implementation by Manzourolajdad et al. (2013)]. The two-cluster average Silhouette index of energy landscape, denoted here as *Sil*, was calculated according to the pipeline used in Quarta et al. (2012) with the exception that we did not account for pseudoknots and only used MFE predictions of the Vienna©Software Package for prediction of structures.

Ribex (Abreu-Goodger and Merino, 2005) riboswitch identification tool that uses annotations was also used for performance comparisons (see Table 3.13). We also tried to explore GC composition information upstream of gathered sequences relative to that in the riboswitch which did not yield significantly better results. Sequence-similarity method such, as BLAST©and profile Hidden Markov Models were also examined as classifiers with the mentioned training and test sets. The pipeline was implemented according to Singh et al. (2009). These methods did not result in significant classification performance even after lowering the corresponding threshold to insignificant values.

3.5.3 Genome-wide scan of the *B. subtilis* and *E. coli* intergenic regions

Bacillus subtilis subsp. subtilis str. 168 (taxid:224308) and *Escherichia coli str. K-12 substr. MG1655* (GenBank©ID: U00096.2) were downloaded from the National Center for Biotechnology Information (NCBI)©(Sayers et al., 2009; Benson et al., 2009). The newer version of *E. coli* str. K-12 genome (gb|U00096.3) was not used, since operon and σ-70 UTR locations were given in the old version. Corresponding locations of *E. coli* riboswitches in the old version were used, where necessary. The operon-location information file for *B. subtilis* was downloaded from Taboada et al. (2012). Candidates consisted of sequence segments of lengths 100 nt, 150 nt, and 157 nt. Each intergenic region was divided into segments of such length such that the most downstream segment in each intergenic region ends at the start codon. Only intergenic regions higher than 150 nt were considered. The same process was applied to the *E. coli*

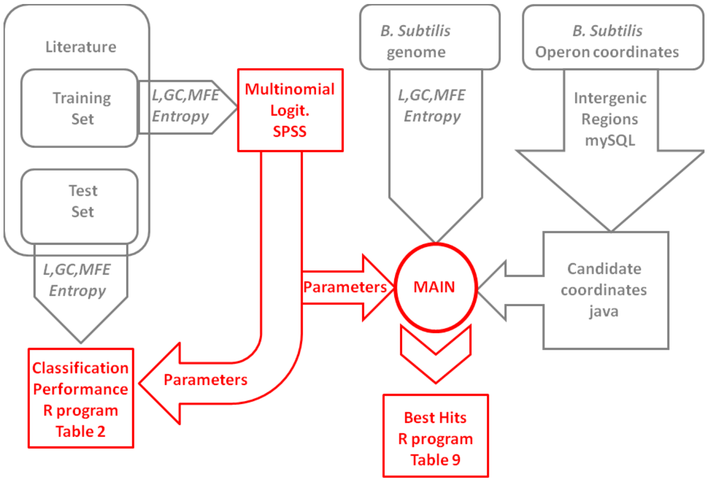

Figure 3.3: Riboswitch Identification Pipeline. Riboswitch classification and identification pipeline.

genome. Operon locations for the *E. coli* genome were downloaded from the RegulonDB website (Salgado et al., 2013). Operon locations in *E. coli* also contained RNA elements in our data set. Hence, results for the genome-wide scan of *E. coli* did not contain any sequence within the operon locations and only contained non annotated regions. MySQL, java, and R programming languages were used in various phases of our pipeline (see figure 3.3).

Chapter 4

Conclusion

In this work, we examined the conformational dynamics of non-coding RNA sequences with a focus on discovering riboswitches using their information-theoretic folding space and stochastic context-free grammar folding models, referred to here as structural entropy. We focused on the folding space of riboswitches mainly because they each typically have potential for forming mutually exclusive secondary structures that are biologically functional. Furthermore, riboswitches do not require protein factors for their regulatory activities, increasing the likelihood that the folding space of a given riboswitch may contain critical information about its potential for having such alternative folds. Models used here included those trained to predict the secondary structures of various RNA sequences, as well as those having minimal number of parameters and assumptions.

As a theoretical contribution, we offered a computationally convenient algorithm for calculating the Shannon entropy of *structurally ambiguous* grammars generalized for all non-stacking SCFG-modeled folding models (Manzourolajdad et al., 2013). In Chapter 2, we also showed that structural entropy can indeed be significantly higher, as well as significantly lower, depending on the family of the RNA and its GC-composition. A fresh approach using a notion of significantly higher entropy of certain ncRNAs with respect to random sequences and structures was then used in chapter 3 as a potential *ab initio* method for characterization and identification of riboswitches versus their antisense sequences and other intergenic regions belonging to the organism hosting the riboswitches. Models and relevant potential features for this goal were based on results of chapter 2. A data set of riboswitches[1] with diverse structures as well as functions was collected with the special consideration to avoiding homologous sequences. Riboswitch sequence segments of different lengths were selected based on the criterion of alternative fold as discussed in Materials and Methods of chapter 3.

The evaluation of folding space of collected riboswitches versus their antisense sequences showed that riboswitches on average have a distinct folding space than their antisense controls. Binomial logistic regression using features such are Length, GC-composition, GU-composition, and an index derived from clustering the RNA energy landscape (Quarta et al., 2012) resulted in 75% true positive rate and 25% false positive rate. In an analysis based on the RNA energy landscape, calculations were less computationally convenient than those for structural entropy while classifiers that incorporated RNA energy landscape clustering features lead to higher performance in sense/antisense discrimination, compared to those that incorporated structural entropy instead. Multinomial logistic regression classifiers incorporating structural entropy and other features were then devised based on known riboswitches, their antisense control, and a collection of short UTR sequences corresponding to σ-70 in *E. coli*. We used regression values of gram-negative dominated data set of riboswitches as our training set and tested it on our gram-positive dominated test set.

The BJK folding model was developed by Knudsen/Hein and originally used in the Pfold package (Knudsen and Hein, 1999, 2003). It was subsequently used in Dowell and Eddy (2004) under the name G6 to predict RNA secondary structure using different training sets for RNA secondary structures. Model parameters used here correspond to the `benchmark`-trained version of this grammar (Dowell and Eddy, 2004). The second model that was incorporated in the regression-based classification was denoted here as RND and is described in chapter 2 under the name RND10. This model consists of a structurally ambiguous simple grammar with symmetric rules and probabilities set according to Manzourolajdad et al. (2013). Classifiers LMFEGCBJK and LMFEGCRND use structural entropy under the BJK and RND models, respectively. The LMFEGCBJK model had significantly better performance in distinguishing riboswitches form their antisense and short UTRs compared to the LMFEGCRND and LMFEGC models. The most challenging part of our evaluation was the choice of length of the riboswitch. One one hand, structural entropy is very sensitive to various sequence and structural features (data not shown). On the other hand,

[1] Certain ribo-regulators having alternative secondary structures such as the tryptophan element were included. We refer to all such sequences in our data set as riboswitches for simplicity.

sense/antisense classification performance was observed to vary differently for different riboswitches upon change of length; performance on some riboswitches was very sensitive to choice of length and location of their corresponding sequence segment while this was not the case for other riboswitches (data not shown). Variable-length sequences were preferred over constant-length data sets, since classification performance for the former was slightly higher than the latter.

Genome-wide scans of the intergenic regions of the *B. subtilis* genome using various choices of classifiers and lengths of sliding windows, revealed that the model LMFEGCRND for sliding window of length 157 nt yields higher performance of riboswitch identification over other choices on lengths and models. Although the more detailed BJK folding model yields better results upon knowledge of riboswitch location (the LMFEGCBJK classifier), when it comes to blind searches, the simple RND folding model (incorporated in LMFEGCRND classifier), outperforms other models. The optimum length for riboswitch identification in the gram negative *E. coli* was found to be 100 nt. The high performance of the RND model and comparison of optimum length of sliding windows between *B. subtilis* and *E. coli* raised several observations about our classifiers.

First, the optimum length of sequence segment used to identify riboswitches based on their structural entropy could be organism dependent. It may be possible that factors such as different rates of transcription by RNA polymerase in different organisms affects the length of the segment of the riboswitch over which alternative structures exist. The number of riboswitches considered, however, is too small to derive any statistically significant conclusion. Second, sequences surrounding riboswitches may be informative of potential for alternative fold and should be taken into consideration. Third, the fact that certain positioning of sequence segments may increase riboswitch identification performance for folding models trained to predict RNA structure (here, the BJK model), could be due to the fact that these models are biased towards certain conformations in the training set and does not necessarily imply their merit. As a larger set becomes available, a better choice of segment length may come available. This argument is strengthened by the fact that the simple RND model with arbitrary parameters was more robust in genome-wide searches blind to the exact position of riboswitches. Given both our genome-wide results and results of chapter 2 for various models, it may that training SCFG models to best predict RNA secondary structure using current optimization methods is done at the cost of losing critical *information* about the folding diversity intrinsic to certain regulatory elements.

Interestingly, the simple RND folding model and using a 200 nt sliding-window scan of 100 nt overlap was shown to be sufficient in identification of extensive secondary structures in *B. subtilis* intergenic regions. A scan of 200 nt sliding window assigns the highest RND entropy value to the unusually long (812 bp) untranslated region upstream of the *cotH* gene. MFE values of corresponding sequence segments did not have such high significance on the genome-wide level. We consider this finding significant, since the RND model was *a priori* designed, and 200 nt was sufficient to show the distinction of the mentioned untranslated region. We conclude that longer secondary structures may have signatures of significantly high structural entropy values within them with respect to other intergenic regions of their organism. The fact that this significance turns out to be high, rather than low, is subject to various speculations about folding space or RNA structure, in general. It may simply be that longer RNAs are more likely to have alternative structures.

4.1 Future Work

In this work, we showed that simple SCFG models can be very informative about the folding space of riboswitches as a whole. Our data set selection was constrained to experimentally validated sequences. Other choices of riboswitches, for instance larger data sets that also include non-validated riboswitches, may lead to better results for the formidable task of *ab initio* riboswitch identification. The approaches presented in this work are only a first step towards this endeavor. In addition, most of our conclusions about the genome-wide significance of structural entropy is based on the performance of an arbitrary RND model. Other choices of parameters aimed at characterizing the folding space of riboswitches or extensively long RNA secondary structures, rather than trained to best predict their final conformation, may give a clearer view of their folding space characteristics. Finally, the Shannon entropy of RNA secondary structural space, is a fairly new quantitative measure and its relationship with other features such as MFE, pseudoknots, sequence length and composition, tertiary interactions, etc... is not very well understood. The reason for its high genome-wide significance for RNA sequences of longer lengths is one interesting question that remains to be answered.

Bibliography

Abreu-Goodger, C., Merino, E., 2005. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Res 33, W690–2.

Adami, C., Ofria, C., Collier, T.C., 2000. Evolution of biological complexity. Proc Natl Acad Sci U S A 97, 4463–8.

Altschul, S.F., Erickson, B.W., 1985. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol Biol Evol 2, 526–38.

Ames, T.D., Breaker, R.R., 2011. Bacterial aptamers that selectively bind glutamine. RNA Biol 8, 82–9.

Ames, T.D., Rodionov, D.A., Weinberg, Z., Breaker, R.R., 2010. A eubacterial riboswitch class that senses the coenzyme tetrahydrofolate. Chem Biol 17, 681–5.

Babitzke, P., Gollnick, P., 2001. Posttranscription initiation control of tryptophan metabolism in *Bacillus subtilis* by the trp RNA-binding attenuation protein (TRAP), anti-TRAP, and RNA structure. J Bacteriol 183, 5795–802.

Babitzke, P., Schaak, J., Yakhnin, A.V., Bevilacqua, P.C., 2003. Role of RNA structure in transcription attenuation in *Bacillus subtilis*: the *trpEDCFBA* operon as a model system. Methods Enzymol 371, 392–404.

Baker, J.L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R.B., Breaker, R.R., 2012. Widespread genetic switches and toxicity resistance proteins for fluoride. Science 335, 233–5.

Barash, D., Sikorski, J., Perry, E.B., Nevo, E., Nudler, E., 2006. Adaptive mutations in RNA-based regulatory mechanisms: computational and experimental investigations. Israel Journal Of Ecology and Evolution 52, 263–79.

Barrandon, C., Spiluttini, B., Bensaude, O., 2008. Non-coding RNAs regulating the transcriptional machiner. Biol. Cell 100, 83–95.

Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J.K., Breaker, R.R., 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc Natl Acad Sci U S A 101, 6421–6.

Batey, R.T., Gilbert, S.D., Montange, R.K., 2004. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature 432, 411–5.

Batey, R.T., Rambo, R.P., Doudna, J.A., 1999. Tertiary motifs in RNA structure and folding. Angew Chem Int Ed Engl 38, 2326–43.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2009. Genbank. Nucleic Acids Res 37, 21.

Bernauer, J., Huang, X., Sim, A.Y., Levitt, M., 2011. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. RNA 17, 1066–1075.

Bocobza, S., Adato, A., Mandel, T., Shapira, M., Nudler, E., Aharoni, A., 2007. Riboswitch-dependent gene regulation and its evolution in the plant kingdom. Genes Dev 21, 2874–9.

Boyapati, V.K., Huang, W., Spedale, J., Aboul-Ela, F., 2012. Basis for ligand discrimination between on and off state riboswitch conformations: the case of the SAM-I riboswitch. RNA 18, 1230–43.

Brannvall, M., Mattsson, J.G., Svard, S.G., Kirsebom, L.A., 1998. RNase p RNA structure and cleavage reflect the primary structure of tRNA genes. J Mol Biol 283, 771–83.

Breaker, R.R., 2012. Riboswitches and the RNA world. Cold Spring Harb Perspect Biol 4, a003566.

Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G., 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 21, 167–95.

Butler, E.B., Xiong, Y., Wang, J., Strobel, S.A., 2011. Structural basis of cooperative ligand binding by the glycine riboswitch. Chem Biol 18, 293–8.

Canny, M.D., Jucker, F.M., Kellogg, E., Khvorova, A., Jayasena, S.D., Pardi, A., 2004. Fast cleavage kinetics of a natural hammerhead ribozyme. J Am Chem Soc 126, 10848–9.

Cech, T.R., Damberger, S.H., Gutell, R.R., 1994. Representation of the secondary and tertiary structure of group I introns. Nat Struct Biol 1, 273–80.

Chan, C.Y., Ding, Y., 2008. Boltzmann ensemble features of RNA secondary structures: a comparative analysis of biological RNA sequences and random shuffles. J Math Biol 56, 93–105.

Chan, P.P., Lowe, T.M., 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37, D93–7.

Chomsky, N., 1959. On certain formal properties of grammars. Information and Control 2, 137–167.

Chowdhury, S., Maris, C., Allain, F.H., Narberhaus, F., 2006. Molecular basis for temperature sensing by an RNA thermometer. EMBO J 25, 2487–97.

Christiansen, M.M., Duffy, K.R., du Pin Calmon, F., Medard, M., 2013. Brute force searching, the typical set and guesswork, in: Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on, pp. 1257–1261.

Clote, P., Ferre, F., Kranakis, E., Krizanc, D., 2005. Structural rna has lower folding energy than random rna of the same dinucleotide frequency. RNA 11, 578–91.

Cloutier, S.C., Wang, S., Ma, W.K., Petell, C.J., Tran, E.J., 2013. Long noncoding RNAs promote transcriptional poising of inducible genes. PLoS Biol 11, e1001715.

Cochrane, J.C., Lipchock, S.V., Strobel, S.A., 2007. Structural investigation of the *GlmS* ribozyme bound to its catalytic cofactor. Chem Biol 14, 97–105.

Corbino, K.A., Barrick, J.E., Lim, J., Welz, R., Tucker, B.J., Puskarz, I., Mandal, M., Rudnick, N.D., Breaker, R.R., 2005. Evidence for a second class of s-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. Genome Biol 6, R70.

Cover, T.M., Thomas, J.A., . Elements of information theory. Wiley-Interscience, Hoboken, N.J. 2nd edition.

Cromie, M.J., Shi, Y., Latifi, T., Groisman, E.A., 2006. An RNA sensor for intracellular mg(2+). Cell 125, 71–84.

Dann, C. E., r., Wakeman, C.A., Sieling, C.L., Baker, S.C., Irnov, I., Winkler, W.C., 2007. Structure and mechanism of a metal-sensing regulatory RNA. Cell 130, 878–92.

D'Haeseleer, P., 2006. What are DNA sequence motifs? Nature Biotechnology 24, 423–5.

Ding, Y., Chan, C.Y., Lawrence, C.E., 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA 11, 1157–66.

Ding, Y., Lawrence, C.E., 2003. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res 31, 7280–301.

Do, C.B., Woods, D.A., Batzoglou, S., 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics 22, e90–8.

Dowell, R.D., Eddy, S.R., 2004. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics 5, 71.

Du, X., Wang, E.D., 2003. Tertiary structure base pairs between D- and TpsiC-loops of *Escherichia coli* tRNA(leu) play important roles in both aminoacylation and editing. Nucleic Acids Res 31, 2865–72.

Durbin, R., 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.

Eddy, S.R., 2001. Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2, 919–29.

Edwards, A.L., Reyes, F.E., Heroux, A., Batey, R.T., 2010. Structural basis for recognition of s-adenosylhomocysteine by riboswitches. RNA 16, 2144–55.

Edwards, T.E., Ferre-D'Amare, A.R., 2006. Crystal structures of the *thi*-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. Structure 14, 1459–68.

Ferre-D'Amare, A.R., 2010. The *glmS* ribozyme: use of a small molecule coenzyme by a gene-regulatory RNA. Q Rev Biophys 43, 423–47.

Freyhult, E., Moulton, V., Clote, P., 2007. RNAbor: a web server for RNA structural neighbors. Nucleic Acids Res 35, W305–9.

Fuchs, R.T., Grundy, F.J., Henkin, T.M., 2006. The S(MK) box is a new SAM-binding RNA for translational regulation of sam synthetase. Nat Struct Mol Biol 13, 226–33.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., Bateman, A., 2009. Rfam: updates to the RNA families database. Nucleic Acids Res 37, D136–40.

Gardner, P.P., Wilm, A., A Washietl, S., 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res 33, 2433–9.

Garst, A.D., Heroux, A., Rambo, R.P., Batey, R.T., 2008. Crystal structure of the lysine riboswitch regulatory mRNA element. J Biol Chem 283, 22347–51.

Giglio, R., Fani, R., Isticato, R., De Felice, M., Ricca, E., Baccigalupi, L., 2011. Organization and evolution of the *cotG* and *cotH* genes of *Bacillus subtilis*. J Bacteriol 193, 6664–73.

Gilbert, S.D., Love, C.E., Edwards, A.L., Batey, R.T., 2007. Mutational analysis of the purine riboswitch aptamer domain. Biochemistry 46, 13297–309.

Gluick, T.C., Gerstner, R.B., Draper, D.E., 1997. Effects of mg2+, k+, and h+ on an equilibrium between alternative conformations of an RNA pseudoknot. J Mol Biol 270, 451–63.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A., 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33, D121–4.

Grundy, F.J., Henkin, T.M., 1998. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. Mol Microbiol 30, 737–49.

Grundy, F.J., Henkin, T.M., 2006. From ribosome to riboswitch: control of gene expression in bacteria by RNA structural rearrangements. Crit Rev Biochem Mol Biol 41, 329–38.

Guerrier-Takada, C., Altman, S., 1993. A physical assay for and kinetic analysis of the interactions between m1 RNA and tRNA precursor substrates. Biochemistry 32, 7152–61.

Hall, M.N., Gabay, J., Debarbouille, M., Schwartz, M., 1982. A role for mRNA secondary structure in the control of translation initiation. Nature 295, 616–8.

Haller, A., Altman, R.B., Souliere, M.F., Blanchard, S.C., Micura, R., 2013. Folding and ligand recognition of the TPP riboswitch aptamer at single-molecule resolution. Proc Natl Acad Sci U S A 110, 4188–93.

Henkin, T.M., 2008. Riboswitch RNAs: using RNA to sense cellular metabolism. Genes Dev 22, 3383–90.

Hofacker, I.L., 2003. Vienna RNA secondary structure server. Nucleic Acids Res 31, 3429–31.

Hollands, K., 2012. Riboswitch control of rho-dependent transcription termination. Proc Natl Acad Sci U S A 109, 5376.

Huang, L., Ishibe-Murakami, S., Patel, D.J., Serganov, A., 2011. Long-range pseudoknot interactions dictate the regulatory response in the tetrahydrofolate riboswitch. Proc Natl Acad Sci U S A 108, 14801–6.

Huynen, M., Gutell, R., Konings, D., 1997. Assessing the reliability of RNA folding using statistical mechanics. J Mol Biol 267, 1104–12.

Kazantsev, A.V., Pace, N.R., 2006. Bacterial RNase P: a new view of an ancient enzyme. Nat Rev Microbiol 4, 729–40.

Klein, D.J., Edwards, T.E., Ferre-D'Amare, A.R., 2009. Cocrystal structure of a class I preQ1 riboswitch reveals a pseudoknot recognizing an essential hypermodified nucleobase. Nat Struct Mol Biol 16, 343–4.

Klein, D.J., Ferre-D'Amare, A.R., 2009. Crystallization of the *glmS* ribozyme-riboswitch. Methods Mol Biol 540, 129–39.

Knudsen, B., Hein, J., 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics 15, 446–54.

Knudsen, B., Hein, J., 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucl Acids Res 31, 3423–8.

Kolter, R., Yanofsky, C., 1982. Attenuation in amino acid biosynthetic operons. Annu Rev Genet 16, 113–34.

Kwon, M., Strobel, S.A., 2008. Chemical basis of glycine riboswitch cooperativity. RNA 14, 25–34.

Lee, E.J., Groisman, E.A., 2012. Control of a *Salmonella* virulence locus by an ATP-sensing leader messenger RNA. Nature 486, 271–5.

Loh, E., Dussurget, O., Gripenland, J., Vaitkevicius, K., Tiensuu, T., Mandin, P., Repoila, F., Buchrieser, C., Cossart, P., Johansson, J., 2009. A trans-acting riboswitch controls expression of the virulence regulator PrfA in *Listeria monocytogenes*. Cell 139, 770–9.

Lu, C., Ding, F., Chowdhury, A., Pradhan, V., Tomsic, J., Holmes, W.M., Henkin, T.M., Ke, A., 2010. SAM recognition and conformational switching mechanism in the *Bacillus subtilis yitJ* S box/SAM-I riboswitch. J Mol Biol 404, 803–18.

Lu, C., Smith, A.M., Fuchs, R.T., Ding, F., Rajashankar, K., Henkin, T.M., Ke, A., 2008. Crystal structures of the SAM-III/S(MK) riboswitch reveal the SAM-dependent translation inhibition mechanism. Nat Struct Mol Biol 15, 1076–83.

Lu, Y., Turner, R.J., Switzer, R.L., 1996. Function of RNA secondary structures in transcriptional attenuation of the *Bacillus subtilis pyr* operon. Proc Natl Acad Sci U S A 93, 14462–7.

Mandal, M., Lee, M., Barrick, J.E., Weinberg, Z., Emilsson, G.M., Ruzzo, W.L., Breaker, R.R., 2004. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. Science 306, 275–9.

Manzourolajdad, A., Wang, Y., Shaw, T.I., Malmberg, R.L., 2013. Information-theoretic uncertainty of SCFG-modeled folding space of the non-coding RNA. Journal of Theoretical Biology 318, 140–163.

Martick, M., Scott, W.G., 2006. Tertiary contacts distant from the active site prime a ribozyme for catalysis. Cell 126, 309–20.

Mathews, D.H., 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA 10, 1178–1190.

McCaskill, J.S., 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29, 1105–19.

Mellin, J.R., Tiensuu, T., Becavin, C., Gouin, E., Johansson, J., Cossart, P., 2013. A riboswitch-regulated antisense RNA in listeria monocytogenes. Proc Natl Acad Sci U S A 110, 13132–7.

Merino, E., Yanofsky, C., 2005. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet 21, 260–4.

Meyer, M.M., Ames, T.D., Smith, D.P., Weinberg, Z., Schwalbach, M.S., Giovannoni, S.J., Breaker, R.R., 2009. Identification of candidate structured RNAs in the marine organism 'Candidatus Pelagibacter ubique'. BMC Genomics 10, 268.

Meyer, M.M., Roth, A., Chervin, S.M., Garcia, G.A., Breaker, R.R., 2008. Confirmation of a second natural preQ1 aptamer class in *Streptococcaceae* bacteria. RNA 14, 685–95.

Miklos, I., Meyer, I.M., Nagy, B., 2005. Moments of the Boltzmann distribution for RNA secondary structures. Bull Math Biol 67, 1031–47.

Mironov, A.S., Gusarov, I., Rafikov, R., Lopez, L.E., Shatalin, K., Kreneva, R.A., Perumov, D.A., Nudler, E., 2002. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. Cell 111, 747–56.

Montange, R.K., Batey, R.T., 2006. Structure of the s-adenosylmethionine riboswitch regulatory mRNA element. Nature 441, 1172–5.

Morris, K.V., 2008. RNA and the Regulation of Gene Expression: A Hidden Layer of Complexity. Caister Academic Press.

Morris, K.V., 2012. Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection. Caister Academic Press.

Mrazek, J., Xie, S., 2006. Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences. Bioinformatics 22, 3099–100.

Naclerio, G., Baccigalupi, L., Zilhao, R., De Felice, M., Ricca, E., 1996. Erratum. *Bacillus subtilis* spore coat assembly requires *cotH* gene expression. J Bacteriol 178, 6407.

Nahvi, A., Sudarsan, N., Ebert, M.S., Zou, X., Brown, K.L., Breaker, R.R., 2002. Genetic control by a metabolite binding mRNA. Chem Biol 9, 1043.

Nawrocki, E.P., Eddy, S.R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933–5.

Nechooshtan, G., 2009. A pH-responsive riboregulator. Genes Dev 23, 2650.

Neidhardt, F.C., 1996. *Escherichia coli* and *Salmonella*: cellular and molecular biology. volume 1. ASM Press, Washington, D.C.. 2nd edition.

Niranjanakumari, S., Stams, T., Crary, S.M., Christianson, D.W., Fierke, C.A., 1998. Protein component of the ribozyme ribonuclease p alters substrate recognition by directly contacting precursor tRNA. Proc Natl Acad Sci U S A 95, 15212–7.

Nocker, A., Hausherr, T., Balsiger, S., Krstulovic, N.P., Hennecke, H., Narberhaus, F., 2001. A mRNA-based thermosensor controls expression of rhizobial heat shock genes. Nucleic Acids Res 29, 4800–7.

Nudler, E., 2006. Flipping riboswitches. Cell 126, 19–22.

Oppenheim, D.S., Yanofsky, C., 1980. Translational coupling during expression of the tryptophan operon of *Escherichia coli*. Genetics 95, 785–95.

Perreault, J., Weinberg, Z., Roth, A., Popescu, O., Chartrand, P., Ferbeyre, G., Breaker, R.R., 2011. Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. PLoS Comput Biol 7, e1002031.

Poiata, E., Meyer, M.M., Ames, T.D., Breaker, R.R., 2009. A variant riboswitch aptamer class for S-adenosylmethionine common in marine bacteria. RNA 15, 2046–56.

Ponty, Y., Termier, M., Denise, A., 2006. GenRGenS: software for generating random genomic sequences and structures. Bioinformatics 22, 1534–5.

Quarta, G., Kim, N., Izzo, J.A., Schlick, T., 2009. Analysis of riboswitch structure and function by an energy landscape framework. J Mol Biol 393, 993–1003.

Quarta, G., Sin, K., Schlick, T., 2012. Dynamic energy landscapes of riboswitches help interpret conformational rearrangements and function. PLoS Comput Biol 8, e1002368.

Ravnum, S., Andersson, D.I., 2001. An adenosyl-cobalamin (coenzyme-B12)-repressed translational enhancer in the *cob* mRNA of *Salmonella typhimurium*. Mol Microbiol 39, 1585–94.

Reeder, J., Steffen, P., Giegerich, R., 2007. pknotsrg: RNA pseudoknot folding including near-optimal structures and sliding windows. Nucleic Acids Res 35, W320–4.

Regulski, E.E., Moy, R.H., Weinberg, Z., Barrick, J.E., Yao, Z., Ruzzo, W.L., Breaker, R.R., 2008. A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. Mol Microbiol 68, 918–32.

Ren, A., Rajashankar, K.R., Patel, D.J., 2012. Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. Nature 486, 85–9.

Repoila, F., Darfeuille, F., 2009. Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. Biol. Cell 101, 117–131.

Ritz, J., Martin, J.S., Laederach, A., 2013. Evolutionary evidence for alternative structure in RNA sequence co-variation. PLoS Comput Biol 9, e1003152.

Rivas, E., Eddy, S.R., 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol 285, 2053–68.

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muniz-Rascado, L., Garcia-Sotelo, J.S., Weiss, V., Solano-Lira, H., Martinez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernandez, S., Alquicira-Hernandez, K., Lopez-Fuentes, A., Porron-Sotelo, L., Huerta, A.M., Bonavides-Martinez, C., Balderas-Martinez, Y.I., Pannier, L., Olvera, M., Labastida, A., Jimenez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chavez, V., Hernandez-Alvarez, A., Morett, E., Collado-Vides, J., 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res 41, D203–13.

Sato, K., Hamada, M., Asai, K., Mituyama, T., 2009. CENTROIDFOLD: a web server for RNA secondary structure prediction. Nucleic Acids Res 37, W277–80.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E., Ye, J., 2009. Database resources of the national center for biotechnology information. Nucleic Acids Res 37, 21.

Scarabino, D., Crisari, A., Lorenzini, S., Williams, K., Tocchini-Valentini, G.P., 1999. tRNA prefers to kiss. EMBO J 18, 4571–8.

Schlax, P.J., Xavier, K.A., Gluick, T.C., Draper, D.E., 2001. Translational repression of the *Escherichia coli* alpha operon mRNA: importance of an mRNA conformational switch and a ternary entrapment complex. J Biol Chem 276, 38494–501.

Schneider, T., Stephens, R., 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–100.

Scott, L.G., Hennig, M., 2008. RNA structure determination by NMR. Methods Mol Biol 452, 29–61.

Serganov, A., Huang, L., Patel, D.J., 2008. Structural insights into amino acid binding and gene control by a lysine riboswitch. Nature 455, 1263–7.

Serganov, A., Huang, L., Patel, D.J., 2009. Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. Nature 458, 233–7.

Serganov, A., Nudler, E., 2013. A decade of riboswitches. Cell 152, 17–24.

Serganov, A., Polonskaia, A., Phan, A.T., Breaker, R.R., Patel, D.J., 2006. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. Nature 441, 1167–71.

Serganov, A., Yuan, Y.R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A.T., Hobartner, C., Micura, R., Breaker, R.R., Patel, D.J., 2004. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. Chem Biol 11, 1729–41.

Shannon, C., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423.

Shaw, T.I., Manzour, A., Wang, Y., Malmberg, R.L., Cai, L., 2011. Analyzing modular RNA structure reveals low global structural entropy in microRNA sequences. J Bioinform Comput Biol 9, 283–98.

Simmonds, P., Karakasiliotis, I., Bailey, D., Chaudhry, Y., Evans, D.J., Goodfellow, I.G., 2008. Bioinformatic and functional analysis of RNA secondary structure elements among different genera of human and animal caliciviruses. Nucleic Acids Res 36, 2530–46.

Singh, P., Bandyopadhyay, P., Bhattacharya, S., Krishnamachari, A., Sengupta, S., 2009. Riboswitch detection using profile hidden markov models. BMC Bioinformatics 10, 325.

Smith, K.D., Lipchock, S.V., Ames, T.D., Wang, J., Breaker, R.R., Strobel, S.A., 2009. Structural basis of ligand binding by a c-di-gmp riboswitch. Nat Struct Mol Biol 16, 1218–23.

Taboada, B., Ciria, R., Martinez-Guerrero, C.E., Merino, E., 2012. ProOpDB: Prokaryotic operon database. Nucleic Acids Res 40, D627–31.

Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M., Mattick, J.S., 2010. Non-coding RNAs: regulators of diseas. Journal of Pathology 220, 126–13.

Thore, S., Leibundgut, M., Ban, N., 2006. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. Science 312, 1208–11.

Tinoco, I., J., Bustamante, C., 1999. How RNA folds. J Mol Biol 293, 271–81.

Tomsic, J., McDaniel, B.A., Grundy, F.J., Henkin, T.M., 2008. Natural variability in S-adenosylmethionine (SAM)-dependent riboswitches: S-box elements in *Bacillus subtilis* exhibit differential sensitivity to SAM in vivo and in vitro. J Bacteriol 190, 823–33.

Tran, T.T., Zhou, F., Marshburn, S., Stead, M., Kushner, S.R., Xu, Y., 2009. *De novo* computational prediction of non-coding RNA genes in prokaryotic genomes. Bioinformatics 25, 2897–905.

Vicens, Q., Mondragon, E., Batey, R.T., 2011. Molecular sensing by the aptamer domain of the fmn riboswitch: a general model for ligand binding by conformational selection. Nucleic Acids Res 39, 8586–98.

Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., Gelfand, M.S., 2003. Regulation of the vitamin b12 metabolism and transport in bacteria by a conserved RNA structural element. RNA 9, 1084–97.

Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., Gelfand, M.S., 2004. Riboswitches: the oldest mechanism for the regulation of gene expression? Trends Genet 20, 44–50.

Wang, Y., Manzour, A., Shareghi, P., Shaw, T.I., Li, Y.W., Malmberg, R.L., Cai, L., 2012. Stable stem enabled shannon entropies distinguish non-coding RNAs from random backgrounds. BMC Bioinformatics 13 Suppl 5, S1.

Watson, P.Y., Fedor, M.J., 2012. The *ydaO* motif is an ATP-sensing riboswitch in *Bacillus subtilis*. Nat Chem Biol 8, 963–5.

Weinberg, Z., Barrick, J.E., Yao, Z., Roth, A., Kim, J.N., Gore, J., Wang, J.X., Lee, E.R., Block, K.F., Sudarsan, N., Neph, S., Tompa, M., Ruzzo, W.L., Breaker, R.R., 2007. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Nucleic Acids Res 35, 4809–19.

Weinberg, Z., Regulski, E.E., Hammond, M.C., Barrick, J.E., Yao, Z., Ruzzo, W.L., Breaker, R.R., 2008. The aptamer core of SAM-IV riboswitches mimics the ligand-binding site of SAM-I riboswitches. RNA 14, 822–8.

Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H., Breaker, R.R., 2010. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol 11, R31.

Westhof, E., Masquida, B., Jossinet, F., 2011. Predicting and modeling RNA architecture. CSH Perspectives 3, a003632.

Wilson, R.C., Smith, A.M., Fuchs, R.T., Kleckner, I.R., Henkin, T.M., Foster, M.P., 2011. Tuning riboswitch regulation through conformational selection. J Mol Biol 405, 926–38.

Winkler, W., Nahvi, A., Breaker, R.R., 2002a. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature 419, 952–6.

Winkler, W., Nahvi, A., Breaker, R.R., 2002b. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature 419, 952–6.

Winkler, W.C., Cohen-Chalamish, S., Breaker, R.R., 2002c. An mRNA structure that controls gene expression by binding FMN. Proc Natl Acad Sci U S A 99, 15908–13.

Winkler, W.C., Nahvi, A., Roth, A., Collins, J.A., Breaker, R.R., 2004. Control of gene expression by a natural metabolite-responsive ribozyme. Nature 428, 281–6.

Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E., Breaker, R.R., 2003. An mRNA structure that controls gene expression by binding s-adenosylmethionine. Nat Struct Biol 10, 701–7.

Yockey, H.P., 2005. Information theory, evolution, and the origin of life. Cambridge University Press.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31, 3406–15.

Zuker, M., Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9, 133–48.

Appendix A

Structural Entropy Derivations

A.1 Structural Entropy of Structurally Ambiguous Non-stacking Grammars

In the case of structurally ambiguous grammars, the left-most derivation constraint makes it possible to uniquely enumerate over left-most derivation trees of structures by avoiding redundant counts of trees: Without loss of generality to all non-stacking grammar rules, we have:
Upon consecutive generation of c_i and c_j substructures in a left-most derivation tree, the following ordering of rules abides by the left-most derivation constraint:

$$\lambda \rightarrow c_i\ \omega; \omega \rightarrow \gamma\ c_j$$

While the following does not:

$$\omega \rightarrow \gamma\ c_j; \gamma \rightarrow c_i\ \delta \tag{A.1}$$

Where c_i and c_j are substructures: $c_i, c_j \in \{a, aYb\}$, and $\lambda, \omega, \gamma,$ and δ are all nonterminals where $\lambda = \omega = \gamma = \delta$. Y is any nonterminal.

By having a closer look at A.1, we can see that when $\omega \rightarrow \gamma\ c_j$ is applied anywhere on the sequence, it puts constraint on the inside probability function deriving γ, since it cannot contain any rule of type $\gamma \rightarrow c_i\ \delta$ at its outermost step. Note: $\gamma = \delta$. We will refer to this particular inside probability function as the left-most derivation inside probability function: $\alpha_l(\gamma, i, j, y)$.
On the other hand, when $\gamma \rightarrow c_i\ \delta$ is applied anywhere on the sequence, it puts constraint on the outside probability function deriving γ, since it cannot contain any rule of type $\omega \rightarrow \gamma\ c_j$ at its innermost step. Note: $\omega = \gamma$. We will refer to this particular outside probability function as the left-most derivation outside probability function: $\beta_l(\gamma, i, j, y)$.

Application of rule of type $\gamma \rightarrow c_i$ does not require any constraint on the last step of the outside probability function that produces γ. Also, substructure Y surrounded by a base-pair: aYb does not put any constraint on the last step of the inside probability function that produces Y.

Applying the above logic yields the structural entropy of ambiguous grammars to be equal to (2.7) with the following modifications for the expression containing the inside and outside probability function coeficients:

For the case of rule: $X \rightarrow aYbZ$:

If $X \neq Z$:

$$\beta(X,i,j,y)\,[\,f(X,aYbZ)\sum_{i+2<k<j-1}\alpha(Y,i+2,k-1,y)\alpha(Z,k+1,j-1,y)]$$

If $X = Z$:

$$\beta_l(X,i,j,y)\,[\,f(X,aYbZ)\sum_{i+2<k<j-1}\alpha(Y,i+2,k-1,y)\alpha(Z,k+1,j-1,y)]$$

For the case of rule: $X \rightarrow YaZb$:
If $X \neq Y$:

$$\beta(X,i,j,y)\,[\,f(X,YaZb)\sum_{i+1<k<j-2}\alpha(Y,i+1,k-1,y)\alpha(Z,k+1,j-2,y)]$$

If $X = Y$:

$$\beta(X,i,j,y)\,[\,f(X,YaZb)\sum_{i+1<k<j-2}\alpha_l(Y,i+1,k-1,y)\alpha(Z,k+1,j-2,y)]$$

For the case of rule: $X \rightarrow aY$:
if $X \neq Y$:

$$\beta(X,i,j,y)f(X,aY)\alpha(Y,i+2,j-1,y)$$

If $X = Y$:

$$\beta_l(X,i,j,y)f(X,aY)\alpha(Y,i+2,j-1,y)$$

For the case of rule: $X \rightarrow Ya$:
If $X \neq Y$:

$$\beta(X,i,j,y)f(X,Ya)\alpha(Y,i+1,j-2,y)$$

If $X = Y$:

$$\beta(X,i,j,y)f(X,Ya)\alpha_l(Y,i+1,j-2,y)$$

Where $\alpha(X,i,j,y)$, $\alpha_l(X,i,j,y)$, $\beta(X,i,j,y)$, and $\beta_l(X,i,j,y)$ are recursively defined as follows:

Note: The *If* statements in the following recursions are not mutually excusive from one another. (A grammar rule can apply to more than one term of the recursion). The following inside-outside probability functions are defined for the X symbol that refers to a particular non-terminal. Hence, symbol X on either side of the ($\rightarrow$) of grammar rules refers to the same non-terminal, here. This is while symbols Y and Z in the grammar rules symbolize any non-terminal.

$$\alpha(X,i,j,y) = \alpha(Y,i+1,j,y) \times p(X \rightarrow aY)$$
$$+\{ifX \neq Y : \alpha(Y,i,j-1,y) \times p(X \rightarrow Ya),\ else\ \alpha_l(Y,i,j-1,y) \times p(X \rightarrow Ya)\}$$
$$+\sum_{i+1<k<j} \alpha(Y,i+1,k-1,y) \times \alpha(Z,k+1,j,y) \times p(X \rightarrow aYbZ)$$
$$+\Bigg[ifX \neq Y : \sum_{i<k<j-1} \alpha(Y,i,k-1,y) \times \alpha(Z,k+1,j-1,y) \times p(X \rightarrow YaZb),$$
$$else \sum_{i<k<j-1} \alpha_l(Y,i,k-1,y) \times \alpha(Z,k+1,j-1,y) \times p(X \rightarrow YaZb)\Bigg]$$
$$+\alpha(Y,i+1,j-1,y) \times p(X \rightarrow aYb)$$

$$\beta(X,i,j,y) = \Bigg[ifY \neq X : \beta(Y,i-1,j,y) \times p(Y \rightarrow aX)$$
$$else\ \beta_l(Y,i-1,j,y) \times p(Y \rightarrow aX)\Bigg]$$
$$+\beta(Y,i,j+1,y) \times p(Y \rightarrow Xa)$$
$$+\Bigg[ifY \neq X : \sum_{0<k<i-1} \beta(Y,k-1,j,y) \times \alpha(Z,k+1,i-1,y) \times p(Y \rightarrow aZbX)$$
$$else \sum_{0<k<i-1} \beta_l(Y,k-1,j,y) \times \alpha(Z,k+1,i-1,y) \times p(Y \rightarrow aZbX)\Bigg]$$
$$+\Big[ifY \neq Z : \sum_{j<k<n_y+1} \beta(Y,i-1,k+1,y) \times \alpha(Z,j+1,k,y) \times p(Y \rightarrow aXbZ)$$
$$else \sum_{j<k<n_y+1} \beta_l(Y,i-1,k+1,y) \times \alpha(Z,j+1,k,y) \times p(Y \rightarrow aXbZ)\Bigg]$$
$$+\Bigg[ifY \neq Z : \sum_{0<k<i} \beta(Y,k-1,j+1,y) \times \alpha(Z,k,i-1,y) \times p(Y \rightarrow ZaXb),$$
$$else \sum_{0<k<i} \beta(Y,k-1,j+1,y) \times \alpha_l(Z,k,i-1,y) \times p(Y \rightarrow ZaXb)\Bigg]$$

$$+ \sum_{j+1<k<n_y+1} \beta(Y, i, k+1, y) \times \alpha(Z, j+1, k-1, y) \times p(Y \rightarrow XaZb)$$

$$+\beta(Y, i-1, j+1, y) \times p(Y \rightarrow aXb)$$

$$
\begin{gathered}
\alpha_l(X,i,j,y) = \{ifX \neq Y : \alpha(Y,i+1,j,y) \times p(X \rightarrow aY),\ else\ 0\} \\
+\{ifX \neq Y : \alpha(Y,i,j-1,y) \times p(X \rightarrow Ya),\ else\ \alpha_l(Y,i,j-1,y) \times p(X \rightarrow Ya)\} \\
+\left\{ifX \neq Z : \sum_{i+1<k<j} \alpha(Y,i+1,k-1,y) \times \alpha(Z,k+1,j,y) \times p(X \rightarrow aYbZ),\ else\ 0\right\} \\
+\left[ifX \neq Y : \sum_{i<k<j-1} \alpha(Y,i,k-1,y) \times \alpha(Z,k+1,j-1,y) \times p(X \rightarrow YaZb),\right. \\
\left. else \sum_{i<k<j-1} \alpha_l(Y,i,k-1,y) \times \alpha(Z,k+1,j-1,y) \times p(X \rightarrow YaZb)\right] \\
+\alpha(Y,i+1,j-1,y) \times p(X \rightarrow aYb)
\end{gathered}
$$

$$
\begin{gathered}
\beta_l(X,i,j,y) = \left[ifY \neq X : \beta(Y,i-1,j,y) \times p(Y \rightarrow aX)\right. \\
\left. else\ \beta_l(Y,i-1,j,y) \times p(Y \rightarrow aX)\right] \\
+\{ifY \neq X : \beta(Y,i,j+1,y) \times p(Y \rightarrow Xa),\ else\ 0\} \\
+\left[ifY \neq X : \sum_{0<k<i-1} \beta(Y,k-1,j,y) \times \alpha(Z,k+1,i-1,y) \times p(Y \rightarrow aZbX)\right. \\
\left. else \sum_{0<k<i-1} \beta_l(Y,k-1,j,y) \times \alpha(Z,k+1,i-1,y) \times p(Y \rightarrow aZbX)\right] \\
+\left[ifY \neq Z : \sum_{j<k<n_y+1} \beta(Y,i-1,k+1,y) \times \alpha(Z,j+1,k,y) \times p(Y \rightarrow aXbZ)\right. \\
\left. else \sum_{j<k<n_y+1} \beta_l(Y,i-1,k+1,y) \times \alpha(Z,j+1,k,y) \times p(Y \rightarrow aXbZ)\right] \\
+\left[ifY \neq Z : \sum_{0<k<i} \beta(Y,k-1,j+1,y) \times \alpha(Z,k,i-1,y) \times p(Y \rightarrow ZaXb),\right. \\
\left. else \sum_{0<k<i} \beta(Y,k-1,j+1,y) \times \alpha_l(Z,k,i-1,y) \times p(Y \rightarrow ZaXb)\right] \\
+\left\{ifY \neq X : \sum_{j+1<k<n_y+1} \beta(Y,i,k+1,y) \times \alpha(Z,j+1,k-1,y) \times p(Y \rightarrow XaZb),\ else\ 0\right\}
\end{gathered}
$$

$$+\beta(Y, i-1, j+1, y) \times p(Y \rightarrow aXb)$$

With initialization:

$$\alpha_l(X, i, i, y) = \alpha(X, i, i, y) = p(X \rightarrow a),\ \forall i$$

$$\beta_l(X, 0, n_y+1, y) = \beta(X, 0, n_y+1, y) = 1,\ \ if X = S_0,\ else\ 0$$

$$\beta_l(X, i, n_y+1, y) = \beta(X, i, n_y+1, y),\ \forall i$$

A.2 Data Collection

Sequences were downloaded from Rfam 10.0. The SEED sequences were downloaded for the following RNAs:

Table A.1: Downloaded sequences from Rfam 10.0

Accession	Name	Type	No. Seed	Average Length	Average % Identity
RF00002	5_8S_rRNA	rRNA	61	152	70
RF00001	5S_rRNA	rRNA	712	116	59
RF00013	6S	rRNA	154	180	45
RF00174	Cobalamin	riboswitch	431	203	54
RF00050	FMN	riboswitch	146	136	72
RF00234	glmS	riboswitch	17	182	59
RF00504	Glycine	riboswitch	44	100	55
RF00163	Hammerhead_1	riboswitch	30	59	70
RF00168	Lysine	riboswitch	47	183	51
RF01056	Mg_sensor	riboswitch	4	114	78
RF00802	mir-207	miRNA	7	78	85
RF00838	mir-252	miRNA	6	79	71
RF00717	mir-315	miRNA	16	82	75
RF00770	mir-330	miRNA	6	97	91
RF00842	MIR403	miRNA	11	106	68
RF00711	mir-449	miRNA	57	91	58
RF00776	mir-540	miRNA	3	82	78
RF00849	mir-60	miRNA	4	72	89
RF00957	mir-663	miRNA	14	89	77
RF00844	mir-67	miRNA	16	65	79
RF00917	mir-708	miRNA	24	83	78
RF00830	mir-74	miRNA	6	95	67
RF01055	MOCO_RNA_motif	riboswitch	179	141	59
RF00522	PreQ1	riboswitch	42	45	67
RF01054	preQ1-II	riboswitch	14	104	68
RF00167	Purine	riboswitch	133	100	56
RF01480	rli52	snRNA/riboswitch	6	95	95
RF01481	rli53	snRNA/riboswitch	5	173	97
RF01491	rli54	snRNA/riboswitch	5	283	99
RF01482	rli55	snRNA/riboswitch	3	100	99
RF01483	rli56	snRNA/riboswitch	6	181	96
RF01485	rli61	snRNA/riboswitch	4	106	99
RF01486	rli62	snRNA/riboswitch	2	172	98
RF01057	SAH_riboswitch	riboswitch	52	84	63
RF00521	SAM_alpha	riboswitch	40	79	71
RF00634	SAM-IV	riboswitch	40	115	73
RF00065	snoR9	snoRNA	5	127	85
RF00072	SNORA75	snoRNA	6	135	74
RF00067	SNORD15	snoRNA	18	124	59
RF00068	SNORD21	snoRNA	5	92	73
RF00069	SNORD24	snoRNA	14	77	71
RF00070	SNORD29	snoRNA	10	73	76
RF00071	SNORD73	snoRNA	25	70	77
RF00059	TPP	riboswitch	115	111	56
RF00005	tRNA	tRNA	967	73	45
RF00030	MRP	RNase	67	321	46
RF00009	nuclear	RNase P	117	312	48
RF00011	bact. type B	RNase P	114	367	68
RF00010	bact. type A	RNase P	306	380	62

A.3 Sensitivity and Specificity (PPV) of models to Annotated Secondary Structure of ncRNAs

Table A.2: Structural Entropy p-Value vs. Accuracy for Bralibase Predictions. Percent(%) of Sensitivity (Sen.) and Specificity (Sp.) of SCFG Models to Bralibase Annotated Secondary Structures. Number of sequences and structures: One annotated secondary structure for each alignment. All alignments have five sequences in them. g2intron (460 sequences; 92 structures), rRNA (445 sequences; 89 structures), tRNA (490 sequences; 98 structures), and U5 (540 sequences; 108 structures).

Mixed80-trained	BJK Sen.	BJK Sp.	RUN Sen.	RUN Sp.	IVO Sen.	IVO Sp.
g2intron	64	55	23	16	3	5
rRNA	42	43	16	13	2	2
tRNA	77	78	30	25	4	7
U5	64	59	11	8	1	2
Benchmark-trained	BJK Sen.	BJK Sp.	RUN Sen.	RUN Sp.	IVO Sen.	IVO Sp.
g2intron	66	57	29	19	3	3
rRNA	42	41	19	15	2	3
tRNA	76	76	33	27	6	7
U5	64	57	18	12	1	1
Rfam5-trained	BJK Sen.	BJK Sp.	RUN Sen.	RUN Sp.	IVO Sen.	IVO Sp.
g2intron	52	44	6	5	2	4
rRNA	38	36	12	10	2	2
tRNA	64	64	22	21	4	7
U5	71	60	6	4	1	2

Table A.3: Structural Entropy p-Value vs. Accuracy for Rfam Predictions. Percent(%) of Sensitivity (Sen.) and Specificity (Sp.) of SCFG Models to Rfam Annotated Secondary Structures.

Mixed80-trained	BJK Sen.	BJK Sp.	RUN Sen.	RUN Sp.	IVO Sen.	IVO Sp.
miRNA	74	66	57	46	2	4
riboswitch	51	39	13	8	2	2
RNase MRP	46	28	19	10	2	1
RNase P	50	41	9	6	2	2
rRNA	42	40	23	17	2	2
tRNA	73	74	30	25	5	8
snoRNA	45	14	60	12	7	3
Benchmark-trained	BJK Sen.	BJK Sp.	RUN Sen.	RUN Sp.	IVO Sen.	IVO Sp.
miRNA	77	66	64	50	4	4
riboswitch	49	38	14	9	2	2
RNase MRP	43	26	21	11	2	1
RNase P	49	40	10	7	2	2
rRNA	42	39	27	20	3	3
tRNA	71	71	34	27	7	7
snoRNA	43	13	50	10	10	3
Rfam5-trained	BJK Sen.	BJK Sp.	RUN Sen.	RUN Sp.	IVO Sen.	IVO Sp.
miRNA	77	63	49	44	2	4
riboswitch	43	32	9	6	2	2
RNase MRP	37	21	15	9	1	1
RNase P	40	31	6	5	2	2
rRNA	39	36	14	12	2	2
tRNA	58	60	23	21	5	8
snoRNA	39	12	43	10	8	3

A.4 Generating random structures

In order to generate random sequences with structure, we used the GenRGenS Software package (Ponty et al., 2006). For each grammar, we generated a pool of random sequences of length 93. The procedure for generating random sequences with structure are as follows: SCFGs were converted to their equivalent Weighted Context-Free Grammar (WCFG) format for this purpose. We then filtered random sequences to find the desired nucleotide composition sets. Standard deviation for each nucleotide in each cluster is 0.02. Standard deviation of nucleotide composition for ncRNA sequences in each cluster is roughly 0.05. Converting SCFGs to their equivalent WCFG together with filtering of sequences causes the posterior probabilities of grammar rules to be different their prior probabilities defined by their corresponding SCFG. In order to reduce the effects of such difference of probabilities and also have a more general understanding of the behavior of each CFG design, we combined all random sequences generated by different parameter sets of a given model. Having one single pool of random sequences for each grammar and for each cluster, we calculated structural entropy of these sequences using the different parameter sets, of their corresponding model, separately.

A.5 Short ncRNA Sequence Clustering

`Cluster 1 (high GC content):`

The nucleotide composition of sequences in this cluster are $0.192, 0.294, 0.328, 0.186$ for A, C, G, and U, respectively. Standard deviation for each nucleotide is roughly 0.05. The length of ncRNAs is between 88 and 98 nucleotides. Sequences in this cluster: miRNAs (43 sequences), tRNAs (4 sequences), riboswitch (24 sequences: 9 Glycine, 1 Hammerhead, 2 preQ1-II, 1 Purine, 11 TPP), snoRNA (4 sequences) and rRNAs (2 5SrRNA sequences).

`Cluster 2 (low GC content):`

The nucleotide composition of sequences in this cluster are $0.331, 0.169, 0.186, 0.315$ for A, C, G, and U, respectively. Standard deviation for each nucleotide is roughly 0.05. The length of ncRNAs is between 88 and 98 nucleotides. Sequences in this cluster: miRNAs (4 sequences), riboswitch (41 sequences: 4 preQ1-II, 21 Purine, 6 rli52, and 10 TPP) and snoRNA (5 sequences).

`Cluster 3 (medium GC content):`

The nucleotide composition of sequences in this cluster are $0.246, 0.231, 0.278, 0.245$ for A, C, G, and U, respectively. Standard deviation for each nucleotide is roughly 0.05. The length of ncRNAs is between 88 and 98 nucleotides. Sequences in this cluster: miRNAs (16 sequences), tRNA (6 sequences), riboswitch (19 sequences: 9 Glycine, 8 SAH, and 2 TPP) and rRNA (1 5SrRNA sequence).

A.6 Structural Entropy P-values of Short Non-coding RNAs Against Random Sequences with Structure

Table A.4: Structural Entropy p-Values of miRNA and Riboswitches with Different GC-comp. Under All Models. Structural entropy p-values of riboswitch and miRNA sequences against random sequences with structure under various folding models (See A.5 for details about clusters and sequences). Corresponding values represent percentage of sequences having p-values higher than 0.95. Labels H, L, and A represent high, low and average GC−compositions, respectively.

Grammar:	RUN	RUN	RUN	RUN	RUN	RUN	RUN	RUN	RUN
Training:	Mixed80	Mixed80	Mixed80	Benchmark	Benchmark	Benchmark	Rfam5	Rfam5	Rfam5
GC−comp.	H	L	A	H	L	A	H	L	A
miRNA	16	0	0	12	0	0	12	0	0
Riboswitch	58	51	11	58	51	11	63	78	21
Grammar:	IVO	IVO	IVO	IVO	IVO	IVO	IVO	IVO	IVO
Training:	Mixed80	Mixed80	Mixed80	Benchmark	Benchmark	Benchmark	Rfam5	Rfam5	Rfam5
GC−comp.	H	L	A	H	L	A	H	L	A
miRNA	33	0	75	49	0	63	47	0	63
Riboswitch	42	80	42	54	88	47	58	93	53
Grammar:	BJK	BJK	BJK	BJK	BJK	BJK	BJK	BJK	BJK
Training:	Mixed80	Mixed80	Mixed80	Benchmark	Benchmark	Benchmark	Rfam5	Rfam5	Rfam5
GC−comp.	H	L	A	H	L	A	H	L	A
miRNA	0	0	0	0	0	0	0	0	0
Riboswitch	50	10	0	50	15	0	71	29	0
Grammar:	RND1	RND1	RND1	RND10	RND10	RND10			
Training:	None	None	None	None	None	None			
GC−comp.	H	L	A	H	L	A			
miRNA	9	0	63	0	0	25			
Riboswitch	21	76	26	17	51	5			

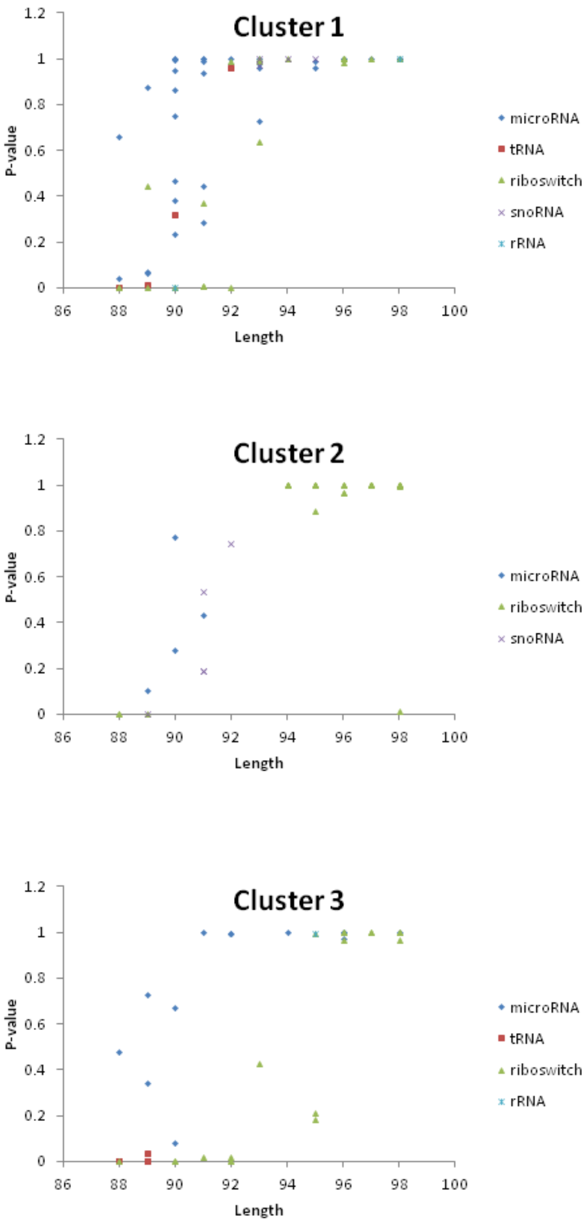

Figure A.1: Structural Entropy p-Values Under Inaccurate Modeling. Structural entropy p-values of short ncRNA sequences against random sequences with structure under the IVO model (See A.5 for information about clusters and sequences). Benchmark training set was used. Other training sets yield similar results (See A.4).

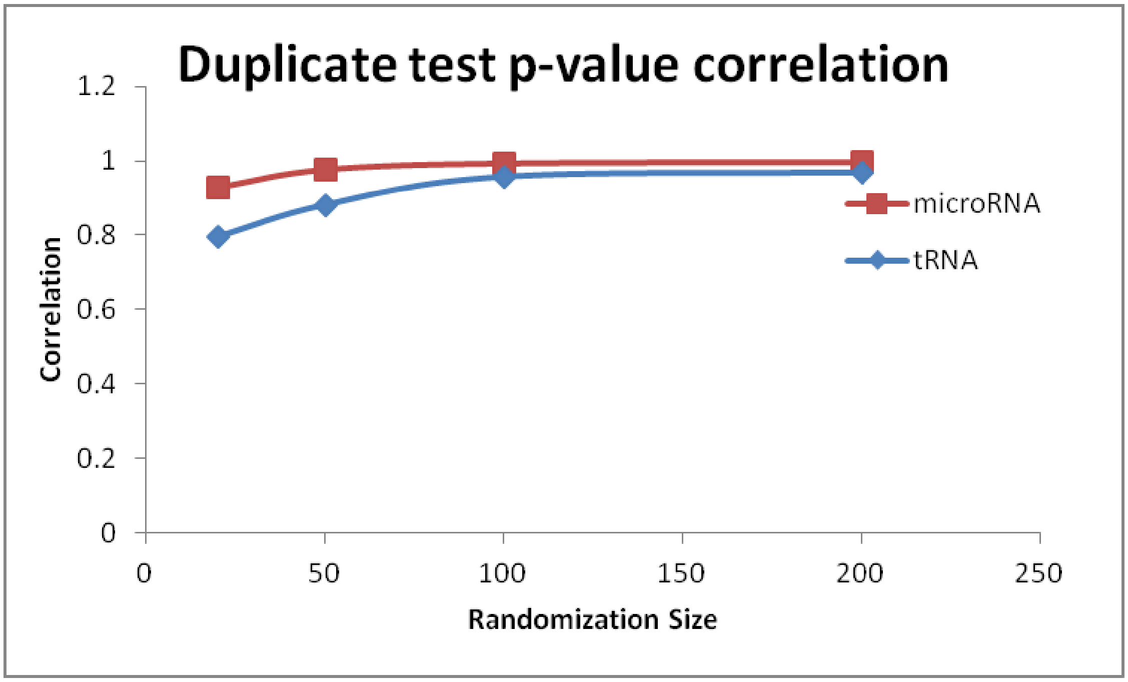

Figure A.2: P-Value Stability Test. Dinucleotide shuffling tests were performed for miRNAs (170 sequences) and tRNAs with lengths between 75 and 100 nucleotides (145 sequences). BJK grammar model was used with trained parameters based on the mixed80 data set. Various numbers of random shuffles of 20, 50, 100, and 200 were used for p-value calculation of individual sequences. The test was then repeated for a second time. The plot is the correlation between corresponding p-values obtained in the first and the second test. Micro RNA average p-value is 0.070 ± 0.001 in all cases. tRNA average p-value is 0.155 ± 0.005 in all cases.

A.7 P-value Stability Test for Dinucleotide Shuffling

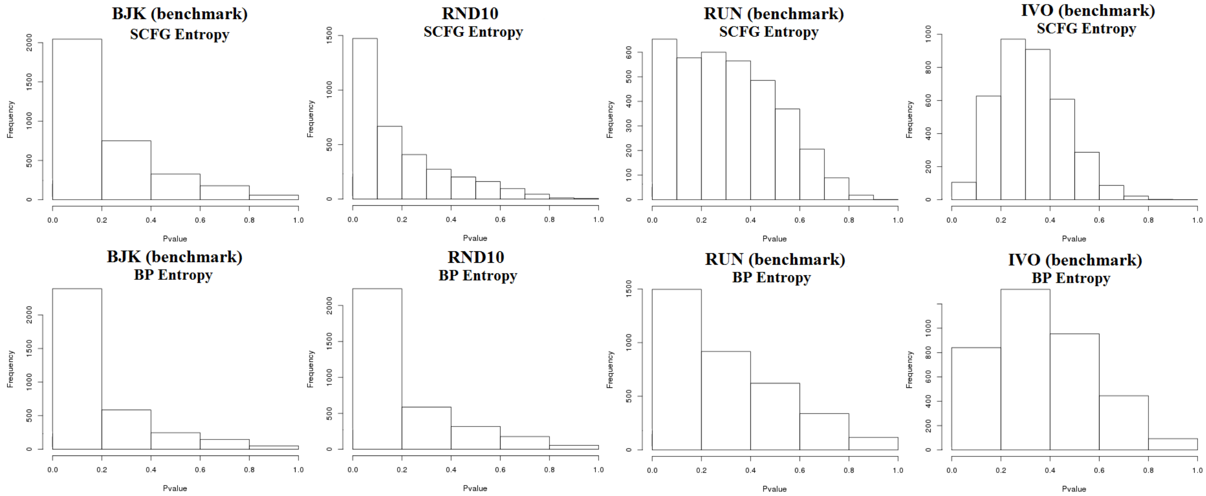

Figure A.3: Structural and Base-pairing Entropy p-Values under Single-nucleotide Randomization. Structural entropy (top row) and base-pairing entropy (bottom row) empirical p-values of all ncRNAs against random sequences: Total of 447 long sequences (24 bacterial type B RNase P, 117 nuclear RNase P, and 306 bacterial type A RNase P) were excluded due to high computational complexity. GenRGenS (Ponty et al., 2006) was used to generate random sequence ensembles for each individual sequence separately. Random sequences were of the same length and single nucleotide distribution as the original sequence. Size of the random ensemble is proportional to the original sequence length.

A.8 Structural Entropy Empirical P-values for single nucleotide composition and dinucleotide shuffling tests

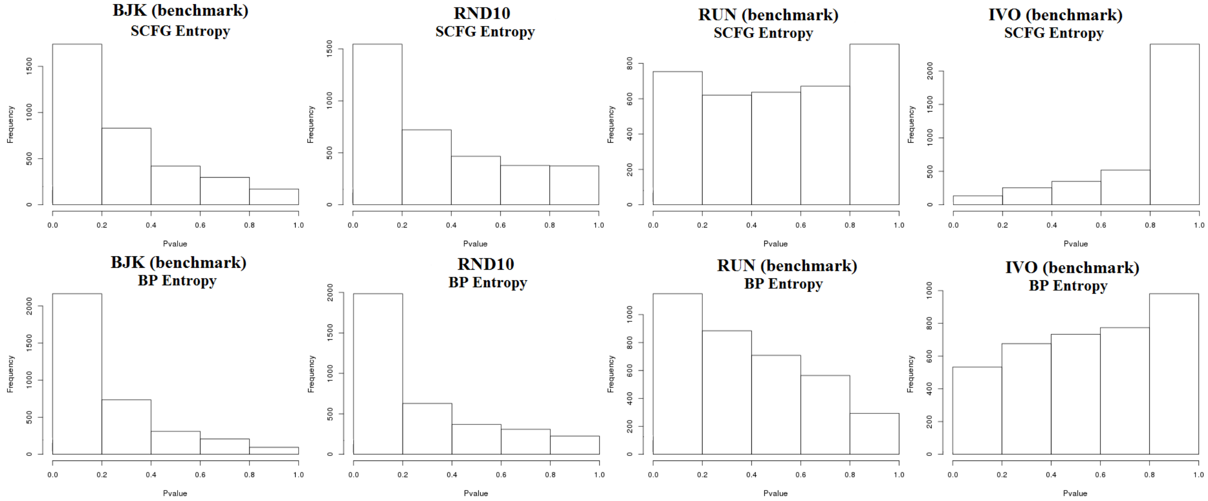

Figure A.4: Structural and Base-pairing Entropy p-Values under Di-nucleotide Randomization. Structural entropy (top row) and base-pairing entropy (bottom row) empirical p-values of all ncRNAs against random shuffles: Total of 447 long sequences (24 bacterial type B RNase P, 117 nuclear RNase P, and 306 bacterial type A RNase P) were excluded due to high computational complexity. (Altschul and Erickson, 1985) was used to generate shuffled random sequence ensembles for each individual sequence, separately. Random sequences were of the same length and dinucleotide distribution as the original sequence. Size of the random ensemble is proportional to the original sequence length.

A.9 Structural Entropy and Total Base-pairing Entropy

A.9.1 Relationship between structural and total-pairing entropy

In the following, we show that in the case of structurally unambiguous grammars, total pairing entropy is the upper bound of structural entropy. We know that for structurally unambiguous grammar spaces there is a one-to-one relationship between all possible structures π on sequence y and pairing/non-pairing representations of all nucleotide so long as pairing/non-pairing representations are valid pseudoknot-free secondary structure representations. If we assign the same probability to valid pairing/non-pairing representations as their corresponding π and zero for non-valid ones, then the entropy of both sets of probabilities will be equal:

$$H(\Pi|y) = H(I^M|y)$$

Where $I^M = [I_{1,2} \dots I_{i,j} \dots I_{n_y - 1, n_y}]$ is a random variable vector of the size M whose every element is binary. Value M is the total number of unique pairs of nucleotides definable on sequence y, which is choose 2 from n_y. $|I^M| = 2^M$.

The independence bound on entropy states that the total uncertainty about multiple random events is always less than or equal to the sum of their individual uncertainties (Cover and Thomas, Theorem 2.6.6,

Table A.5: Structural Entropy p-Values Average Under All Models. Average structural entropy p-values for ncRNA families against dinucleotide shuffles (Di.) and single nucleotide composition random sequences (Sing.) under BJK, RUN, and IVO model

Mixed80-trained	BJK (Di.)	RUN (Di.)	IVO (Di.)	BJK (Sing.)	RUN (Sing.)	IVO (Sing.)
miRNA	0.068	0.115	0.758	0.045	0.086	0.287
riboswitch	0.288	0.611	0.729	0.212	0.399	0.362
RNase	0.171	0.553	0.648	0.163	0.389	0.36
RNaseP	0.281	0.847	0.581	0.26	0.502	0.381
rRNAs	0.281	0.428	0.693	0.242	0.285	0.326
tRNA	0.174	0.558	0.891	0.122	0.353	0.403
snoRNA	0.525	0.671	0.783	0.449	0.357	0.285
Benchmark-trained	BJK (Di.)	RUN (Di.)	IVO (Di.)	BJK (Sing.)	RUN (Sing.)	IVO (Sing.)
miRNA	0.065	0.097	0.814	0.042	0.057	0.255
riboswitch	0.311	0.596	0.773	0.226	0.368	0.332
RNase	0.178	0.522	0.692	0.167	0.353	0.335
RNaseP	0.299	0.783	0.544	0.274	0.436	0.348
rRNAs	0.286	0.411	0.736	0.241	0.251	0.292
tRNA	0.187	0.525	0.914	0.127	0.293	0.352
snoRNA	0.528	0.616	0.839	0.439	0.297	0.259
Rfam5-trained	BJK (Di.)	RUN (Di.)	IVO (Di.)	BJK (Sing.)	RUN (Sing.)	IVO (Sing.)
miRNA	0.059	0.221	0.919	0.037	0.069	0.143
riboswitch	0.361	0.671	0.887	0.255	0.334	0.22
RNase	0.171	0.556	0.821	0.15	0.275	0.248
RNaseP	0.335	0.866	0.661	0.27	0.519	0.256
rRNAs	0.324	0.481	0.891	0.259	0.199	0.182
tRNA	0.216	0.588	0.977	0.13	0.209	0.19
snoRNA	0.496	0.753	0.968	0.354	0.305	0.16

Table A.6: Structural Entropy p-Values Average Under a Random Model. Average structural entropy p-values for ncRNA families against dinucleotide shuffles (Di.) and single nucleotide composition random sequences (Sing.) under RND10 Model

ncRNA Family	RND10 (Di.)	RND10 (Sing.)
miRNA	0.076	0.029
riboswitch	0.473	0.288
RNase	0.115	0.076
RNaseP	0.512	0.404
rRNAs	0.257	0.141
tRNA	0.209	0.091
snoRNA	0.476	0.24

pg. 30).

$$H(X_1 \dots X_N) \leq \sum_{i=1}^{N} H(X_i)$$

Substituting for pairing uncertainties gives

$$H(\Pi|y) \leq \texttt{TP Entropy}(y)$$

A.9.2 Random fold and random sequence fold

Random fold refers to uniformly distributed probability assignments to all possible folds given both a sequence and a folding model that can satisfy such a folding distribution (if they exist!). Let's call the corresponding variables y^*, G^*, Θ^*, and $\Pi(y^*)$. Maximum Entropy Theorem (Cover and Thomas, pg. 409) states that the Shannon folding entropy of such a distribution is higher than any other folding distribution so

Table A.7: Kolmogorov-Smirnov Distance of p-Values Under All Models. Kolmogorov-Smirnov test statistic distance of dinucleotide-shuffled base-pairing (BP) and structural (SCFG) entropy p-values of RNA families were calculated. Below is the sum of all pairwise distances for each folding model. RNA families used: miRNA, riboswitch, RNase MRP, bacterial type B RNase P (64 sequences), rRNA, snoRNA, and tRNA.

Parameter Set	BJK(BP)	RUN(BP)	IVO(BP)	BJK(SCFG)	RUN(SCFG)	IVO(SCFG)
Mixed80-trained	8.485828	9.990327	5.605391	8.370136	8.783326	5.580663
Benchmark-trained	8.621376	9.881117	5.631987	8.544915	8.294112	6.488109
Rfam5-trained	8.834326	9.218774	5.817303	8.711496	8.473245	8.293196

Table A.8: Kolmogorov-Smirnov Distance of p-Values Under a Random Model. Kolmogorov-Smirnov test statistic distance of dinucleotide-shuffled base-pairing (BP) and structural (SCFG) entropy p-values of RNA families were calculated. Below is the sum of all pairwise distances for each folding model. RNA families used: miRNA, riboswitch, RNase MRP, bacterial type B RNase P (64 sequences), rRNA, snoRNA, and tRNA. (Red.) refers to values obtained using the not-left-most-derivation-restricted inside and outside probability functions. (Left.) refers to values corresponding to left-most restricted inside and outside probability functions (see A.1).

	RND1(BP)	RND10(BP)	RND1(SCFG)	RND10(SCFG)		RND1(SCFG)	RND10(SCFG)
Red.	5.085642	9.931111	5.637075	10.161575	Left.	5.40582	9.988141

long as the sizes of the possible folds are kept equal.

$$H(\Pi|y^*, G^*, \Theta^*) \geq H(\Pi|y, G, \Theta) \quad \forall y, G, \Theta, \;\; where \; |\Pi(y)| = |\Pi(y^*)|$$

Since $\Pi(y^*)$ is a uniformly distributed discrete random variable,

$$p(\pi|y^*) = \frac{1}{|\Pi(y^*)|}, \quad \forall \pi \in \Pi(y^*)$$

The maximum entropy theorem also implies that the random sequence variable constructed of *iid*[1] nucleotides with uniform alphabet distribution will have higher entropy than any other sequences of the same lengh and alphabet. Let's call this random sequence space Y^{**}.

$$H(Y^{**}) \geq H(Y) \quad \forall Y, \;\; where |Y| = |Y^{**}|$$

Since Y^{**} is a uniformly distributed discrete variable,

$$p(y^{**}) = \frac{1}{|Y^{**}|}, \quad \forall y^{**} \in Y^{**}$$

However, the notion of random sequence fold, used in this paper, refers to the folding distribution of the random sequence. The maximum entropy theorem does not guarantee maximum folding entropy for the random sequence under an arbitrary model $H(\Pi|Y^{**}, G, \theta)$ or even under the model (G^*, Θ^*), $H(\Pi|Y^{**}, G^*, \theta^*)$ nor does it guarantee maximum folding entropy for a *typical*[2] instance of the random sequence y^{**} under

[1] iid is the shorthand for independent and identically distributed random variables

[2] The typical set is a set of sequences whose probability is close to one (Cover and Thomas, pg. 62).

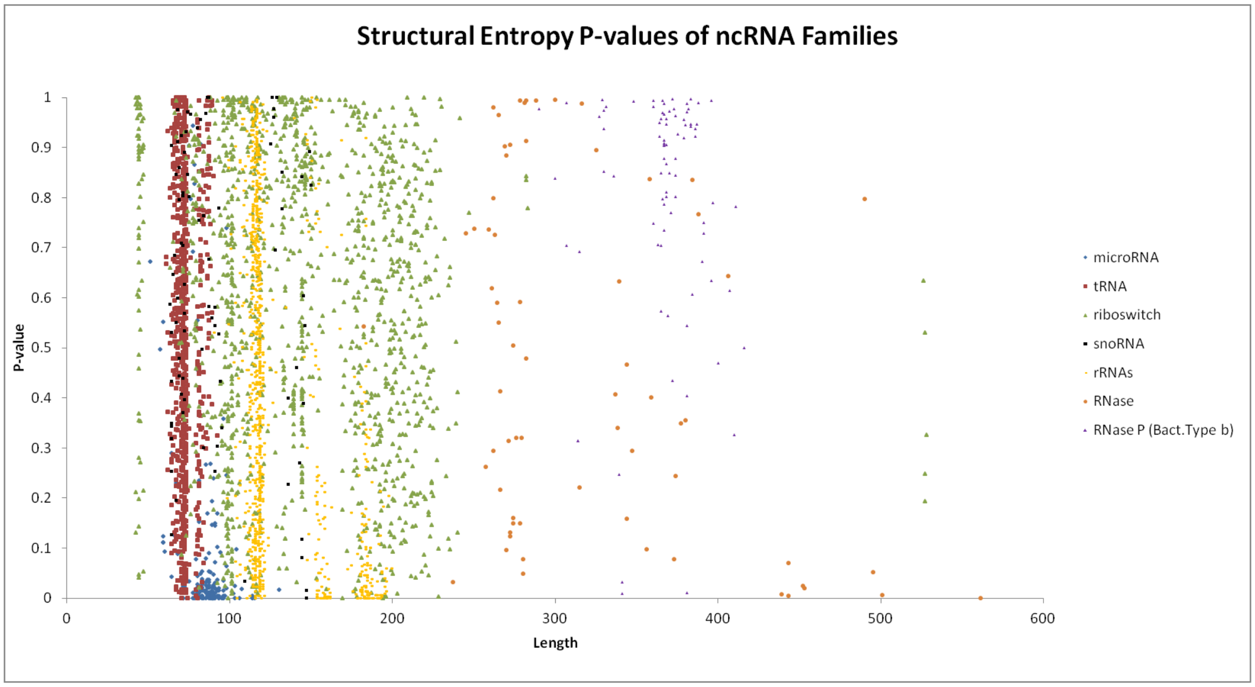

Figure A.5: Structural Entropy p-Value vs. Length. Average structural entropy p-values of ncRNA sequences are plotted with respect to their length. RUN (benchmark) model was used. P-values were obtained by performing dinucleotide-preserved shuffling test.

the folding model (G^*, Θ^*), $H(\Pi|y^{**}, G^*, \Theta^*)$. The next section contains an example which shows that less conserved primary structure does not necessarily lead to less conserved secondary structure.

Entropy of a mapped random variable

The probability assignment that maximizes the entropy of a random variable does not necessarily maximize the entropy of all probabilistic functions defined over it:

$$\begin{aligned} \exists f(.), X_1, X_2: \quad & H(f(X_1)) < H(f(X_2)) \\ & where \;\; H(X_1) > H(X_2) \end{aligned} \tag{A.2}$$

Consider the following scenario:
Let X_1 and X_2 be two distinct probability assignments for the binary random variable $\{0, 1\}$.

$$p_{X_1}(0) = 1/2 \quad p_{X_1}(1) = 1/2$$

$$p_{X_2}(0) = 3/5 \quad p_{X_2}(1) = 2/5$$

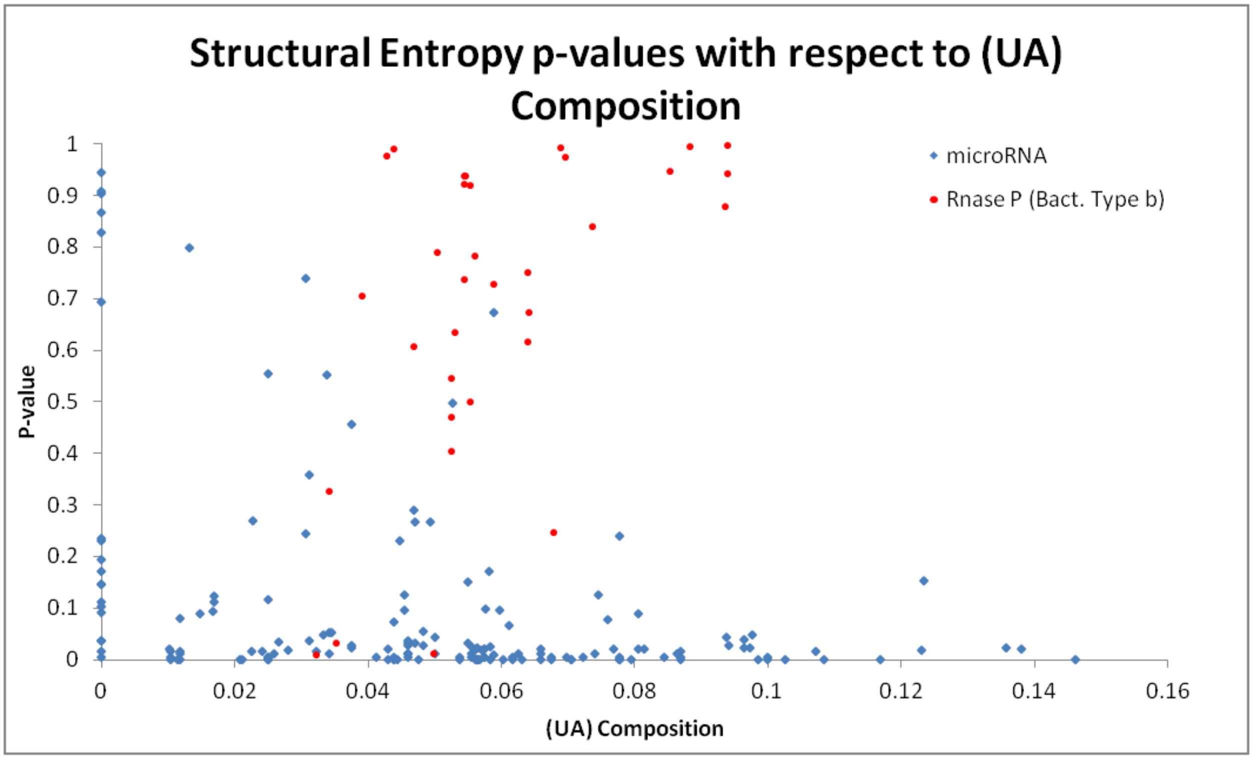

Figure A.6: Structural Entropy p-Value vs. UA-comp. Structural entropy p-values of miRNA and 34 bacterial type B RNase P sequences are plotted with respect to their `UA`-dinucleotide composition. RUN (benchmark) grammar was used as folding space model. P-values empirically calculated from dinucleotide shuffling test.

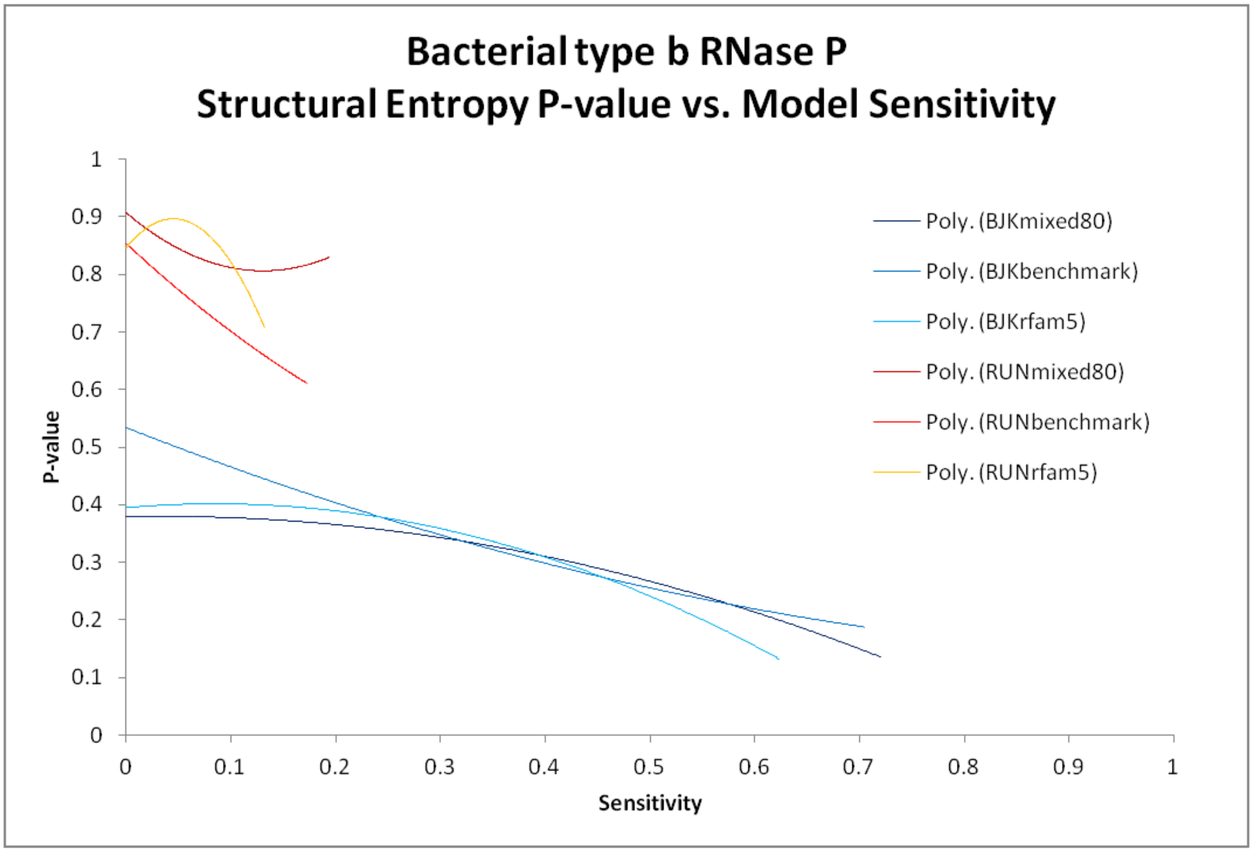

Figure A.7: Structural Entropy p-Value of RNase P vs. Model Sensitivity. Structural entropy p-values of Bacterial type B RNase P sequences against folding model sensitivity to their secondary structure. Dinucleotide shuffling was used to calculate p-values. 2-order polynomial trendline of p-values are shown for each grammar model.

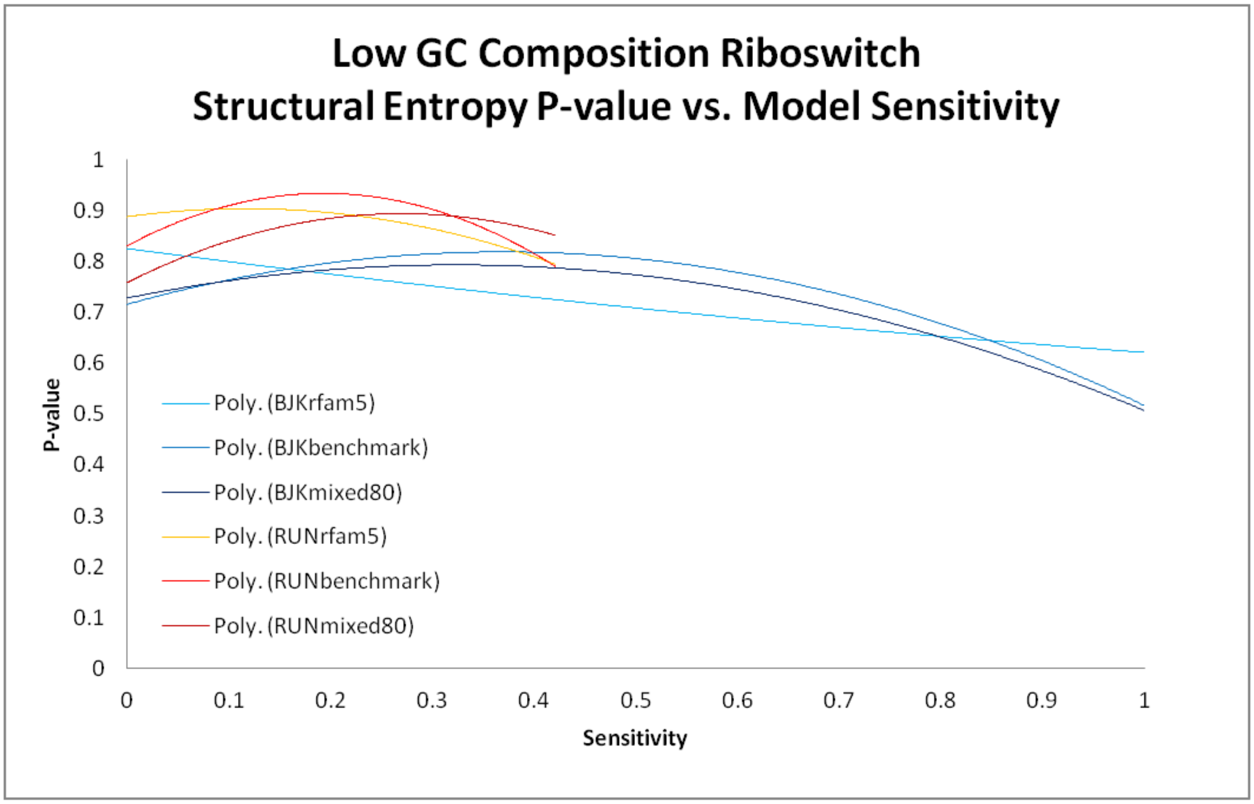

Figure A.8: Structural Entropy p-Value of Low-GC Riboswitches vs. Model Sensitivity. Structural entropy p-values of low-GC composition riboswitch sequences of length 93 ± 5 against folding model sensitivity to their secondary structure. Riboswitch sequences belong to cluster 2 (See A.5 for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See A.4 for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

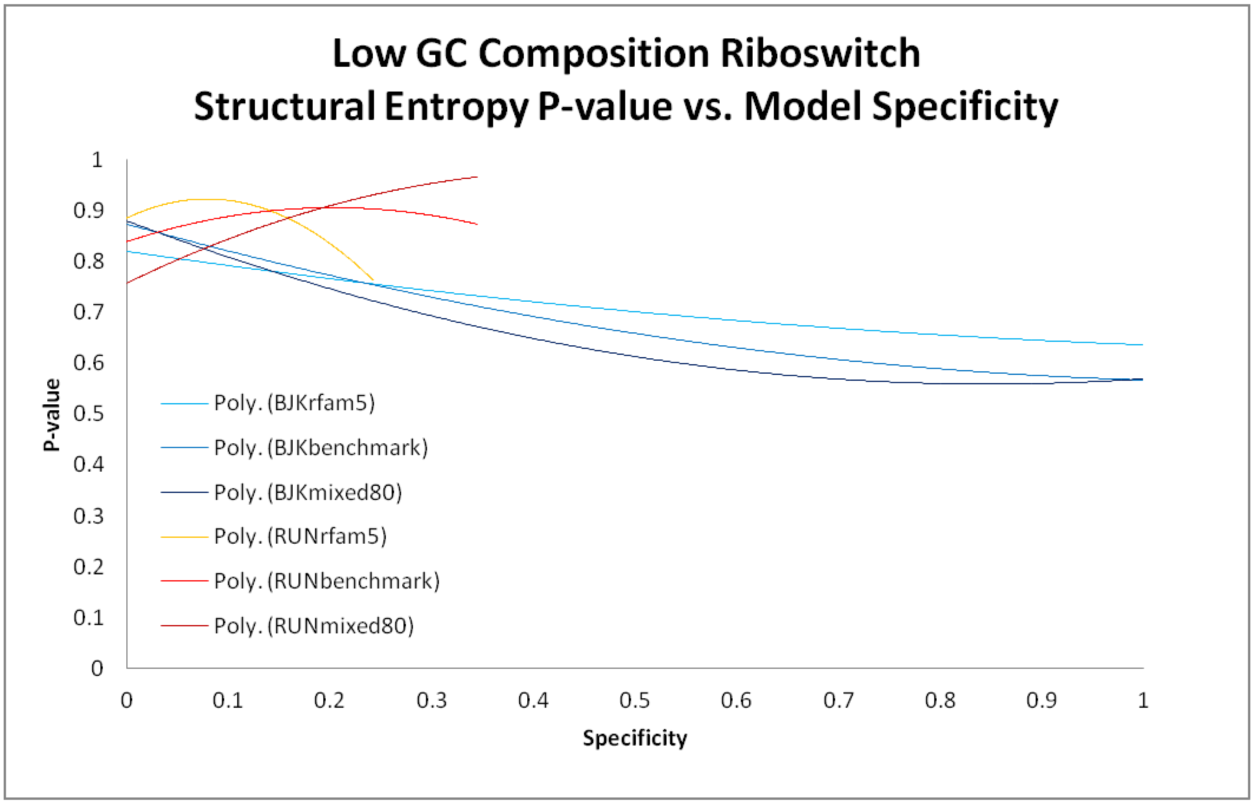

Figure A.9: Structural Entropy p-Value of Low-GC Riboswitches vs. Model Specificity. Structural entropy p-values of low-GC composition riboswitch sequences of length 93 ± 5 against folding model specificity to their secondary structure. Riboswitch sequences belong to cluster 2 (See A.5 for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See A.4 for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

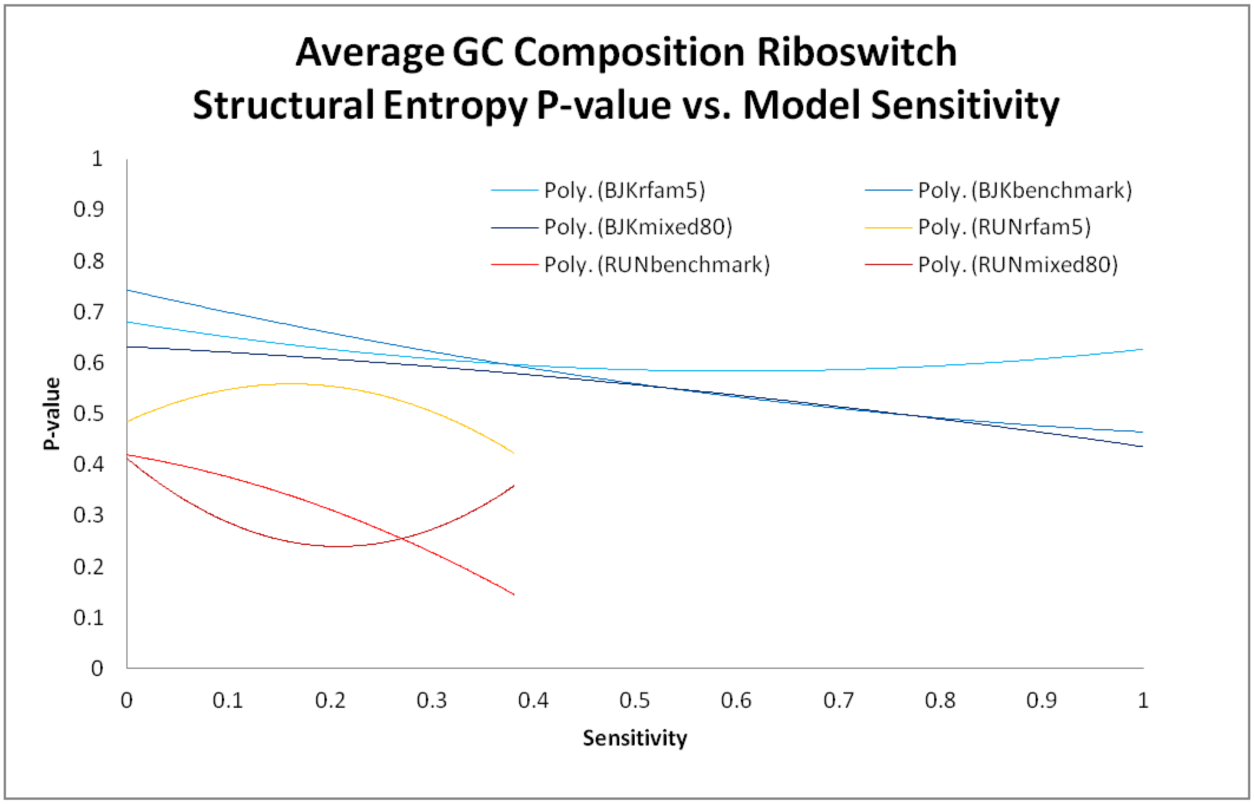

Figure A.10: Structural Entropy p-Value of Ave.-GC Riboswitches vs. Model Sensitivity. Structural entropy p-values of average-GC composition riboswitch sequences of length 93 ± 5 against folding model sensitivity to their secondary structure. Riboswitch sequences belong to cluster 3 (See A.5 for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See A.4 for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

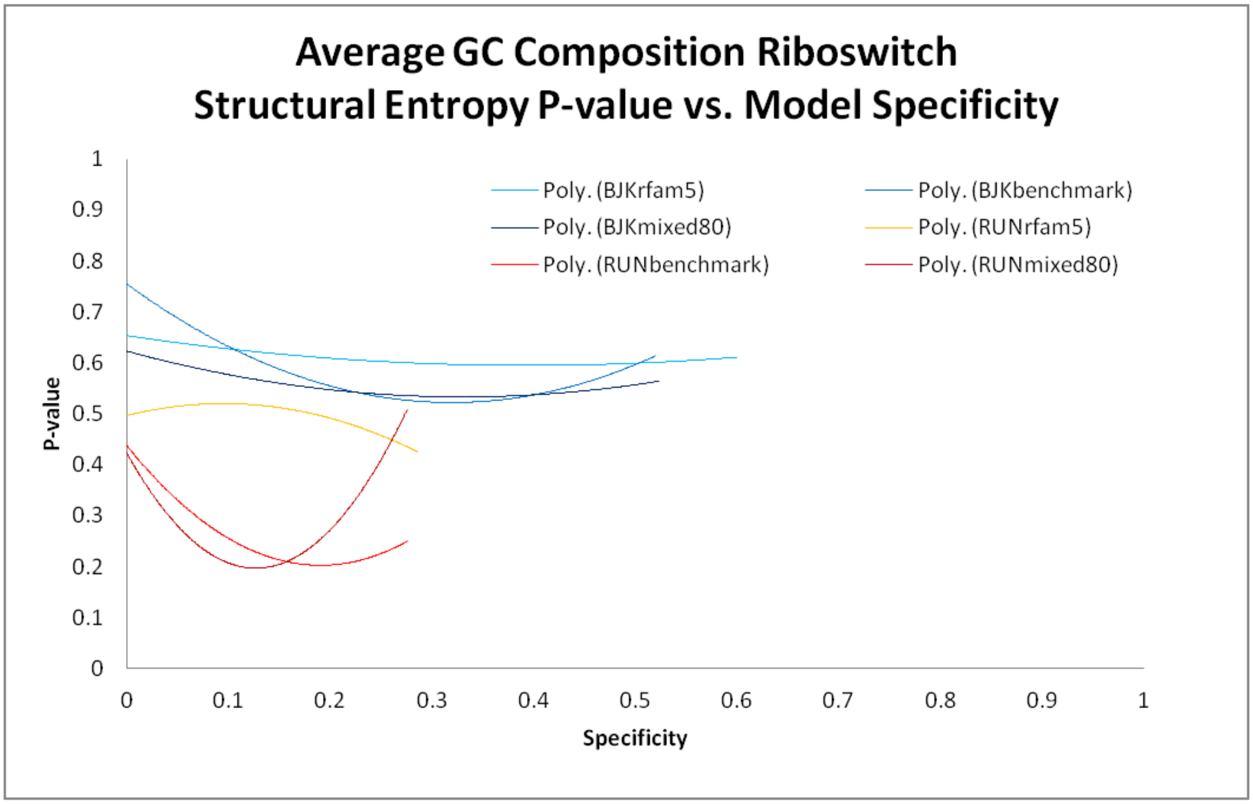

Figure A.11: Structural Entropy p-Value of Ave.-GC Riboswitches vs. Model Specificity. Structural entropy p-values of average-GC composition riboswitch sequences of length 93 ± 5 against folding model specificity to their secondary structure. Riboswitch sequences belong to cluster 3 (See A.5 for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See A.4 for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

Table A.9: Structural Entropy p-Value Correlation. Correlation of structural entropy p-values and dinucleotide composition for miRNA and 34 bacterial type B RNAse P sequences. RUN (benchmark) was used as folding model. P-values were obtained from dinucleotide shuffling test

	AA	CA	GA	UA
miRNA	-0.28	-0.35	0.12	-0.36
RNase P (bact. Type B)	0.39	0	-0.14	0.53
	AC	CC	GC	UC
miRNA	-0.22	0.21	0.14	0.33
RNase P (bact. Type B)	0.4	-0.49	-0.41	-0.45
	AG	CG	GG	UG
miRNA	-0.01	0.22	0.44	-0.23
RNase P (bact. Type B)	0.22	-0.4	-0.33	-0.31
	AU	CU	GU	UU
miRNA	-0.33	0.29	-0.18	-0.07
RNase P (bact. Type B)	0.21	-0.24	0.3	0.41

And let the following probabilistic binary function $f(.) \in \{0, 1\}$ definable over the binary random variable have the following assignment:

$$p(f(.)) : \left\{ \begin{array}{ll} p(f(0) = 0) = 2/3, & p(f(0) = 1) = 1/3 \\ p(f(1) = 0) = 1/4, & p(f(1) = 1) = 3/4 \end{array} \right\}$$

Entropy values for X_1 and X_2 are

$$\begin{array}{l} H(X_1) = -(1/2)\log(1/2) - (1/2)\log(1/2) = 1 \\ H(X_2) = -(3/5)\log(3/5) - (2/5)\log(2/5) = 0.971 \end{array}$$

While Entropy values for $f(X_1)$ and $f(X_2)$ random variables are,

$$\begin{array}{r} H(f(X_1)) = -((1/2)(2/3) + (1/2)(1/4))\log\left((1/2)(2/3) + (1/2)(1/4)\right) \\ -((1/2)(1/3) + (1/2)(3/4))\log\left((1/2)(1/3) + (1/2)(3/4)\right) = 0.995 \end{array}$$

$$\begin{array}{r} H(f(X_2)) = -((3/5)(2/3) + (2/5)(1/4))\log\left((3/5)(2/3) + (2/5)(1/4)\right) \\ -((3/5)(1/3) + (2/5)(3/4))\log\left((3/5)(1/3) + (2/5)(3/4)\right) = 1 \end{array}$$

Hence, function $f(.)$, X_1, and X_2 satisfy A.2.
In fact, the above argument can be generalized to the following binary streams:
Let X_1^L and X_2^L be binary stream random variables of length L consisting of iid variables X_1 and X_2, respectively. Also, let $F(.)$ be a binary stream random variable consisting of iid transition probabilities $f(.)$ such that:

$$p(F(x^L)) = \prod_{i=1}^{L} p(f(x_i))$$

The corresponding entropy values for the above random variables are:

$$\begin{array}{l} H(X_1^L) = L \\ H(X_2^L) = 0.971L \\ H(F(X_1^L)) = 0.995L \\ H(F(X_2^L)) = L \end{array}$$

The above scenario of binary streams (x_1^L and x_2^L) and their mapping functions $F(x^L)$ can be easily extended to RNA sequences and folding spaces, without loss of generality (i.e., $F(x^L)$ is not a valid folding model and used here only as an example). Hence, we have found at least one model under which a typical sequence x_2^L is slightly more conserved than x_1^L while its folding space is more diverse.

A.9.3 An Example of a High Structural Entropy P-value of a Hypothetical Micro RNA Sequence Against the Di-nucleotide Shuffling Test under a Single Stem-Loop SCFG Model

Consider the following 31-nucleotide long hypothetical micro RNA sequence with a single hairpin loop:

```
GGGGGGGGGGGCGCGCGCGCGCCCCCCCCCC
((((((((((...........))))))))))
```

The following SCFG has arbitrary assigned rule probabilities and attempts to capture the structural features of the hypothetical miRNA enforcing a single stem-loop structure:

$$S \rightarrow aXb\;(1) \qquad X \rightarrow aXb\;\;(0.5)|aL\;\;(0.5) \quad L \rightarrow aL\;(0.9)|a\;(0.1)$$

Where non-terminal S is the starting non-terminal, non-terminal X is the stem generation nonterminal, and non-terminal L denotes the generation of the loop.

We also assign 0.5 base-pairing probabilities to `G-C` and `C-G` and zero to other pairings. Also, loop generation probabilities are equally divided amongst all four nucleotides.

The Structural Entropy of the hypothetical miRNA is 0.631783 while all 100 di-nucleotide shuffled sequences have lower folding entropy. The following are various scenarios of di-nucleotide shuffled sequences and their corresponding CYK-based predicted structure and folding entropy:

```
GGGGGGGGGCCGCGCGGCCCCGCCCCCCGGC
(.............................); 0
GGGCCCCCCGGGCCGGGGGGGCGCCCGCGCC
((...........................)); 0.453339
GGGGGCGGCGGCCGCGCCCCCCCGGGGGCCC
(((.........................))); 0.58333
GCCCCCCCCCGCCGCGGGGGGCGGGGCGGGC
((((.......................)))); 0.619343
GCCCCCCGCGGCGCCCCGGGCCGGGGGGGGC
(((((((.................))))))); 0.631615
```

Appendix B

Riboswitch Classification

B.1 Data and Classification Results

Table B.1: Genomic locations of collected sequences. Column `ID` corresponds riboswitches in Table 3.1.

ID	Accession	start	end	strand	Length
ID01	U00096.3	3442440	3442547	-	108
ID02	NC_000964.3	486099	486230	+	132
ID03	AE017180.2	2773395	2773492	+	98
ID04	CP000860.1	1860063	1860186	-	124
ID05	U00096.3	4163564	4163632	+	69
ID06	BA000040.2	5279368	5279482	-	115
ID07	AE006468.1	2113803	2113897	-	95
ID08	CP000075.1	1675079	1675157	-	79
ID09	CP000702.1	1794825	1794895	+	71
ID10	AE017194.1	4815592	4815665	+	74
ID11	AE009951.2	2496	2668	+	173
ID12	U00096.3	3184455	3184718	-	264
ID13	NC_000964.3	2431380	2431615	-	236
ID14	AE009951.2	963901	963988	-	89
ID15	NC_000964.3	2549381	2549501	-	121
ID16	AE000512.1	1519015	1519250	-	236
ID17	NC_000964.3	2910878	2911045	-	170
ID18	CP001363.1	4712312	4712483	+	172
ID19	U00096.3	4467416	4467525	+	110
ID20	NC_000964.3	1395622	1395825	+	204
ID21	U00096.3	816923	817041	+	119
ID22	U00096.3	3238486	3238569	+	84
ID23	CP003959.1	4635235	4635309	+	75
ID24	AE007317.1	904178	904257	+	80
ID25	NC_000964.3	1439279	1439338	+	60
ID26	AE016796.2	504379	504491	+	113
ID27	NC_000964.3	626329	626426	-	98
ID28	NC_000964.3	2320055	2320196	-	142
ID29	U55047.1	3107	3215	+	109
ID30	U00096.3	3867416	3867488	-	73
ID31	BA000012.4	1943727	1943820	-	94
ID32	AY316747.1	197909	198004	+	96
ID33	AP012279.1	5017601	5017677	-	135
ID34	AL646052.1	1348529	1348625	+	97
ID35	AE008691.1	1750249	1750372	-	124
ID36	NC_000964.3	1180646	1180802	-	157
ID37	AE007869.2	2703460	2703559	+	100
ID38	CP000725.1	1038292	1038371	+	80
ID39	CP003726.1	618415	618496	+	82
ID40	NC_003888.3	2308634	2308770	-	137
ID41	AE000516.2	3723565	3723713	+	149
ID42	AAYC01000001.1	142052	142099	+	48
ID43	ABID01000011.1	17036	17084	-	49
ID44	CP000084.1	1005827	1005879	+	53
ID45	CP000084.1	1127359	1127423	-	65
ID46	FP929059.1	95139	95281	-	144
ID47	NC_009706.1	3903929	3904072	+	144
ID48	U00096.3	2185279	2185426	-	148
ID49	NC_000964.3	1242265	1242422	+	158
ID50	U00096.3	1322975	1323055	-	81
ID51	NC_000964.3	2377419	2377559	-	141
ID52	CP000148.1	1157816	1157926	-	111

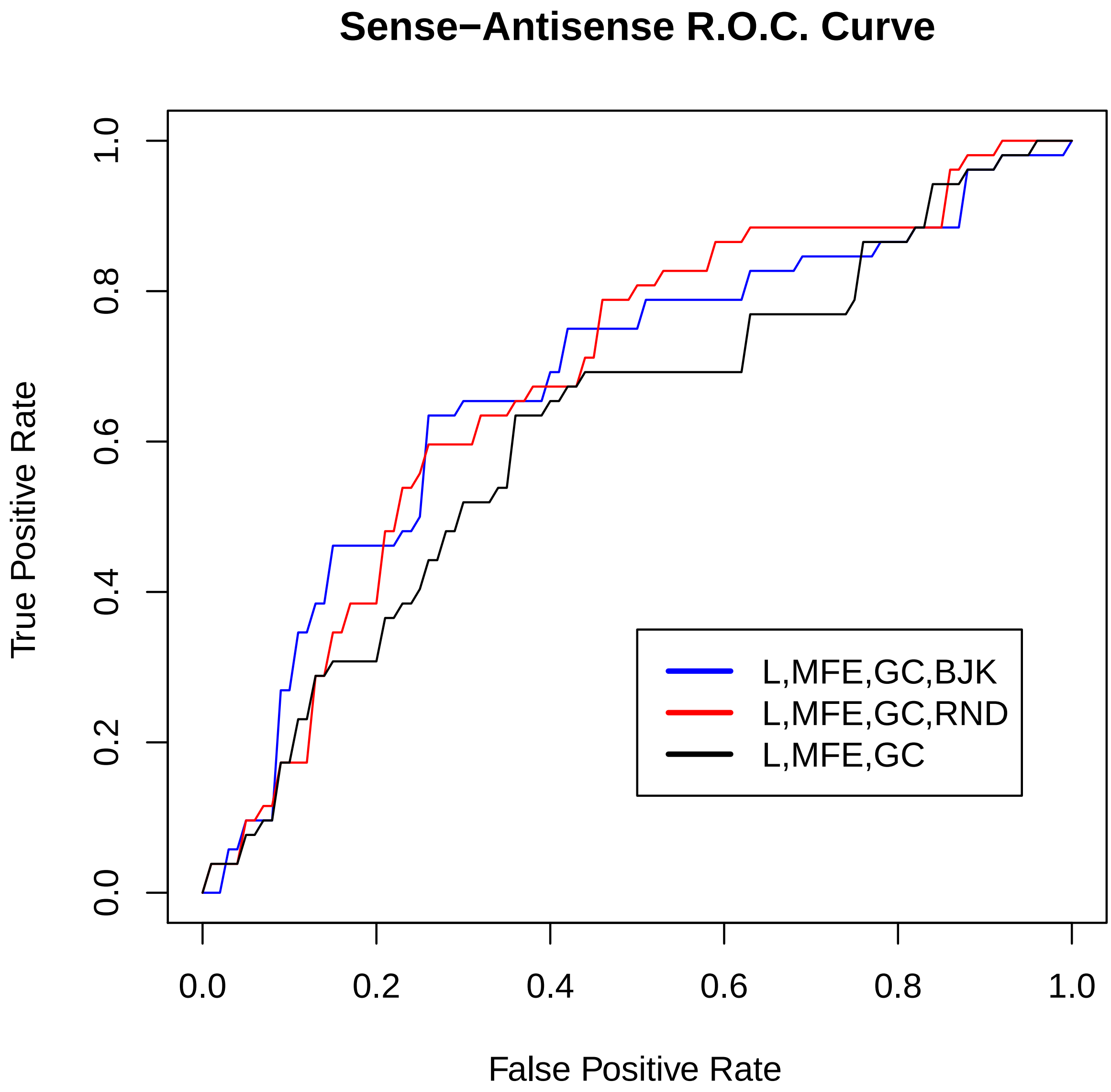

Figure B.1: Classification ROC Curve. ROC curves of 104-fold binomial simple logistic classifiers on all the 52 riboswitch sequences and their antisense sequences. classifier features shown in legend. Weka©open source software package used to assess probability distributions.

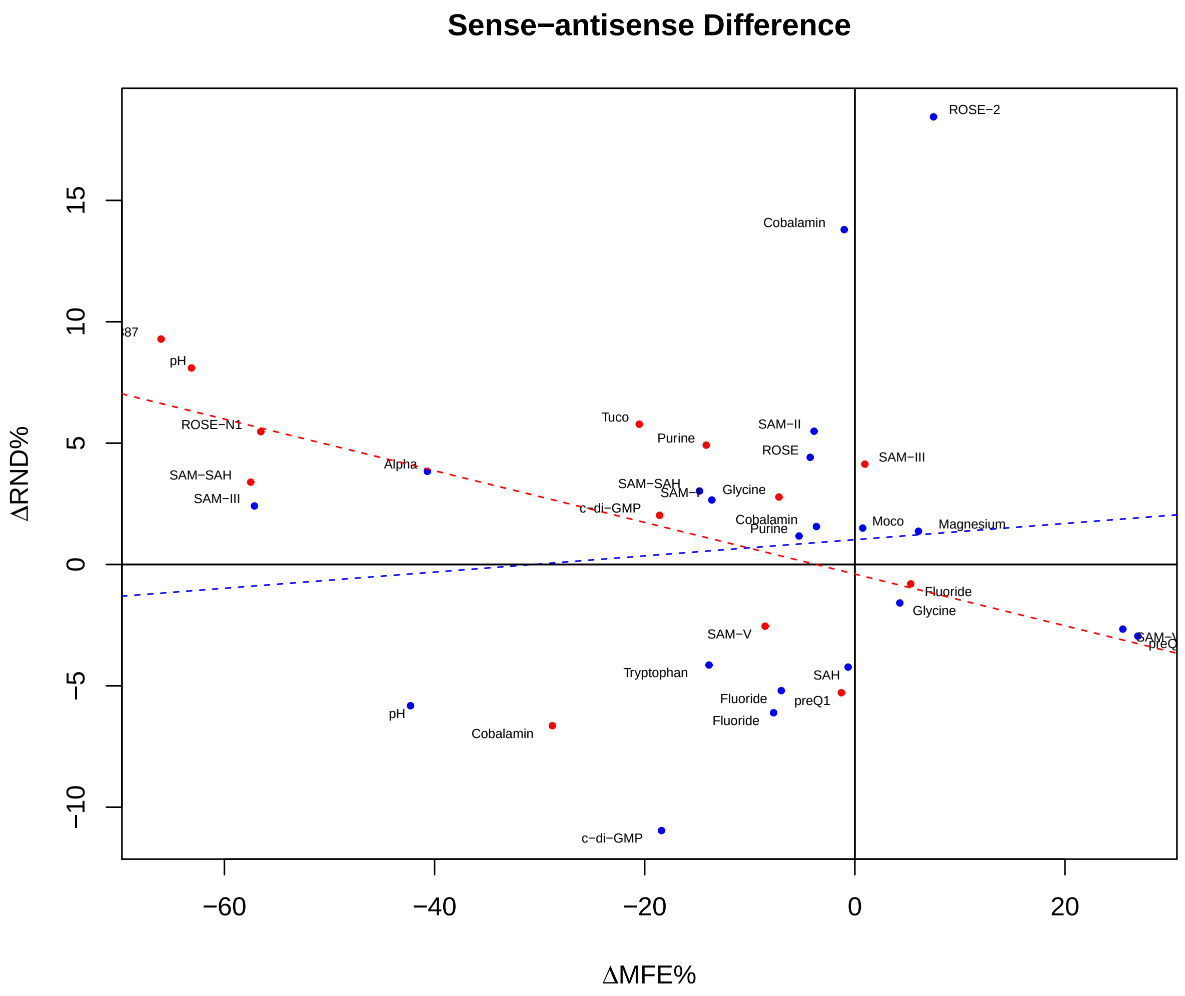

Figure B.2: Sense-antisense Differential Entropy of Shorter Riboswitches. Sense-Antisense differential entropy ($\Delta\, Entropy\% = 100 \times (Entropy_{sense} - Entropy_{antisense})/Entropy_{antisense}$) of sequences in the training and test sets with lengths less than 125nt have been shown against the minimum free energy difference between the sense and the antisense ($\Delta\, MFE\% = 100 \times (MFE_{sense} - MFE_{antisense})/abs(MFE_{antisense})$) under the RND model. Blue represents the training set while red represents the test set. Trendlines are shown as dashed lines. GC-composition average and standard deviations for the training and test sets are 0.51 ± 0.10 and 0.49 ± 0.09, respectively.

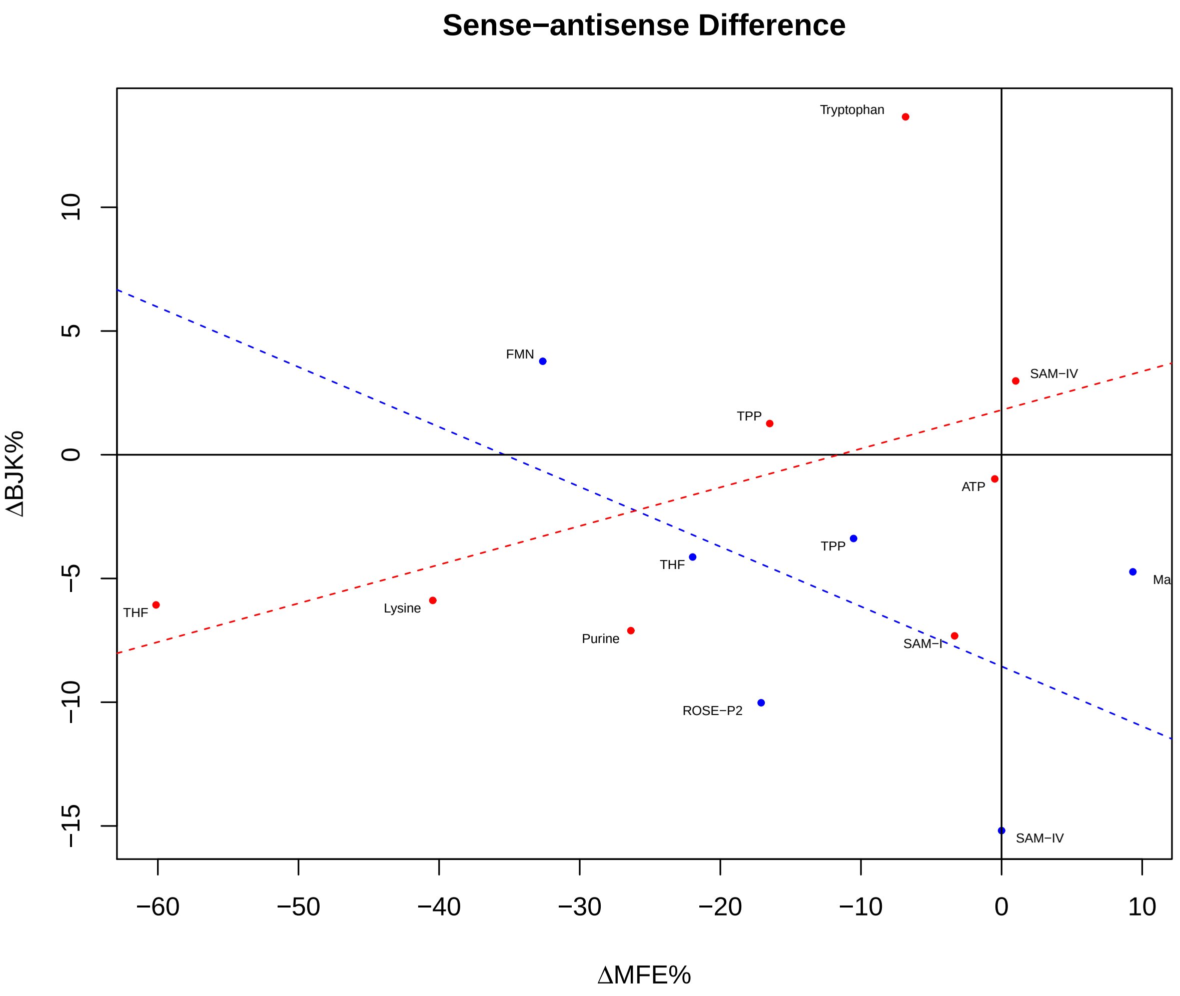

Figure B.3: Sense-antisense Differential Entropy of Longer Riboswitches. Sense-Antisense differential entropy ($\Delta\, Entropy\% = 100 \times (Entropy_{sense} - Entropy_{antisense})/Entropy_{antisense}$) of sequences in the training and test sets with lengths between 125nt and 175nt have been shown against the minimum free energy difference between the sense and the antisense ($\Delta\, MFE\% = 100 \times (MFE_{sense} - MFE_{antisense})/abs(MFE_{antisense})$) under the BJK model. Blue represents the training set while red represents the test set. Trendlines are shown as dashed lines. GC-composition average and standard deviations for the training and test sets are 0.55 ± 0.11 and 0.49 ± 0.10, respectively.

Table B.2: Classification Performance Using Cross Validation. 104-fold binomial logistic classifiers on all of the 52 riboswitch sequences and their antisense sequences. classifier features shown in legend. Weka©open source software package used. Features *L,MFE,GC,GU,GCU* and *U* denote length, MFE, GC-composition, and Uracil frequency, respectively. Features *RND* and *BJK* denote structural entropy of the RND and BJK models, respectively. as defined in (Huynen et al., 1997). Feature *BJKbp* denotes base-pairing entropy as defined in (Huynen et al., 1997). Feature *Sil* denotes the two-cluster average Silhouette index of energy landscape as calculated in (Quarta et al., 2012).

Classifier	TP Rate	FP Rate	MCC	R.O.C. Area
{*L,GC,GU,Sil*}	0.750	0.250	0.500	0.826
{*L,GC,GU*}	0.673	0.327	0.346	0.700
{*L,GC,GU,BJK*}	0.644	0.356	0.289	0.691
{*L,GC,GU,BJKbp*}	0.654	0.346	0.309	0.690
{*L,GC,GU,RND*}	0.654	0.346	0.308	0.689
{*L,MFE,GC,GU,RND*}	0.673	0.327	0.346	0.714
{*L,MFE,GC,GU*}	0.654	0.346	0.308	0.707
{*L,MFE,GC,GU,BJK*}	0.663	0.337	0.327	0.703
{*L,MFE,GC,GU,BJKbp*}	0.663	0.337	0.327	0.701
{*L,MFE,GC,GU,Sil*}	0.625	0.375	0.250	0.697
{*L,MFE,GU*}	0.663	0.337	0.327	0.710
{*L,MFE,GU,RND*}	0.663	0.337	0.327	0.702
{*L,MFE,GU,BJKbp*}	0.663	0.337	0.32	0.701
{*L,MFE,GU,Sil*}	0.654	0.346	0.308	0.701
{*L,MFE,GU,BJK*}	0.644	0.356	0.289	0.699
{*L,MFE,GC,RND*}	0.663	0.337	0.327	0.708
{*L,MFE,GC,BJK*}	0.663	0.337	0.327	0.703
{*L,MFE,GC,BJKbp*}	0.635	0.365	0.269	0.683
{*L,MFE,GC*}	0.606	0.394	0.212	0.650
{*L,MFE,GC,Sil*}	0.635	0.365	0.270	0.644
{*L,MFE,GCU,RND*}	0.644	0.356	0.289	0.693
{*L,MFE,GCU,BJK*}	0.625	0.375	0.250	0.617
{*L,MFE,GCU,BJKbp*}	0.596	0.404	0.193	0.595
{*L,MFE,GCU*}	0.587	0.413	0.174	0.581
{*L,MFE,GCU,Sil*}	0.548	0.452	0.097	0.554

B.2 *Bacillus subtilis* Classification Results

B.3 *Bacillus subtilis* Genome-wide Scan Results

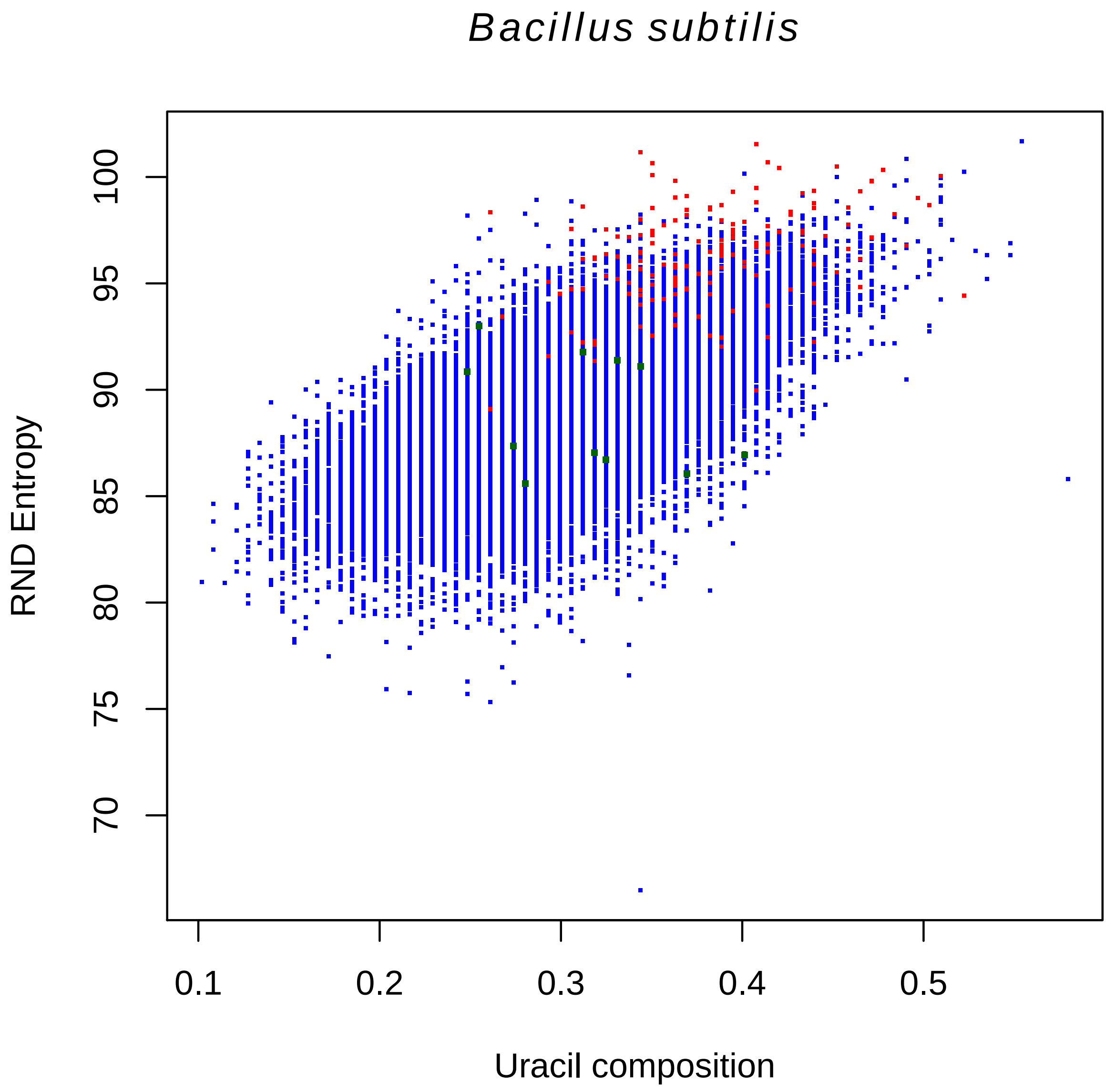

Figure B.4: Structural Entropy vs. Uracil-comp. in *B. subtilis*. Entropy Distribution of the 157 nt window-scan results. 28340 candidate segments of *B. subtilis* against Uracil composition. Blue denotes all segments. Red denotes those with classification probabilities under the LMFEGCRND are higher than 0.8. Green denotes the eleven bonafide riboswitches of the test set that are in *B. subtilis*.

Table B.3: Short UTR Collection. 30 randomly chosen untranslated regions of lengths less than 80 nt corresponding to the σ-70 transcription factor binding sites in *Escherichia coli* str. K-12 substr. MG1655 (GenBank©ID: U00096.2). Column `Start` denotes start of the binding site. `End` denotes the downstream start codon. `Gene` denotes the name of the first gene in the corresponding mRNA. `Length` denotes the length of the UTR.

Start	End	Strand	Gene	Length
42325	42403	+	*fixA*	79
246641	246712	+	*yafL*	72
570070	570116	+	*ybcL*	47
848134	848173	-	*dps*	40
879876	879950	+	*dacC*	75
989579	989637	-	*pncB*	59
1108480	1108558	+	*mdoG*	79
1331812	1331879	+	*cysB*	68
1397550	1397576	-	*fnr*	27
1570069	1570096	-	*gadB*	28
1732381	1732459	+	*mepH*	79
1927731	1927756	-	*yebE*	26
2039370	2039399	+	*zinT*	30
2268700	2268748	+	*rtn*	49
2380676	2380735	+	*elaD*	60
2541550	2541579	-	*cysP*	30
2823813	2823854	+	*srlA*	42
2982146	2982216	-	*kduI*	71
3134393	3134425	-	*pitB*	33
3276888	3276936	+	*kbaZ*	49
3467875	3467918	-	*chiA*	44
3651959	3651984	+	*slp*	26
3735493	3735520	+	*malS*	28
3845190	3845221	-	*uhpT*	32
3909548	3909591	-	*pstS*	44
4028994	4029036	-	*fadB*	43
4213425	4213501	+	*aceB*	77
4244442	4244487	-	*malE*	46
4358054	4358129	-	*cadB*	76
4492620	4492646	+	*indK*	27

Table B.4: Riboswitch Statistics. Average and standard deviation values of Length, MFE, and GC-compositions of the training and test sets. Column `Sense` denotes riboswitches. Column `UTR` denotes *E. coli* UTR sequences collected.

Total	L	MFE	GC	std(L)	std(MFE)	std(GC)
Sense	117.04	-42.63	0.51	47.81	21.54	0.09
antisense	117.04	-37.73	0.51	47.81	19.55	0.09
UTR	49.53	-6.48	0.37	19.37	5.77	0.08

Train	L	MFE	GC	std(L)	std(MFE)	std(GC)
Sense	114.1	-41.05	0.52	49.27	23.83	0.1
antisense	114.1	-37.73	0.52	49.27	21.32	0.1
UTR	48.18	-5.4	0.35	18.27	5.1	0.08

Test	L	MFE	GC	std(L)	std(MFE)	std(GC)
Sense	120.74	-44.63	0.49	46.71	18.6	0.09
antisense	120.74	-37.74	0.49	46.71	17.54	0.09
UTR	51.31	-7.9	0.39	21.35	6.47	0.08

Table B.5: Classification Performance for Different Choices of Length. Sub-section `Variable Length` refers to results of actual sequence lenghts for both training and test sets (equal number of varying sequence lengths of 100, 150, and 200 from *E. coli* UTR chosen as negative set). Sub-sections `100`, `150`, and `200` refers to results where all sequences in the training and test sets have a constant length. Column `Features` denotes features used from the training set. $TP\%$ denotes percentage of true positives. $FP_1\%$ and $FP_2\%$ represent the percentages of antisense sequences and *E. coli* UTRs that are misclassified as riboswitches, respectively. Sensitivity denotes overall percentage of correctly classified sequences. Sig. denotes significant (less than 0.05 in the training set) features of the multinomial classifier.

Variable Length					
Features	$TP\%$	$FP_1\%$	$FP_2\%$	Sensitivity	Sig.
L,MFE,GC,RND	69.6	39.1	7.7	61	MFE,GC
L,MFE,GC,BJK	87.0	34.8	0.0	71.2	GC
L,MFE,GC	87.0	30.4	0.0	76.3	L,MFE
100					
Features	$TP\%$	$FP_1\%$	$FP_2\%$	Sensitivity	Sig.
MFE,GC,RND	69.6	26.1	7.7	66.1	-
MFE,GC,BJK	65.2	21.7	7.7	64.4	-
MFE,GC	56.5	21.7	15.4	64.4	-
150					
Features	$TP\%$	$FP_1\%$	$FP_2\%$	Sensitivity	Sig.
MFE,GC,RND	69.6	26.1	23.1	57.6	MFE,RND
MFE,GC,BJK	69.6	39.1	7.7	59.3	MFE
MFE,GC	69.6	39.1	0.0	64.4	-
200					
Features	$TP\%$	$FP_1\%$	$FP_2\%$	Sensitivity	Sig.
MFE,GC,RND	65.2	34.8	7.7	62.7	MFE
MFE,GC,BJK	78.3	34.8	7.7	66.1	MFE
MFE,GC	82.6	39.1	0.0	76.3	MFE

Table B.6: Classification Performance for Different Choices of Length in *B. subtilis*. Classifier Performance on the eleven *B. subtilis* riboswitches. Constant length of 100 nt, 150 nt, 157 nt, and 200 nt used for test. $TP\%$ denotes percentage of true positives. $FP_2\%$ represent the overall percentage of sequences that are classified as riboswitches within the *B. subtilis* genome. 50 nt, 75 nt, and 100 nt overlaps were used for for sliding windows of lengths 100 nt, 150 nt, and 200 nt, respectively. No overlaps were used for the 157 nt window. True Positive sequences were according to maximum overlap with the location of the actual length of riboswitches.

100nt window	$TP\%$	$FP_2\%$
LMFEGCRND	63.6	29.8
LMFEGCBJK	27.3	15.4
LMFEGC	18.2	9.0

150nt window	$TP\%$	$FP_2\%$
LMFEGCRND	72.7	20.5
LMFEGCBJK	63.6	9.6
LMFEGC	45.5	3.2

157nt window	$TP\%$	$FP_2\%$
LMFEGCRND	81.8	19.5
LMFEGCBJK	54.5	8.3
LMFEGC	63.6	2.1

200nt window	$TP\%$	$FP_2\%$
LMFEGCRND	72.7	14.2
LMFEGCBJK	45.5	6.7
LMFEGC	18.2	1.3

Table B.7: Top Classification Hits in *B. subtilis*. Top 50 hits of the forward and reverse strands of the *B. subtilis* intergenic regions using no-overlap 157 nt window and under the LMFEGCRND model. The ranking of each hit is denoted in column R. Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the multinomial regression likelihood of being a riboswitch under the LMFEGCRND model.

R	Start	End	Strand	Upstream Operon	Upstream Gene	Dist. to Upstream	Uracil	Dist. to Downstream	Downstream Gene	Downstream Operon	Probability
1	3717569	3717725	reverse	BSU36100	ywrD	-1486	0.4076	550	cotH	BSU36060	0.943
2	3717412	3717568	reverse	BSU36100	ywrD	-1643	0.4076	393	cotH	BSU36060	0.935
3	4134175	4134331	reverse	BSU40230	yydA	-182	0.3439	79	yydB	BSU40220	0.931
4	3714883	3715039	forward	BSU36030	ywrK	-859	0.3949	2277	cotG	BSU36070	0.922
5	748990	749146	forward	BSU06780	yeeC	-2912	0.414	707	yeeG	BSU06820	0.919
6	3666640	3666796	reverse	BSU35680	ggaB	-490	0.4968	1335	mnaA	BSU35660	0.908
7	3866327	3866483	reverse	BSU37690	ywfG	-1881	0.3503	79	eutD	BSU37660	0.903
8	681153	681309	forward	BSU06260	ydjN	-201	0.3885	5731	yeaB	BSU06320	0.899
9	2987548	2987704	reverse	BSU29200	accA	-104	0.4268	79	pfkA	BSU29190	0.898
10	1680274	1680430	forward	BSU16080	ylqH	63	0.3822	79	sucC	BSU16090	0.897
11	2730227	2730383	reverse	BSU26730	yrdF	-254	0.4204	79	azlB	BSU26720	0.896
12	2316268	2316424	reverse	BSU22040	ypbQ	-99	0.363	236	ypbR	BSU22030	0.896
13	2219985	2220141	forward	BSU20929	yoyI	-6828	0.4204	2277	yonP	BSU21030	0.896
14	688027	688183	forward	BSU06320	yeaB	-114	0.3885	79	yeaC	BSU06330	0.893
15	243578	243734	forward	BSU02170	ybfB	-5370	0.363	236	purT	BSU02230	0.89
16	984466	984622	reverse	BSU09120	yhcK	-1189	0.3885	79	cspB	BSU09100	0.889
17	2376780	2376936	forward	BSU22510	ypjC	-15199	0.4395	16564	ypzI	BSU22869	0.888
18	748205	748361	forward	BSU06780	yeeC	-2127	0.4395	1492	yeeG	BSU06820	0.886
19	3421066	3421222	reverse	BSU33340	sspJ	-320	0.3312	79	lysP	BSU33330	0.885
20	2093235	2093391	forward	BSU19200	desR	-852	0.4331	4789	yoyB	BSU19259	0.88
21	3941212	3941368	reverse	BSU38430	gspA	-3269	0.4777	1649	ywbA	BSU38390	0.879
22	1493630	1493786	forward	BSU14230	ykuV	-230	0.3503	79	rok	BSU14240	0.879
23	2531945	2532101	forward	BSU24210	yqiG	-14308	0.3439	9028	yqhQ	BSU24490	0.879
24	746478	746634	forward	BSU06780	yeeC	-400	0.5095	3219	yeeG	BSU06820	0.878
25	2096100	2096256	reverse	BSU19230	yocJ	-171	0.4268	393	yocI	BSU19220	0.877
26	300673	300829	forward	BSU02770	yccK	-1196	0.3822	79	ycdB	BSU02790	0.875
27	3373963	3374119	reverse	BSU32890	yusQ	-2575	0.4076	393	fadM	BSU32850	0.874
28	3686143	3686299	forward	BSU35770	tagC	-1298	0.4586	2591	gerBA	BSU35800	0.87
29	1335487	1335643	reverse	BSU12820	spoIISB	-12876	0.5032	13895	xre	BSU12510	0.868
30	4139318	4139474	forward	BSU40240	yycS	-3475	0.3567	864	rapG	BSU40300	0.865
31	1268672	1268828	forward	BSU11970	yjcS	-1	0.4458	79	yjdA	BSU11980	0.865
32	3685829	3685985	forward	BSU35770	tagC	-984	0.414	2905	gerBA	BSU35800	0.865
33	3681213	3681369	forward	BSU35670	gtaB	-14627	0.363	79	tagA	BSU35750	0.864
34	1122705	1122861	forward	BSU10490	sipV	55	0.4013	79	yhjG	BSU10500	0.86
35	3671690	3671846	reverse	BSU35700	tagH	-1795	0.3694	393	ggaA	BSU35690	0.859
36	2160701	2160857	forward	BSU20000	yosU	-1938	0.4395	9028	yosA	BSU20190	0.859
37	1097850	1098006	reverse	BSU10230	yhfH	-191	0.465	79	gltT	BSU10220	0.858
38	1467020	1467176	forward	BSU13960	ykwC	-342	0.414	707	pbpH	BSU13980	0.857
39	191850	192006	forward	BSU01590	ybaS	-12186	0.3057	2277	trnSL-Glu2	BSU_tRNA_75	0.856
40	20723	20879	forward	BSU00120	yaaE	-86	0.3185	79	serS	BSU00130	0.856
41	2691445	2691601	reverse	BSU26240	yqaO	-1121	0.3439	79	yqaQ	BSU26220	0.852
42	3158851	3159007	reverse	BSU30890	ytxO	-328	0.363	3376	ytdA	BSU30850	0.852
43	1958206	1958362	forward	BSU18190	yngC	-9863	0.3949	44353	iseA	BSU18380	0.852
44	557716	557872	forward	BSU05100	yddT	-188	0.3503	79	ydzN	BSU05109	0.851
45	3907629	3907785	reverse	BSU38100	ywcH	-2594	0.3567	393	ywcI	BSU38080	0.851
46	1926523	1926679	forward	BSU17950	yneJ	-1482	0.3949	79	citB	BSU18000	0.851
47	1017271	1017427	forward	BSU09400	spoVR	-139	0.363	1649	lytE	BSU09420	0.85
48	1493595	1493751	reverse	BSU14250	yknT	-729	0.4522	1649	ykuT	BSU14210	0.85
49	2477743	2477899	forward	BSU23830	yqjL	66	0.4076	1335	zwf	BSU23850	0.849
50	2769617	2769773	reverse	BSU27160	cypB	-4194	0.3185	1021	yrhP	BSU27100	0.849
51	2739991	2740147	reverse	BSU26830	yrpE	-1287	0.3694	3533	aadK	BSU26790	0.849
52	644384	644540	forward	BSU05940	gcp	-7	0.3694	2120	moaC	BSU05960	0.848
53	4039599	4039755	forward	BSU39100	yxiO	-23552	0.4713	1806	hutP	BSU39340	0.847
54	2203622	2203778	forward	BSU20580	yoqM	-7279	0.363	79	yopS	BSU20780	0.847
55	3014345	3014501	reverse	BSU29460	moaB	-90	0.3694	79	argG	BSU29450	0.847
56	749147	749303	forward	BSU06780	yeeC	-3069	0.363	550	yeeG	BSU06820	0.846
57	665425	665581	forward	BSU06130	ydjC	-677	0.3439	1963	gutB	BSU06150	0.846
58	2106272	2106428	reverse	BSU19360	odhB	-1154	0.3949	79	yocR	BSU19340	0.846
59	226409	226565	forward	BSU02050	ybdO	-82	0.3885	79	ybxG	BSU02060	0.844
60	2106333	2106489	forward	BSU19330	sodF	-1353	0.3885	79	yocS	BSU19350	0.844
61	308175	308331	forward	BSU02840	ycdG	48	0.3503	79	adcA	BSU02850	0.843
62	2678925	2679081	forward	BSU26050	yqdB	-427	0.363	12639	yqaP	BSU26230	0.843
63	3875571	3875727	reverse	BSU37760	rocC	-130	0.3121	79	ywfA	BSU37750	0.843
64	2433680	2433836	reverse	BSU23340	ypuB	-384	0.3885	236	ypzJ	BSU23328	0.843
65	2879134	2879290	reverse	BSU28190	engB	-669	0.4013	79	hemA	BSU28170	0.842
66	1533806	1533962	forward	BSU14610	pdhD	-445	0.3503	236	ykzW	BSU14629	0.841
67	368137	368293	forward	BSU03360	yciC	-802	0.3312	1021	yckC	BSU03390	0.841
68	447000	447156	forward	BSU03930	gdh	-792	0.3121	2120	ycnL	BSU03970	0.84
69	3726630	3726786	forward	BSU36160	ywqM	-2216	0.4331	7144	ywqB	BSU36270	0.84
70	543132	543288	reverse	BSU05000	yddK	-2955	0.4904	11697	immR	BSU04820	0.84
71	3268320	3268476	forward	BSU31810	yuzE	-4017	0.4522	8557	yukF	BSU31920	0.84
72	2065804	2065960	forward	BSU18960	yozM	-348	0.3758	8557	yobN	BSU19020	0.839
73	45296	45452	reverse	BSU01550	gerD	-113140	0.3822	236	abrB	BSU00370	0.837
74	2048779	2048935	reverse	BSU18810	yobA	-1092	0.363	550	yoaZ	BSU18790	0.836

75	3153718	3153874	reverse	BSU30850	ytdA	-938	0.3439	79	menF	BSU30830	0.836
76	3388260	3388416	reverse	BSU33040	fumC	-685	0.3312	393	yuzO	BSU33029	0.834
77	205252	205408	forward	BSU01820	adaB	-283	0.3248	79	ndhF	BSU01830	0.834
78	469269	469425	forward	BSU04160	mtlR	24	0.3121	79	ydaB	BSU04170	0.834
79	1868460	1868616	forward	BSU17360	ymzA	-7	0.3885	79	nrdI	BSU17370	0.834
80	3746069	3746225	forward	BSU36380	rapD	-577	0.3822	2905	ywoH	BSU36440	0.833
81	3467327	3467483	reverse	BSU33800	opuCD	-140	0.3376	79	sdpR	BSU33790	0.832
82	1264932	1265088	reverse	BSU11990	yjdB	-4722	0.5223	79	yjcM	BSU11910	0.832
83	1903262	1903418	reverse	BSU17690	yncM	-170	0.3822	1963	cotU	BSU17670	0.831
84	4204441	4204597	reverse	BSU40960	parB	-1036	0.414	79	yyaD	BSU40940	0.831
85	1017114	1017270	forward	BSU09400	spoVR	18	0.3376	1806	lytE	BSU09420	0.831
86	2709577	2709733	reverse	BSU26490	yrkJ	-346	0.3503	236	yrkK	BSU26480	0.829
87	955738	955894	forward	BSU08780	ygaJ	-74	0.3822	79	thiC	BSU08790	0.828
88	554386	554542	reverse	BSU05130	ydeB	-5686	0.4395	1963	lrpB	BSU05060	0.828
89	3988764	3988920	reverse	BSU38860	galE	-1105	0.293	79	yxkD	BSU38840	0.825
90	2186812	2186968	reverse	BSU20420	yorD	-94	0.2611	79	yorE	BSU20410	0.825
91	2926840	2926996	reverse	BSU28630	pheT	-89	0.3185	1021	yshA	BSU28610	0.823
92	2054401	2054557	reverse	BSU18840	xynA	-119	0.3822	550	pps	BSU18830	0.822
93	610963	611119	reverse	BSU05660	ydgI	-1149	0.3121	2277	dinB	BSU05630	0.822
94	3457144	3457300	reverse	BSU33700	opuBD	-2583	0.3185	707	yvzC	BSU33650	0.821
95	736435	736591	reverse	BSU06740	yefB	-2481	0.3376	3690	yerO	BSU06700	0.82
96	2061478	2061634	reverse	BSU18930	yobH	-1953	0.3439	864	yozJ	BSU18900	0.82
97	2262616	2262772	reverse	BSU21440	bdbB	-2530	0.3885	15151	youB	BSU21329	0.819
98	4118717	4118873	reverse	BSU40110	bglA	-2370	0.3758	5574	glxK	BSU40040	0.818
99	4204755	4204911	reverse	BSU40960	parB	-722	0.3949	393	yyaD	BSU40940	0.818
100	3648264	3648420	reverse	BSU35530	tagO	-311	0.363	1806	degS	BSU35500	0.818

Table B.8: Top Classification Hits in *B. subtilis* Uracil-comp. Constrained. Top 50 hits of the forward and reverse strands of the *B. subtilis* intergenic regions using no-overlap 157 nt window and under the LMFEGCRND model. Uracil composition constrained to that of the range of known riboswitches in *B. subtilis* (between 0.2484 and 0.40127). The ranking of each hit is denoted in column R. Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the multinomial regression likelihood of being a riboswitch under the LMFEGCRND model.

R	Start	End	Strand	Upstream Operon	Upstream Gene	Dist. to Upstream	Uracil	Dist. to Downstream	Downstream Gene	Downstream Operon	Probability
1	4134175	4134331	reverse	BSU40230	yydA	-182	0.3439	79	yydB	BSU40220	0.931
2	3714883	3715039	forward	BSU36030	ywrK	-859	0.3949	2277	cotG	BSU36070	0.922
3	3866327	3866483	reverse	BSU37690	ywfG	-1881	0.3503	79	eutD	BSU37660	0.903
4	681153	681309	forward	BSU06260	ydjN	-201	0.3885	5731	yeaB	BSU06320	0.899
5	1680274	1680430	forward	BSU16080	ylqH	63	0.3822	79	sucC	BSU16090	0.897
6	2316268	2316424	reverse	BSU22040	ypbQ	-99	0.363	236	ypbR	BSU22030	0.896
7	688027	688183	forward	BSU06320	yeaB	-114	0.3885	79	yeaC	BSU06330	0.893
8	243578	243734	forward	BSU02170	ybfB	-5370	0.363	236	purT	BSU02230	0.89
9	984466	984622	reverse	BSU09120	yhcK	-1189	0.3885	79	cspB	BSU09100	0.889
10	3421066	3421222	reverse	BSU33340	sspJ	-320	0.3312	79	lysP	BSU33330	0.885
11	1493630	1493786	forward	BSU14230	ykuV	-230	0.3503	79	rok	BSU14240	0.879
12	2531945	2532101	forward	BSU24210	yqiG	-14308	0.3439	9028	yqhQ	BSU24490	0.879
13	300673	300829	forward	BSU02770	yccK	-1196	0.3822	79	ycdB	BSU02790	0.875
14	4139318	4139474	forward	BSU40240	yycS	-3475	0.3567	864	rapG	BSU40300	0.865
15	3681213	3681369	forward	BSU35670	gtaB	-14627	0.363	79	tagA	BSU35750	0.864
16	3671690	3671846	reverse	BSU35700	tagH	-1795	0.3694	393	ggaA	BSU35690	0.859
17	191850	192006	forward	BSU01590	ybaS	-12186	0.3057	2277	trnSL-Glu2	BSU_tRNA_75	0.856
18	20723	20879	forward	BSU00120	yaaE	-86	0.3185	79	serS	BSU00130	0.856
19	2691445	2691601	reverse	BSU26240	yqaO	-1121	0.3439	79	yqaQ	BSU26220	0.852
20	3158851	3159007	reverse	BSU30890	ytxO	-328	0.363	3376	ytdA	BSU30850	0.852
21	1958206	1958362	forward	BSU18190	yngC	-9863	0.3949	44353	iseA	BSU18380	0.852
22	557716	557872	forward	BSU05100	yddT	-188	0.3503	79	ydzN	BSU05109	0.851
23	3907629	3907785	reverse	BSU38100	ywcH	-2594	0.3567	393	ywcI	BSU38080	0.851
24	1926523	1926679	forward	BSU17950	yneJ	-1482	0.3949	79	citB	BSU18000	0.851
25	1017271	1017427	forward	BSU09400	spoVR	-139	0.363	1649	lytE	BSU09420	0.85
26	2769617	2769773	reverse	BSU27160	cypB	-4194	0.3185	1021	yrhP	BSU27100	0.849
27	2739991	2740147	reverse	BSU26830	yrpE	-1287	0.3694	3533	aadK	BSU26790	0.849
28	644384	644540	forward	BSU05940	gcp	-7	0.3694	2120	moaC	BSU05960	0.848
29	2203622	2203778	forward	BSU20580	yoqM	-7279	0.363	79	yopS	BSU20780	0.847
30	3014345	3014501	reverse	BSU29460	moaB	-90	0.3694	79	argG	BSU29450	0.847
31	749147	749303	forward	BSU06780	yeeC	-3069	0.363	550	yeeG	BSU06820	0.846
32	665425	665581	forward	BSU06130	ydjC	-677	0.3439	1963	gutB	BSU06150	0.846
33	2106272	2106428	reverse	BSU19360	odhB	-1154	0.3949	79	yocR	BSU19340	0.846
34	226409	226565	forward	BSU02050	ybdO	-82	0.3885	79	ybxG	BSU02060	0.844
35	2106333	2106489	forward	BSU19330	sodF	-1353	0.3885	79	yocS	BSU19350	0.844
36	308175	308331	forward	BSU02840	ycdG	48	0.3503	79	adcA	BSU02850	0.843
37	2678925	2679081	forward	BSU26050	yqdB	-427	0.363	12639	yqaP	BSU26230	0.843
38	3875571	3875727	reverse	BSU37760	rocC	-130	0.3121	79	ywfA	BSU37750	0.843
39	2433680	2433836	reverse	BSU23340	ypuB	-384	0.3885	236	ypzJ	BSU23328	0.843
40	1533806	1533962	forward	BSU14610	pdhD	-445	0.3503	236	ykzW	BSU14629	0.841
41	368137	368293	forward	BSU03360	yciC	-802	0.3312	1021	yckC	BSU03390	0.841
42	447000	447156	forward	BSU03930	gdh	-792	0.3121	2120	ycnL	BSU03970	0.84
43	2065804	2065960	forward	BSU18960	yozM	-348	0.3758	8557	yobN	BSU19020	0.839
44	45296	45452	reverse	BSU01550	gerD	-113140	0.3822	236	abrB	BSU00370	0.837
45	2048779	2048935	reverse	BSU18810	yobA	-1092	0.363	550	yoaZ	BSU18790	0.836
46	3153718	3153874	reverse	BSU30850	ytdA	-938	0.3439	79	menF	BSU30830	0.836
47	3388260	3388416	reverse	BSU33040	fumC	-685	0.3312	393	yuzO	BSU33029	0.834
48	205252	205408	forward	BSU01820	adaB	-283	0.3248	79	ndhF	BSU01830	0.834
49	469269	469425	forward	BSU04160	mtlR	24	0.3121	79	ydaB	BSU04170	0.834
50	1868460	1868616	forward	BSU17360	ymzA	-7	0.3885	79	nrdI	BSU17370	0.834
51	3746069	3746225	forward	BSU36380	rapD	-577	0.3822	2905	ywoH	BSU36440	0.833
52	3467327	3467483	reverse	BSU33800	opuCD	-140	0.3376	79	sdpR	BSU33790	0.832
53	1903262	1903418	reverse	BSU17690	yncM	-170	0.3822	1963	cotU	BSU17670	0.831
54	1017114	1017270	forward	BSU09400	spoVR	18	0.3376	1806	lytE	BSU09420	0.831
55	2709577	2709733	reverse	BSU26490	yrkJ	-346	0.3503	236	yrkK	BSU26480	0.829
56	955738	955894	forward	BSU08780	ygaJ	-74	0.3822	79	thiC	BSU08790	0.828
57	1283149	1283305	forward	BSU12100	yjeA	-539	0.3758	236	yjfC	BSU12130	0.827
58	3988764	3988920	reverse	BSU38860	galE	-1105	0.293	79	yxkD	BSU38840	0.825
59	200120	200276	forward	BSU01770	glmM	-198	0.2611	79	glmS	BSU01780	0.825
60	2186812	2186968	reverse	BSU20420	yorD	-94	0.2611	79	yorE	BSU20410	0.825
61	2926840	2926996	reverse	BSU28630	pheT	-89	0.3185	1021	yshA	BSU28610	0.823
62	2054401	2054557	reverse	BSU18840	xynA	-119	0.3822	550	pps	BSU18830	0.822
63	610963	611119	reverse	BSU05660	ydgI	-1149	0.3121	2277	dinB	BSU05630	0.822
64	3457144	3457300	reverse	BSU33700	opuBD	-2583	0.3185	707	yvzC	BSU33650	0.821
65	736435	736591	reverse	BSU06740	yefB	-2481	0.3376	3690	yerO	BSU06700	0.82
66	2061478	2061634	reverse	BSU18930	yobH	-1953	0.3439	864	yozJ	BSU18900	0.82
67	3268477	3268633	forward	BSU31810	yuzE	-4174	0.3949	8400	yukF	BSU31920	0.82
68	3107044	3107200	forward	BSU30340	ytvA	30	0.2675	1492	yttA	BSU30360	0.82
69	2262616	2262772	reverse	BSU21440	bdbB	-2530	0.3885	15151	youB	BSU21329	0.819
70	4118717	4118873	reverse	BSU40110	bglA	-2370	0.3758	5574	glxK	BSU40040	0.818
71	4204755	4204911	reverse	BSU40960	parB	-722	0.3949	393	yyaD	BSU40940	0.818
72	252357	252513	forward	BSU02320	ybfP	36	0.3822	79	ybfQ	BSU02330	0.818

73	3648264	3648420	reverse	BSU35530	tagO	-311	0.363	1806	degS	BSU35500	0.818
74	850053	850209	forward	BSU07750	yflA	-3789	0.3694	236	treP	BSU07800	0.817
75	255279	255435	forward	BSU02330	ybfQ	-1718	0.2994	2434	ybgA	BSU02370	0.816
76	1541729	1541885	forward	BSU14680	ykzC	-2958	0.3376	79	ylaA	BSU14710	0.816
77	909862	910018	forward	BSU08330	yfiN	-658	0.3503	79	estB	BSU08350	0.816
78	4109617	4109773	reverse	BSU40030	yxaB	-1253	0.3885	79	yxaD	BSU40010	0.813
79	3252983	3253139	forward	BSU31660	mrpG	-538	0.3949	4632	yuzC	BSU31730	0.811
80	4066294	4066450	reverse	BSU39600	yxeC	-234	0.3567	864	yxeF	BSU39570	0.81
81	1923077	1923233	forward	BSU17910	yneF	-231	0.3248	79	ccdA	BSU17930	0.809
82	1540787	1540943	forward	BSU14680	ykzC	-2016	0.3503	1021	ylaA	BSU14710	0.809
83	3665472	3665628	forward	BSU35650	lytR	-1192	0.3376	79	gtaB	BSU35670	0.808
84	1679031	1679187	reverse	BSU17060	ymzD	-101508	0.3503	7458	ylqB	BSU15960	0.808
85	2698717	2698873	reverse	BSU26360	yqaD	-714	0.363	79	yqaF	BSU26340	0.808
86	3604725	3604881	reverse	BSU35100	yvlD	-1958	0.363	236	yvmC	BSU35070	0.808
87	3354671	3354827	forward	BSU32650	yurS	-105	0.3057	17820	yuzL	BSU32849	0.807
88	3052234	3052390	forward	BSU29710	acuC	-9600	0.3503	2434	ytoQ	BSU29850	0.806
89	188867	189023	forward	BSU01590	ybaS	-9203	0.3185	5260	trnSL-Glu2	BSU_tRNA_75	0.806
90	245389	245545	reverse	BSU02340	gltP	-8050	0.363	1806	ybfI	BSU02220	0.805
91	1445373	1445529	reverse	BSU13810	ykvS	-2210	0.3439	2748	ykvN	BSU13760	0.804
92	2249114	2249270	reverse	BSU21440	bdbB	-16032	0.3439	1649	youB	BSU21329	0.803
93	3918262	3918418	reverse	BSU38190	galT	-752	0.3057	79	qoxA	BSU38170	0.801
94	933760	933916	reverse	BSU08620	yfhP	-618	0.3439	5574	sspK	BSU08550	0.8
95	201248	201404	reverse	BSU01800	alkA	-1220	0.293	7301	ybbK	BSU01720	0.8
96	3684268	3684424	reverse	BSU35780	lytD	-479	0.3439	3376	tagD	BSU35740	0.8
97	2739834	2739990	reverse	BSU26830	yrpE	-1444	0.3439	3376	aadK	BSU26790	0.799
98	2252097	2252253	reverse	BSU21440	bdbB	-13049	0.3949	4632	youB	BSU21329	0.798
99	1601271	1601427	reverse	BSU15640	yloA	-34781	0.3503	24100	ylbP	BSU15100	0.797
100	2111609	2111765	reverse	BSU19380	yojO	-149	0.3439	79	sucA	BSU19370	0.796

Table B.9: Top Entropy Hits in *B. subtilis* Forward Strand. Significant hits of the forward and reverse strands (only showing forward strand here) of the *B. subtilis* intergenic regions having significantly high RND entropy (p-Val.<0.0500) and LMFEGCRND probability higher than 0.8. 100 nt overlap used for 200 nt scan (44847 segments). Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the likelihood of being a riboswitch under the LMFEGCRND model. Negative values indicate distance to upstream operon. Columns `Upsream/Downstream Operon` show gene ID within the operon.

B. subtilis	Start	End	Strand	Upstream Operon	Upstream Gene	Dist. to Upstream	MFE	MFE p. Val.	GC	RND	RND p. Val.	Uracil	Dist. to Downstream	Downstream Gene	Downstream Operon	Probability
200nt	3714838	3715037	forward	BSU36030	ywrK	-794	-49.70	-	0.3300	126.0619965	-	0.3850	2302	cotG	BSU36070	0.9275143743
200nt	3359769	3359968	forward	BSU32650	yurS	-5183	-66.40	-	0.4600	123.4530029	-	0.3450	12702	yuzL	BSU32849	0.9236087203
200nt	243592	243791	forward	BSU02170	ybfB	-5364	-57.00	-	0.4450	126.5479965	-	0.3700	202	purT	BSU02230	0.9204539061
200nt	2093202	2093401	forward	BSU19200	desR	-799	-55.20	-	0.4350	126.5859985	-	0.4400	4802	yoyB	BSU19259	0.9146069884
200nt	749075	749274	forward	BSU06780	yeeC	-2977	-58.19	-	0.4450	125.6159973	-	0.3900	602	yeeG	BSU06820	0.9128865004
200nt	1467005	1467204	forward	BSU13960	ykwC	-307	-60.40	-	0.4100	123.2139969	-	0.4150	702	pbpH	BSU13980	0.9070840478
200nt	2281367	2281566	forward	BSU21620	yokE	95	-44.60	-	0.2600	124.4599991	-	0.4300	202	yokD	BSU21630	0.9058990479
200nt	850067	850266	forward	BSU07750	yflA	-3783	-57.20	-	0.4000	124.1029968	-	0.3800	202	treP	BSU07800	0.9058393836
200nt	1466905	1467104	forward	BSU13960	ykwC	-207	-56.93	-	0.3900	123.7220001	-	0.4200	802	pbpH	BSU13980	0.9029595852
200nt	3759694	3759893	forward	BSU36530	bcrC	-467	-62.40	-	0.3850	121.3089981	-	0.4500	902	ywnH	BSU36560	0.9023656249
200nt	3268355	3268554	forward	BSU31810	yuzE	-4032	-44.66	-	0.3750	127.7630005	-	0.4450	8502	yukF	BSU31920	0.8946693540
200nt	2073039	2073238	forward	BSU18960	yozM	-7563	-52.00	-	0.3250	123.0869980	-	0.4450	1302	yobN	BSU19020	0.8944090009
200nt	748975	749174	forward	BSU06780	yeeC	-2877	-58.80	-	0.4750	125.1039963	-	0.3750	702	yeeG	BSU06820	0.8876969814
200nt	432172	432371	forward	BSU03780	phrC	-1988	-51.70	-	0.3050	122.0960007	-	0.3450	102	yclN	BSU03800	0.8850299716
200nt	4039583	4039782	forward	BSU39100	yxiO	-23516	-53.00	-	0.4350	125.9430008	-	0.4350	1802	hutP	BSU39340	0.8843178153
200nt	531587	531786	forward	BSU_tRNA_51	trnS-Leu2	-2066	-46.60	-	0.2650	122.5979996	-	0.3850	102	sacV	BSU04830	0.8803396225
200nt	665366	665565	forward	BSU06130	ydjC	-598	-64.50	-	0.4900	122.9229965	-	0.3300	2002	gutB	BSU06150	0.8790112138
200nt	3714938	3715137	forward	BSU36030	ywrK	-894	-41.70	-	0.3350	126.8040009	-	0.3550	2202	cotG	BSU36070	0.8768669963
200nt	3685812	3686011	forward	BSU35770	tagC	-947	-51.80	-	0.3900	124.4950027	-	0.4150	2902	gerBA	BSU35800	0.8747953176
200nt	2093302	2093501	forward	BSU19200	desR	-899	-61.40	-	0.4400	122.2649994	-	0.3700	4702	yoyB	BSU19259	0.8738172650
200nt	1012447	1012646	forward	BSU09360	yhdC	-598	-51.20	-	0.3650	123.8079987	-	0.4100	3102	spoVR	BSU09400	0.8727893829
200nt	955595	955794	forward	BSU08780	ygaJ	89	-59.79	-	0.4050	121.6190033	-	0.3600	202	thiC	BSU08790	0.8710888028
200nt	3685912	3686111	forward	BSU35770	tagC	-1047	-53.60	-	0.3950	123.7610016	-	0.3950	2802	gerBA	BSU35800	0.8705116510
200nt	4134651	4134850	forward	BSU40190	fbp	-4508	-58.20	-	0.4900	125.0049973	-	0.3900	602	yycS	BSU40240	0.8676616549
200nt	45433	45632	forward	BSU00360	yabC	-535	-42.70	-	0.3000	124.7969971	-	0.4200	102	metS	BSU00380	0.8666248322
200nt	2531951	2532150	forward	BSU24210	yqiG	-14294	-62.90	-	0.5050	123.5699997	-	0.3450	9002	yqhQ	BSU24490	0.8666005135
200nt	1540786	1540985	forward	BSU14680	ykzC	-1995	-57.52	-	0.4350	123.3629990	-	0.3750	1002	ylaA	BSU14710	0.8663020134
200nt	1406312	1406511	forward	BSU13390	ykoT	-1721	-71.55	-	0.5450	121.3809967	-	0.3250	3502	ykoX	BSU13430	0.8654530644
200nt	4029283	4029482	forward	BSU39100	yxiO	-13216	-50.90	-	0.3850	124.3079987	-	0.4000	12102	hutP	BSU39340	0.8653051257
200nt	1526531	1526730	forward	BSU14550	ykrA	-273	-63.20	-	0.4800	122.5350037	-	0.3350	602	ykyA	BSU14570	0.8648978472
200nt	2220040	2220239	forward	BSU20929	yoyI	-6863	-38.72	-	0.3800	129.0460052	-	0.3700	2202	yonP	BSU21030	0.8647925854
200nt	192105	192304	forward	BSU01590	ybaS	-12421	-51.20	-	0.4350	125.8399963	-	0.4100	2002	trnSL-Glu2	BSU_tRNA_75	0.8642573953
200nt	1780406	1780605	forward	BSU17050	mutL	-87	-52.60	-	0.3550	122.5059967	-	0.3650	1402	pksA	BSU17080	0.8627771139
200nt	4037783	4037982	forward	BSU39100	yxiO	-21716	-57.74	-	0.4400	123.2249985	-	0.3950	3602	hutP	BSU39340	0.8606380820
200nt	1264357	1264556	forward	BSU_tRNA_83	trnSL-Val2	-1397	-33.05	-	0.2700	127.4010010	-	0.4200	602	yjcN	BSU11920	0.8592507839
200nt	847767	847966	forward	BSU07750	yflA	-1483	-59.62	-	0.4500	122.6080017	-	0.3550	2502	treP	BSU07800	0.8554052114
200nt	226266	226465	forward	BSU02050	ybdO	81	-37.20	-	0.2300	124.2030029	-	0.3550	202	ybxG	BSU02060	0.8548350334
200nt	3052246	3052445	forward	BSU29710	acuC	-9592	-61.90	-	0.4900	123.0039978	-	0.3400	2402	ytoQ	BSU29850	0.8543979526
200nt	530887	531086	forward	BSU_tRNA_51	trnS-Leu2	-1366	-42.20	-	0.3250	125.3190002	-	0.4000	802	sacV	BSU04830	0.8526363969
200nt	2617117	2617316	forward	BSU25220	antE	-13743	-59.90	-	0.4500	122.3389969	-	0.4200	3502	yqeW	BSU25420	0.8511158824
200nt	2221540	2221739	forward	BSU20929	yoyI	-8363	-46.30	-	0.2850	122.2210007	-	0.3450	702	yonP	BSU21030	0.8502966762
200nt	2054178	2054377	forward	BSU18820	yobB	-3127	-48.70	-	0.3450	123.2819977	-	0.4000	2002	yobD	BSU18850	0.8502687216
200nt	3042946	3043145	forward	BSU29710	acuC	-292	-49.94	-	0.4350	125.8190002	-	0.4150	11702	ytoQ	BSU29850	0.8499680161
200nt	2780909	2781108	forward	BSU27150	yrhK	-7164	-47.80	-	0.3250	122.8980026	-	0.3750	202	yrhE	BSU27220	0.8482846022
200nt	2723337	2723536	forward	BSU26630	yrdQ	-594	-44.00	-	0.2950	123.3580017	-	0.4300	2402	gltR	BSU26670	0.8465546370
200nt	683762	683961	forward	BSU06260	ydjN	-2790	-59.40	-	0.4800	123.3310013	-	0.3600	3102	yeaB	BSU06320	0.8446838856
200nt	3052346	3052545	forward	BSU29710	acuC	-9692	-56.62	-	0.4400	123.0849991	-	0.3800	2302	ytoQ	BSU29850	0.8442399502
200nt	2405829	2406028	forward	BSU22869	ypzI	-12171	-60.30	-	0.4650	122.4049988	-	0.4150	3802	fer	BSU23040	0.8431282043
200nt	579341	579540	forward	BSU05329	ydzO	-10	-47.40	-	0.3850	124.9130020	-	0.3050	102	aseR	BSU05330	0.8431255221
200nt	748875	749074	forward	BSU06780	yeeC	-2777	-60.60	-	0.4600	122.0989990	-	0.3950	802	yeeG	BSU06820	0.8426845670
200nt	339225	339424	forward	BSU03130	nadE	-20	-63.60	-	0.4700	121.1809998	-	0.3250	702	aroK	BSU03150	0.8414211273
200nt	1251377	1251576	forward	BSU11730	cotO	-1930	-35.49	-	0.3200	127.4260025	-	0.4800	702	yjcA	BSU11790	0.8401102424
200nt	3640353	3640552	forward	BSU35210	yvkA	-19956	-52.60	-	0.3600	121.8539963	-	0.4250	6302	yvyE	BSU35510	0.8399478197
200nt	3686112	3686311	forward	BSU35770	tagC	-1247	-51.99	-	0.3750	122.5879974	-	0.4300	2602	gerBA	BSU35800	0.8393257856
200nt	1494405	1494604	forward	BSU14240	rok	56	-52.70	-	0.4100	123.4729996	-	0.4150	1002	mobA	BSU14260	0.8389207721
200nt	373532	373731	forward	BSU03410	bglC	-1741	-56.90	-	0.4500	123.1060028	-	0.3800	2402	hxlR	BSU03470	0.8382304311
200nt	3686212	3686411	forward	BSU35770	tagC	-1347	-45.70	-	0.4000	125.9280014	-	0.4250	2502	gerBA	BSU35800	0.8377878666
200nt	374532	374731	forward	BSU03410	bglC	-2741	-56.99	-	0.3750	120.5049973	-	0.3400	1402	hxlR	BSU03470	0.8375294805
200nt	1540186	1540385	forward	BSU14680	ykzC	-1395	-61.11	-	0.4650	121.7969971	-	0.3500	1602	ylaA	BSU14710	0.8347978592
200nt	360837	361036	forward	BSU03270	ycgT	-6870	-59.70	-	0.5000	123.5510025	-	0.3300	2002	nasA	BSU03330	0.8347288966
200nt	213641	213840	forward	BSU01900	ybcM	-73	-36.30	-	0.2750	125.3079987	-	0.3950	202	skfA	BSU01910	0.8321032524
200nt	739678	739877	forward	BSU06730	yefA	-597	-51.16	-	0.3900	123.1309967	-	0.4150	102	yefC	BSU06750	0.8301935792
200nt	1495005	1495204	forward	BSU14240	rok	-544	-50.26	-	0.4150	124.2959976	-	0.3800	402	mobA	BSU14260	0.8287579417
200nt	1541686	1541885	forward	BSU14680	ykzC	-2895	-43.10	-	0.3250	124.1309967	-	0.3650	102	ylaA	BSU14710	0.8283772469
200nt	1268629	1268828	forward	BSU11970	yjcS	62	-40.70	-	0.2700	123.2060013	-	0.4250	102	yjdA	BSU11980	0.8273611665
200nt	652232	652431	forward	BSU06030	groEL	-265	-37.70	-	0.2950	125.2659988	-	0.4500	1102	ydiM	BSU06040	0.8273396492
200nt	2108093	2108292	forward	BSU19350	yocS	-539	-59.80	-	0.4700	122.2429962	-	0.3700	11202	yojI	BSU19440	0.8269666433
200nt	728532	728731	forward	BSU06640	yerI	-2436	-46.30	-	0.3100	122.2779999	-	0.3600	102	gatC	BSU06670	0.8268005848
200nt	1540686	1540885	forward	BSU14680	ykzC	-1895	-60.72	-	0.4350	120.6729965	-	0.3350	1102	ylaA	BSU14710	0.8265900016
200nt	3746052	3746251	forward	BSU36380	rapD	-540	-57.10	-	0.4350	122.1200027	-	0.3850	2902	ywoH	BSU36440	0.8260388970
200nt	1495105	1495304	forward	BSU14240	rok	-644	-55.30	-	0.3950	121.4899979	-	0.4000	302	mobA	BSU14260	0.8259468675
200nt	1923034	1923233	forward	BSU17910	yneF	-168	-40.60	-	0.2500	122.5039978	-	0.3500	102	ccdA	BSU17930	0.8253148198
200nt	746475	746674	forward	BSU06780	yeeC	-377	-31.51	-	0.2650	126.6760025	-	0.5000	3202	yeeG	BSU06820	0.8248795867
200nt	2625315	2625514	forward	BSU25420	yqeW	-3576	-54.30	-	0.4850	124.8850021	-	0.3550	10402	rpsT	BSU25550	0.8240758777
200nt	4007404	4007603	forward	BSU39020	yxjA	-360	-48.97	-	0.4150	124.6660004	-	0.3950	2902	citH	BSU39060	0.8239642382
200nt	2376722	2376921	forward	BSU22510	ypjC	-15121	-44.40	-	0.3450	124.1429977	-	0.4000	16602	ypzI	BSU22869	0.8239628077
200nt	3640453	3640652	forward	BSU35210	yvkA	-20056	-45.52	-	0.3450	123.6029968	-	0.4400	6202	yvyE	BSU35510	0.8211722970
200nt	530787	530986	forward	BSU_tRNA_51	trnS-Leu2	-1266	-47.10	-	0.3350	122.6100006	-	0.3850	902	sacV	BSU04830	0.8206871748
200nt	4139260	4139459	forward	BSU40240	yycS	-3397	-60.10	-	0.5050	123.1039963	-	0.3450	902	rapG	BSU40300	0.8204663396
200nt	184305	184504	forward	BSU01590	ybaS	-4621	-44.20	-	0.4000	125.9649963	-	0.4400	9802	trnSL-Glu2	BSU_tRNA_75	0.8200225234
200nt	792182	792381	forward	BSU07230	yetM	-656	-69.00	-	0.5400	120.5859985	-	0.3000	402	yetO	BSU07250	0.8170907497
200nt	4007304	4007503	forward	BSU39020	yxjA	-260	-49.30	-	0.4200	124.4300003	-	0.3950	3002	citH	BSU39060	0.8151187301
200nt	3726352	3726551	forward	BSU36160	ywqM	-1918	-63.90	-	0.5250	122.0380020	-	0.3200	7402	ywqB	BSU36270	0.8136813045
200nt	3201391	3201590	forward	BSU31170	yulF	-3400	-50.17	-	0.4150	123.8629990	-	0.3950	11102	tgl	BSU31270	0.8136008978
200nt	3714738	3714937	forward	BSU36030	ywrK	-694	-38.82	-	0.3050	124.7269974	-	0.4300	2402	cotG	BSU36070	0.8135811090
200nt	2160707	2160906	forward	BSU20000	yosU	-1924	-34.70	-	0.3050	126.3850021	-	0.4300	9002	yosA	BSU20190	0.8132891059
200nt	192805	193004	forward	BSU01590	ybaS	-13121	-48.80	-	0.4150	124.4029999	-	0.3650	1302	trnSL-Glu2	BSU_tRNA_75	0.8131257892
200nt	2217640	2217839	forward	BSU20929	yoyI	-4463	-44.10	-	0.2800	121.7109985	-	0.4100	4602	yonP	BSU21030	0.8124790788
200nt	182405	182604	forward	BSU01590	ybaS	-2721	-53.80	-	0.4150	122.3399963	-	0.3550	11702	trnSL-Glu2	BSU_tRNA_75	0.8117757440
200nt	2276877	2277076	forward	BSU21520	yolC	-4287	-38.90	-	0.3250	125.3180008	-	0.4500	3002	yokF	BSU21610	0.8117634654
200nt	1997137	1997336	forward	BSU18190	yngC	-48774	-50.97	-	0.4000	122.9609985	-	0.3250	5402	iseA	BSU18380	0.8112495542
200nt	2276977	2277176	forward	BSU21520	yolC	-4387	-39.30	-	0.3200	124.9540024	-	0.4200	2902	yokF	BSU21610	0.8106592894
200nt	749175	749374	forward	BSU06780	yeeC	-3077	-58.49	-	0.4600	121.9140015	-	0.3700	502	yeeG	BSU06820	0.8099753261
200nt	3726652	3726851	forward	BSU36160	ywqM	-2218	-52.11	-	0.4550	124.3000031	-	0.3950	7102	ywqB	BSU36270	0.8091073036
200nt	1251277	1251476	forward	BSU11730	cotO	-1830	-37.50	-	0.3650	127.1539993	-	0.5150	802	yjcA	BSU11790	0.8088676929
200nt	1405212	1405411	forward	BSU13390	ykoT	-621	-56.26	-	0.4050	120.9140015	-	0.3500	4602	ykoX	BSU13430	0.8086636066
200nt	3268455	3268654	forward	BSU31810	yuzE	-4132	-52.90	-	0.3900	121.7509995	-	0.4200	8402	yukF	BSU31920	0.8081850410
200nt	1467105	1467304	forward	BSU13960	ykwC	-407	-41.17	-	0.3700	125.7789993	-	0.4650	602	pbpH	BSU13980	0.8068280816
200nt	1406412	1406611	forward	BSU13390	ykoT	-1821	-67.24	-	0.5600	121.6240005	-	0.3150	3402	ykoX	BSU13430	0.8052443266
200nt	2897488	2897687	forward	BSU28180	ysxD	-17529	-28.00	-	0.2150	125.7649994	-	0.4100	202	ysnD	BSU28320	0.8036175966
200nt	1474672	1474871	forward	BSU14010	cheV	-57	-50.60	-	0.3800	122.1959991	-	0.3650	3302	ykuF	BSU14060	0.8031342626
200nt	746375	746574	forward	BSU06780	yeeC	-277	-31.50	-	0.2800	126.4580002	-	0.5000	3302	yeeG	BSU06820	0.8004485369

B.4 *Escherichia coli* Genome-wide Scan Results

Table B.10: Top Entropy Hits in *B. subtilis* Reverse Strand. Significant hits of the forward and reverse strands (only showing reverse strand here) of the *B. subtilis* intergenic regions having significantly high RND entropy (p-Val.<0.0500) and LMFEGCRND probability higher than 0.8. 100 nt overlap used for 200 nt scan (44847 segments). Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the multinomial regression likelihood of being a riboswitch under the LMFEGCRND model. Negative values indicate distance to upstream operon. Columns `Upsream/Downstream Operon` show gene ID within the operon.

B. subtilis	Start	End	Strand	Upstream Operon	Upstream Gene	Dist. to Upstream	MFE	MFE p. Val.	GC	RND	RND p. Val.	Uracil	Dist. to Downstream	Downstream Gene	Downstream Operon	Probability
200nt	3717398	3717597	reverse	BSU36100	ywrD	-1637	-51.30	-	0.3650	130.8540039	-	0.3950	399	cotH	BSU36060	0.9702541828
200nt	3717498	3717697	reverse	BSU36100	ywrD	-1537	-50.60	-	0.3500	129.2720032	-	0.4000	499	cotH	BSU36060	0.9603169560
200nt	4066209	4066408	reverse	BSU39600	yxeC	-299	-67.50	-	0.4900	125.4860001	-	0.3650	799	yxeF	BSU39570	0.9434255362
200nt	786306	786505	reverse	BSU07220	yetL	-3247	-72.30	-	0.5600	125.9049988	-	0.3050	499	yetH	BSU07160	0.9432973266
200nt	2249144	2249343	reverse	BSU21440	bdbB	-15982	-49.00	-	0.3700	127.3170013	-	0.3850	1699	youB	BSU21329	0.9216341376
200nt	2596201	2596400	reverse	BSU25170	yqfO	-1202	-42.96	-	0.3650	129.5290070	-	0.4950	799	cshB	BSU25140	0.9206426144
200nt	3717298	3717497	reverse	BSU36100	ywrD	-1737	-45.43	-	0.3750	128.8240051	-	0.4200	299	cotH	BSU36060	0.9199228883
200nt	3671576	3671775	reverse	BSU35700	tagH	-1889	-33.00	-	0.2350	128.6049957	-	0.4550	299	ggaA	BSU35690	0.9113640189
200nt	3717598	3717797	reverse	BSU36100	ywrD	-1437	-50.50	-	0.3850	126.3860016	-	0.4350	599	cotH	BSU36060	0.9073441625
200nt	3373949	3374148	reverse	BSU32890	yusQ	-2569	-67.60	-	0.4800	122.0930023	-	0.3600	399	fadM	BSU32850	0.8957566023
200nt	3666584	3666783	reverse	BSU35680	ggaB	-526	-44.06	-	0.3300	126.5329971	-	0.4650	1299	mnaA	BSU35660	0.8957416415
200nt	3941142	3941341	reverse	BSU38430	gspA	-3319	-43.33	-	0.3950	128.7100067	-	0.4800	1599	ywbA	BSU38390	0.8889677525
200nt	2879134	2879333	reverse	BSU28190	engB	-649	-38.80	-	0.2800	126.4779968	-	0.4100	99	hemA	BSU28170	0.8852627277
200nt	3907615	3907814	reverse	BSU38100	ywcH	-2588	-46.12	-	0.2950	123.9560013	-	0.3600	399	ywcI	BSU38080	0.8837128878
200nt	1248822	1249021	reverse	BSU11740	cotZ	-521	-54.90	-	0.5300	128.1889954	-	0.4000	8799	yjbP	BSU11630	0.8796436787
200nt	2004047	2004246	reverse	BSU18400	yoeD	-116	-40.10	-	0.3150	126.8629990	-	0.3900	199	yoeC	BSU18390	0.8789740205
200nt	3671676	3671875	reverse	BSU35700	tagH	-1789	-35.80	-	0.2600	126.5380020	-	0.3900	399	ggaA	BSU35690	0.8741892576
200nt	2257744	2257943	reverse	BSU21440	bdbB	-7382	-38.35	-	0.3350	128.0359955	-	0.4600	10299	youB	BSU21329	0.8739098310
200nt	2294238	2294437	reverse	BSU21800	ypkP	-1645	-57.60	-	0.3900	122.0520020	-	0.3950	299	ilvA	BSU21770	0.8724481463
200nt	791156	791355	reverse	BSU07240	yetN	-207	-70.84	-	0.5850	123.2480011	-	0.2900	1099	yetL	BSU07220	0.8711011410
200nt	494742	494941	reverse	BSU04430	ydbD	-899	-52.90	-	0.4250	125.0230026	-	0.4350	1399	ydaT	BSU04380	0.8695754409
200nt	1355092	1355291	reverse	BSU12900	htrA	-2745	-71.91	-	0.5400	121.2139969	-	0.3050	2499	ykbA	BSU12860	0.8692007065
200nt	737924	738123	reverse	BSU06740	yefB	-972	-70.30	-	0.5400	121.8629990	-	0.3150	5199	yerO	BSU06700	0.8691427112
200nt	937998	938197	reverse	BSU08700	ygaE	-3071	-54.20	-	0.4350	124.7170029	-	0.4200	1099	yfhS	BSU08640	0.8665797114
200nt	3421066	3421265	reverse	BSU33340	sspJ	-300	-64.50	-	0.4850	122.1780014	-	0.3150	99	lysP	BSU33330	0.8648849726
200nt	2739937	2740136	reverse	BSU26830	yrpE	-1321	-49.69	-	0.4200	125.9660034	-	0.3800	3499	aadK	BSU26790	0.8648592830
200nt	1097850	1098049	reverse	BSU10230	yhfH	-171	-36.90	-	0.2550	125.4639969	-	0.4450	99	gltT	BSU10220	0.8626419306
200nt	3851617	3851816	reverse	BSU37520	ywhD	-470	-68.42	-	0.5500	122.7009964	-	0.3000	2099	speE	BSU37500	0.8624630570
200nt	2724828	2725027	reverse	BSU26660	yrdN	-187	-37.44	-	0.2200	124.0479965	-	0.3800	99	czcD	BSU26650	0.8623370528
200nt	1204503	1204702	reverse	BSU11270	yjzD	-129	-59.84	-	0.5050	124.6380005	-	0.3600	13299	yitU	BSU11140	0.8622197509
200nt	2692345	2692544	reverse	BSU26240	yqaO	-201	-47.40	-	0.3350	123.8980026	-	0.4100	999	yqaQ	BSU26220	0.8618891239
200nt	3108926	3109125	reverse	BSU30370	bceB	-372	-32.52	-	0.3250	129.4949951	-	0.5250	599	yttB	BSU30350	0.8596788645
200nt	1493525	1493724	reverse	BSU14250	yknT	-779	-39.60	-	0.2150	122.8750000	-	0.4750	1599	ykuT	BSU14210	0.8589127660
200nt	2111609	2111808	reverse	BSU19380	yojO	-129	-37.20	-	0.2700	125.5370026	-	0.4050	99	sucA	BSU19370	0.8542534709
200nt	1678852	1679051	reverse	BSU17060	ymzD	-101667	-62.81	-	0.4750	122.0299988	-	0.3300	7299	ylqB	BSU15960	0.8517054319
200nt	4109617	4109816	reverse	BSU40030	yxaB	-1233	-52.14	-	0.3400	121.7229996	-	0.3900	99	yxaD	BSU40010	0.8503260016
200nt	3373849	3374048	reverse	BSU32890	yusQ	-2669	-56.90	-	0.5000	125.1539993	-	0.3500	299	fadM	BSU32850	0.8485122919
200nt	1886780	1886979	reverse	BSU17590	xylR	-3633	-42.30	-	0.3150	124.7409973	-	0.3500	13299	cwlC	BSU17410	0.8470579386
200nt	3153718	3153917	reverse	BSU30850	ytdA	-918	-51.80	-	0.3600	122.3779984	-	0.3500	99	menF	BSU30830	0.8457853198
200nt	984466	984665	reverse	BSU09120	yhcK	-1169	-40.11	-	0.2800	124.3939972	-	0.3750	99	cspB	BSU09100	0.8457109332
200nt	3464551	3464750	reverse	BSU33780	sdpI	-1784	-40.60	-	0.3100	125.1740036	-	0.4450	1399	opuBA	BSU33730	0.8446103930
200nt	3684271	3684470	reverse	BSU35780	lytD	-456	-43.90	-	0.3450	125.0149994	-	0.3500	3399	tagD	BSU35740	0.8443253636
200nt	2770775	2770974	reverse	BSU27160	cypB	-3016	-41.20	-	0.3550	126.4410019	-	0.4050	2199	yrhP	BSU27100	0.8441781402
200nt	1666288	1666487	reverse	BSU15960	ylqB	-4779	-56.60	-	0.4500	123.3949966	-	0.3650	10599	rpmB	BSU15820	0.8431691527
200nt	2343862	2344061	reverse	BSU22330	ypoC	-303	-43.50	-	0.3800	126.2779999	-	0.4850	3199	yppC	BSU22300	0.8418609500
200nt	3158854	3159053	reverse	BSU30890	ytxO	-305	-56.71	-	0.4200	122.1419983	-	0.3650	3399	ytdA	BSU30850	0.8375009298
200nt	4129648	4129847	reverse	BSU40200	yydD	-831	-66.70	-	0.5050	120.9280014	-	0.3350	2099	yydF	BSU40180	0.8359215260
200nt	4134175	4134374	reverse	BSU40230	yydA	-162	-30.70	-	0.2450	126.6350021	-	0.3400	99	yydB	BSU40220	0.8344783783
200nt	245362	245561	reverse	BSU02340	gltP	-8057	-57.10	-	0.4500	122.8600006	-	0.3500	1799	ybfI	BSU02220	0.8332447410
200nt	3996389	3996588	reverse	BSU38970	yxjF	-4051	-40.50	-	0.3750	126.9710007	-	0.4250	4899	yxkA	BSU38870	0.8313001394
200nt	2739837	2740036	reverse	BSU26830	yrpE	-1421	-64.40	-	0.4950	121.3069992	-	0.3300	3399	aadK	BSU26790	0.8294041157
200nt	1903178	1903377	reverse	BSU17690	yncM	-234	-35.20	-	0.2700	125.4830017	-	0.4250	1899	cotU	BSU17670	0.8289884329
200nt	933665	933864	reverse	BSU08620	yfhP	-693	-62.04	-	0.4700	121.3809967	-	0.3750	5499	sspK	BSU08550	0.8283531070
200nt	2054430	2054629	reverse	BSU18840	xynA	-70	-52.10	-	0.3850	122.5159988	-	0.3450	599	pps	BSU18830	0.8281581998
200nt	737824	738023	reverse	BSU06740	yefB	-1072	-69.00	-	0.5350	120.7480011	-	0.3100	5099	yerO	BSU06700	0.8277365565
200nt	198226	198425	reverse	BSU01800	alkA	-4222	-30.81	-	0.3750	130.7449951	-	0.5150	4299	ybbK	BSU01720	0.8267450333
200nt	3604668	3604867	reverse	BSU35100	yvlD	-1995	-39.30	-	0.2750	123.9219971	-	0.3550	199	yvmC	BSU35070	0.8267388940
200nt	3098465	3098664	reverse	BSU30310	ytwF	-3637	-48.30	-	0.3850	123.9609985	-	0.4100	1999	ytaP	BSU30250	0.8252500296
200nt	419514	419713	reverse	BSU03690	yczF	-150	-54.00	-	0.4400	123.4860001	-	0.3300	1899	dtpT	BSU03670	0.8242135644
200nt	2221888	2222087	reverse	BSU21080	yonI	-6769	-36.60	-	0.3550	127.6279984	-	0.3150	599	yonR	BSU21020	0.8236665726
200nt	2434023	2434222	reverse	BSU23340	ypuB	-21	-56.65	-	0.4750	123.5520020	-	0.3050	599	ypzJ	BSU23328	0.8226841688
200nt	3241980	3242179	reverse	BSU31590	yufS	-4073	-57.60	-	0.4600	122.6460037	-	0.3500	5099	yufK	BSU31510	0.8222519755
200nt	1700752	1700951	reverse	BSU17060	ymzD	-79767	-52.70	-	0.4300	123.5100021	-	0.3150	29199	ylqB	BSU15960	0.8189544678
200nt	2709820	2710019	reverse	BSU26490	yrkJ	-83	-48.40	-	0.3600	122.8440018	-	0.3550	499	yrkK	BSU26480	0.8178487420
200nt	153939	154138	reverse	BSU01550	gerD	-4477	-50.70	-	0.4100	123.6019974	-	0.3700	108899	abrB	BSU00370	0.8176639676
200nt	3239080	3239279	reverse	BSU31590	yufS	-6973	-58.30	-	0.4150	120.6800003	-	0.3750	2199	yufK	BSU31510	0.8171101809
200nt	3467327	3467526	reverse	BSU33800	opuCD	-120	-32.10	-	0.2350	125.1719971	-	0.3550	99	sdpR	BSU33790	0.8168275952
200nt	543114	543313	reverse	BSU05000	yddK	-2953	-43.99	-	0.3600	124.5599976	-	0.4700	11699	immR	BSU04820	0.8156080246
200nt	3108826	3109025	reverse	BSU30370	bceB	-472	-37.52	-	0.3400	126.4919968	-	0.4900	499	yttB	BSU30350	0.8153505921
200nt	3334388	3334587	reverse	BSU32470	pucE	-1264	-54.50	-	0.4850	124.4670029	-	0.3050	5899	pucH	BSU32410	0.8130649924
200nt	3684371	3684570	reverse	BSU35780	lytD	-356	-42.00	-	0.3250	124.1009979	-	0.3650	3499	tagD	BSU35740	0.8130072355
200nt	881307	881506	reverse	BSU08120	yfjF	-4438	-52.36	-	0.4850	125.3059998	-	0.3900	9099	yfjQ	BSU08000	0.8121696115
200nt	2926798	2926997	reverse	BSU28630	pheT	-111	-60.46	-	0.4950	122.2929993	-	0.3200	999	yshA	BSU28610	0.8096395731
200nt	4066309	4066508	reverse	BSU39600	yxeC	-199	-55.60	-	0.4600	123.0559998	-	0.3600	899	yxeF	BSU39570	0.8090547919
200nt	3688648	3688847	reverse	BSU35830	ywtG	-3786	-49.93	-	0.3300	120.9260025	-	0.3550	199	yvyI	BSU35790	0.8084035516
200nt	1668788	1668987	reverse	BSU15960	ylqB	-2279	-51.60	-	0.4050	122.7809982	-	0.4150	13099	rpmB	BSU15820	0.8080439568
200nt	3334288	3334487	reverse	BSU32470	pucE	-1364	-62.06	-	0.4700	120.7320023	-	0.2950	5799	pucH	BSU32410	0.8073683381
200nt	3723732	3723931	reverse	BSU36170	ywqL	-589	-55.50	-	0.4550	122.8669968	-	0.3300	499	ywqN	BSU36150	0.8070057034
200nt	899787	899986	reverse	BSU08340	padR	-9312	-54.60	-	0.3900	121.0039978	-	0.3550	10199	yfjA	BSU08170	0.8061554432
200nt	2813434	2813633	reverse	BSU27540	yrvM	-110	-48.20	-	0.3850	123.3710022	-	0.4200	1399	cymR	BSU27520	0.8043378592
200nt	3458216	3458415	reverse	BSU33700	opuBD	-1491	-44.09	-	0.3500	123.8420029	-	0.4200	1799	yvzC	BSU33650	0.8041363358
200nt	3918262	3918461	reverse	BSU38190	galT	-732	-37.50	-	0.2900	124.4739990	-	0.3100	99	qoxA	BSU38170	0.8040998578
200nt	2576788	2576987	reverse	BSU24950	pstBB	-323	-53.30	-	0.3700	120.7910004	-	0.4100	699	yqgL	BSU24920	0.8040402532
200nt	2316211	2316410	reverse	BSU22040	ypbQ	-136	-50.40	-	0.3350	120.7730026	-	0.3850	199	ypbR	BSU22030	0.8038558364
200nt	2249344	2249543	reverse	BSU21440	bdbB	-15782	-43.38	-	0.3500	124.1179962	-	0.4650	1899	youB	BSU21329	0.8037469983
200nt	2333112	2333311	reverse	BSU22210	yprB	-113	-43.21	-	0.3300	123.4840012	-	0.4200	199	cotD	BSU22200	0.8028732538
200nt	2116670	2116869	reverse	BSU19420	yojK	-282	-43.45	-	0.3300	123.3679962	-	0.3150	99	cwlS	BSU19410	0.8022140265
200nt	3941242	3941441	reverse	BSU38430	gspA	-3219	-42.84	-	0.4300	126.9950027	-	0.4600	1699	ywbA	BSU38390	0.8019362092
200nt	1248722	1248921	reverse	BSU11740	cotZ	-621	-64.39	-	0.5300	121.6650009	-	0.3550	8699	yjbP	BSU11630	0.8019282818
200nt	3014345	3014544	reverse	BSU29460	moaB	-70	-39.90	-	0.2550	122.2269974	-	0.3300	99	argG	BSU29450	0.8009542227
200nt	2542225	2542424	reverse	BSU24510	yqhO	-115	-57.14	-	0.4800	122.8730011	-	0.4100	1499	yqhR	BSU24480	0.8008475304
200nt	2096086	2096285	reverse	BSU19230	yocJ	-165	-39.50	-	0.2800	123.2190018	-	0.4150	399	yocI	BSU19220	0.8003621101
200nt	4107928	4108127	reverse	BSU40010	yxaD	-1158	-51.01	-	0.4150	123.1330032	-	0.3800	99	yxaF	BSU39990	0.8002312183
200nt	2659771	2659970	reverse	BSU25880	yqxJ	-3681	-49.60	-	0.3800	122.5149994	-	0.3600	1099	yqcI	BSU25820	0.8001416922

Table B.11: Top Classification Hits in *E. coli*. Top 50 hits of the forward and reverse strands of the *E. coli* intergenic regions using 50 nt-overlap 100 nt window and under the LMFEGCRND model. The ranking of each hit is denoted in column R. Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the multinomial regression likelihood of being a riboswitch under the LMFEGCRND model. Positions are according to gb|U00096.2 version of *E. coli* and not gb|U00096.3 version.

R	Start	End	Strand	Upstream Operon	Dist. to Upstream	Uracil	Dist. to Downstream	Downstream Operon	Probability
1	384006	384105	forward	insC-1,insCD-1,insD-1	-2154	0.52	402	tauA,tauB,tauC,tauD	0.942
2	237185	237284	forward	aspV	-129	0.47	102	yafT	0.934
3	2777119	2777218	forward	yfjX,yfjY,yfjZ,ypjF,ypjJ	-1266	0.38	7252	ygaQ_1,ygaQ_2	0.925
4	2304856	2304955	forward	eco	-2392	0.45	6202	micF	0.923
5	83968	84067	forward	setA,sgrS,sgrT	-5120	0.49	352	leuO	0.92
6	2902496	2902595	reverse	queE	-224	0.48	4249	ygcW	0.918
7	294815	294914	forward	yagJ	-3311	0.43	7352	yagU	0.914
8	4554566	4554665	forward	uxuR	-1145	0.48	402	iraD	0.913
9	405479	405578	forward	yaiI	16	0.45	102	aroL,aroM,yaiA	0.908
10	4570237	4570336	forward	yjiS	-250	0.38	152	yjiT	0.906
11	754000	754099	forward	nei,ybgI,ybgJ,ybgK,ybgL	-8002	0.4	352	sdhA[1]	0.905
12	2054653	2054752	reverse	asnW	-1349	0.44	3349	yeeL_1,yeeL_2	0.905
13	2202241	2202340	reverse	yehS	-7458	0.44	10099	mrp	0.903
14	330995	331094	forward	betT	-226	0.52	552	yahA	0.9
15	3183291	3183390	reverse	glgS	-6421	0.44	599	ribB,sroG	0.9
16	384056	384155	forward	insC-1,insCD-1,insD-1	-2204	0.43	352	tauA,tauB,tauC,tauD	0.898
17	4570187	4570286	forward	yjiS	-200	0.42	202	yjiT	0.898
18	557285	557384	forward	cysS	-2017	0.35	102	sfmA	0.894
19	3190062	3190161	reverse	sibD	-2632	0.37	149	glgS	0.893
20	1543575	1543674	forward	nhoA	-10633	0.45	1802	fdnG,fdnH,fdnI	0.89
21	2190295	2190394	reverse	yehE	-193	0.46	99	yehA,yehB,yehC,yehD	0.89
22	3181507	3181606	reverse	ribB,sroG	-279	0.43	1099	ygiD	0.89
23	3834703	3834802	reverse	nlpA	-2446	0.37	799	yicI,yicJ	0.889
24	1753166	1753265	reverse	ynhG	-2530	0.45	49	ydhZ	0.888
25	819916	820015	forward	ybhL	-56	0.41	52	ybhM	0.887
26	2901746	2901845	reverse	queE	-974	0.51	3499	ygcW	0.887
27	651208	651307	forward	ybdR	-8610	0.49	202	dpiA,dpiB	0.886
28	584973	585072	forward	appY	-1271	0.57	9802	cusA,cusB,cusC,cusF	0.886
29	2362398	2362497	reverse	ais	-593	0.48	149	yfaZ	0.886
30	1596214	1596313	forward	osmC	-41085	0.39	3252	lsrA,lsrB,lsrC,lsrD,lsrF,lsrG,tam	0.882
31	522335	522434	forward	ybbA,ybbP	-232	0.39	102	rhsD,ybbC,ybbD,ylbH	0.881
32	3490420	3490519	reverse	php,yhfS,yhfT,yhfU,yhfW,yhfX	-12488	0.39	149	ppiA	0.88
33	1986023	1986122	reverse	yecH	-1203	0.45	49	isrB	0.879
34	3217299	3217398	reverse	ygjH	-1589	0.42	249	aer	0.878
35	2714626	2714725	forward	eamB	-545	0.45	102	ung	0.877
36	3984255	3984354	forward	aslB	-1990	0.42	152	glmZ	0.877
37	2651611	2651710	reverse	sseB	-519	0.5	99	C0614	0.877
38	4516300	4516399	forward	insO-2,yjhV,yjhW	-8095	0.39	202	insA-7	0.876
39	3886253	3886352	reverse	purP	-6993	0.31	4549	dnaA,dnaN,recF	0.876
40	1577414	1577513	forward	osmC	-22285	0.41	22052	lsrA,lsrB,lsrC,lsrD,lsrF,lsrG,tam	0.875
41	776349	776448	reverse	zitB	-6707	0.39	11299	mngR	0.874
42	1543625	1543724	forward	nhoA	-10683	0.46	1752	fdnG,fdnH,fdnI	0.871
43	29201	29300	forward	dapB	43	0.49	402	carA,carB	0.871
44	1542975	1543074	forward	nhoA	-10033	0.39	2402	fdnG,fdnH,fdnI	0.871
45	3768179	3768278	reverse	yibH,yibI	-38	0.41	8249	yibF	0.871
46	3631114	3631213	forward	yhhI	-7528	0.4	1702	yhiM	0.87
47	2166486	2166585	forward	cyaR	-1213	0.41	202	yegS	0.87
48	4578972	4579071	forward	symR	-989	0.38	5952	mrr	0.869
49	522235	522334	forward	ybbA,ybbP	-132	0.39	202	rhsD,ybbC,ybbD,ylbH	0.869
50	2383795	2383894	reverse	yfbN	-1888	0.48	99	yfbK	0.869
51	4577258	4577357	forward	yjiV	-2331	0.36	552	symR	0.868
52	153855	153954	forward	yadD	-5936	0.37	8202	hrpB	0.868
53	3665704	3665803	reverse	yhjA	-61	0.42	149	gadA,gadW,gadX	0.868
54	3651672	3651771	reverse	hdeA,hdeB,yhiD	-1557	0.45	499	insH-11	0.868
55	4619642	4619741	forward	deoA,deoB,deoC,deoD	32	0.38	102	yjjJ	0.867
56	1645875	1645974	reverse	ynfP	-4828	0.45	49	dicC,ydfW,ydfX	0.867
57	4554516	4554615	forward	uxuR	-1095	0.42	452	iraD	0.866
58	4539860	4539959	forward	fimB	-229	0.41	152	fimE	0.866
59	1588711	1588810	reverse	hipA,hipB	-118	0.4	199	yneL	0.866
60	3181557	3181656	reverse	ribB,sroG	-229	0.33	1149	ygiD	0.865
61	269657	269756	reverse	insH-1	-3619	0.45	299	perR	0.864
62	3925028	3925127	forward	cbrB,cbrC	-28347	0.47	102	asnA	0.863
63	4538580	4538679	forward	yjhR	-4477	0.43	352	fimB	0.863
64	4077095	4077194	reverse	fdhE,fdoG,fdoH,fdoI	-1178	0.39	5549	yihS,yihT,yihU	0.863
65	1210379	1210478	reverse	iraM	-475	0.42	1549	stfE,tfaE	0.862
66	4213351	4213450	forward	metA	-70	0.37	102	aceA,aceB,aceK	0.861
67	578853	578952	forward	essD,rrrD,rzpD	-1013	0.49	202	ybcW	0.861
68	3755951	3756050	reverse	selA,selB	-40	0.33	149	yiaY	0.861
69	924768	924867	forward	clpA	44	0.36	7002	lrp	0.86
70	2520650	2520749	reverse	xapA,xapB	-52	0.42	199	xapR	0.86
71	582454	582553	forward	tfaX	-122	0.44	402	appY	0.859
72	1049984	1050083	forward	insA-4,insAB-4,insB-4	-182	0.38	652	cspG	0.859
73	3767704	3767803	forward	yibG	-994	0.36	2552	mtlA,mtlD,mtlR	0.859
74	2383745	2383844	reverse	yfbN	-1938	0.45	49	yfbK	0.859

75	1005025	1005124	forward	pyrD	25	0.46	102	zapC	0.858
76	157105	157204	forward	yadD	-9186	0.34	4952	hrpB	0.858
77	1669567	1669666	reverse	mdtI,mdtJ	-1228	0.46	1999	ynfL	0.858
78	3800263	3800362	forward	rfaD,waaC,waaF,waaL	-3984	0.35	6252	coaD,waaA	0.856
79	2784960	2785059	reverse	ygaU	-9350	0.41	1149	ileY	0.856
80	2468130	2468229	reverse	yfdK,yfdL,yfdM,yfdN,yfdO	-920	0.49	5149	mlaA	0.855
81	2991883	2991982	reverse	ygeK,ygeL	-550	0.39	3549	yqeK	0.855
82	2859337	2859436	reverse	nlpD,rpoS	-5195	0.42	99	ygbI	0.855
83	3490370	3490469	reverse	php,yhfS,yhfT,yhfU,yhfW,yhfX	-12538	0.32	99	ppiA	0.854
84	4220501	4220600	forward	aceA,aceB,aceK	-2097	0.37	1302	metH	0.853
85	715871	715970	reverse	potE,speF	-249	0.37	99	ybfG,ybfH	0.853
86	2461957	2462056	reverse	mlaA	-268	0.45	3049	yfcZ	0.852
87	2876502	2876601	reverse	cas1,cas2,casA,casB,casC,casD,casE	-40	0.34	2199	cysC,cysD,cysN	0.851
88	4084875	4084974	forward	fdhD	46	0.42	102	yiiG	0.85
89	4535630	4535729	forward	yjhR	-1527	0.31	3302	fimB	0.849
90	1635392	1635491	reverse	gnsB	-192	0.42	1049	nohA,tfaQ,ydfN	0.849
91	3497220	3497319	reverse	php,yhfS,yhfT,yhfU,yhfW,yhfX	-5688	0.33	6949	ppiA	0.849
92	3266938	3267037	reverse	garK,garL,garP,garR,rnpB	-1251	0.35	1899	tdcA,tdcB,tdcC,tdcD,tdcE,tdcF,tdcG	0.848
93	2267568	2267667	reverse	yejG	-8298	0.34	4299	yeiW	0.848
94	3580990	3581089	reverse	ggt	-2065	0.4	1999	ryhB	0.848
95	1811219	1811318	reverse	cedA	-177	0.5	99	ydjO	0.845
96	583210	583309	reverse	envY,ompT	-644	0.4	1299	ybcY	0.845
97	3576850	3576949	reverse	yhhW	-74	0.33	149	gntK,gntR,gntU	0.844
98	3266488	3266587	reverse	garK,garL,garP,garR,rnpB	-1701	0.43	1449	tdcA,tdcB,tdcC,tdcD,tdcE,tdcF,tdcG	0.843
99	2055603	2055702	reverse	asnW	-399	0.33	4299	yeeL_1,yeeL_2	0.843
100	2739273	2739372	reverse	rimM,rplS,rpsP,trmD	-2883	0.3	149	aroF,tyrA	0.842

[1] Table B.11: Complete list of genes in this operon is sdhA,sdhB,sdhC,sdhD,sucA,sucB,sucC,sucD.

Table B.12: Top Classification Hits in *E. coli* Uracil-comp. Constrained. Top 50 hits of the forward and reverse strands of the *E. coli* intergenic regions that have Uracil composition within the range of known riboswitches in *E. coli* (between 0.23 and 0.34). 50 nt-overlap 100 nt window used. The ranking of each hit is denoted in column R. Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the multinomial regression likelihood of being a riboswitch under the LMFEGCRND model. Positions are according to gb|U00096.2 version of *E. coli* and not gb|U00096.3 version.

R	Start	End	Strand	Upstream Operon	Dist. to Upstream	Uracil	Dist. to Downstream	Downstream Operon	Probability
1	3886253	3886352	reverse	purP	-6993	0.31	4549	dnaA,dnaN,recF	0.876
2	3181557	3181656	reverse	ribB,sroG	-229	0.33	1149	ygiD	0.865
3	3755951	3756050	reverse	selA,selB	-40	0.33	149	yiaY	0.861
4	3490370	3490469	reverse	php,yhfS,yhfT,yhfU,yhfW,yhfX	-12538	0.32	99	ppiA	0.854
5	4535630	4535729	forward	yjhR	-1527	0.31	3302	fimB	0.849
6	3497220	3497319	reverse	php,yhfS,yhfT,yhfU,yhfW,yhfX	-5688	0.33	6949	ppiA	0.849
7	3576850	3576949	reverse	yhhW	-74	0.33	149	gntK,gntR,gntU	0.844
8	2055603	2055702	reverse	asnW	-399	0.33	4299	yeeL_1,yeeL_2	0.843
9	2739273	2739372	reverse	rimM,rplS,rpsP,trmD	-2883	0.3	149	aroF,tyrA	0.842
10	2698570	2698669	reverse	acpS,era,pdxJ,recO,rnc	-21	0.31	399	shoB	0.84
11	3945101	3945200	reverse	hdfR	-1	0.3	5799	hsrA,yieP	0.835
12	2739223	2739322	reverse	rimM,rplS,rpsP,trmD	-2933	0.31	99	aroF,tyrA	0.832
13	3453521	3453620	reverse	bfd,bfr	-10701	0.32	149	gspA,gspB	0.825
14	4274265	4274364	reverse	soxS	-769	0.32	1249	yjcB	0.822
15	1467282	1467381	forward	ydbA_1	-1259	0.32	52	insI-2	0.82
16	3116880	3116979	forward	pheV	-8368	0.33	2452	C0719	0.819
17	790896	790995	forward	aroG	-4939	0.26	3052	acrZ	0.817
18	2777069	2777168	forward	yfjX,yfjY,yfjZ,ypjF,ypjJ	-1216	0.33	7302	ygaQ_1,ygaQ_2	0.817
19	3189691	3189790	reverse	glgS	-21	0.31	6999	ribB,sroG	0.817
20	1694096	1694195	reverse	uidR	-341	0.33	49	uidA,uidB,uidC	0.815
21	2823699	2823798	reverse	norR	-5049	0.31	149	mltB	0.81
22	1805258	1805357	reverse	yniB	-1414	0.32	1199	ydiY	0.809
23	3382541	3382640	reverse	yhcO	-1289	0.33	299	mdh	0.809
24	1868697	1868796	reverse	yoaI	-3356	0.29	4249	mipA	0.808
25	2386449	2386548	reverse	nuoA[1]	-1572	0.31	49	yfbN	0.807
26	3617057	3617156	reverse	rbbA,yhhJ,yhiI	-6596	0.3	7349	yhhS	0.806
27	2815604	2815703	forward	micA	-2660	0.29	8202	gutM,gutQ,srlA,srlB,srlD,srlE,srlR	0.805
28	2943865	2943964	reverse	mltA	-189	0.31	49	tcdA	0.804
29	1983499	1983598	forward	uspC	-5245	0.31	1402	ftnB	0.803
30	150155	150254	forward	yadD	-2236	0.32	11902	hrpB	0.801
31	2553093	2553192	forward	amiA,hemF	-898	0.33	3652	intZ	0.8
32	2876452	2876551	reverse	cas1,cas2,casA,casB,casC,casD,casE	-90	0.31	2149	cysC,cysD,cysN	0.8
33	1217949	1218048	reverse	ymgD,ymgG	-3530	0.33	3299	bluF	0.8
34	3346238	3346337	reverse	elbB,mtgA	-816	0.33	8199	mlaB,mlaC,mlaD,mlaE,mlaF	0.798
35	1174989	1175088	reverse	ycfZ,ymfA	-4664	0.3	649	ycfT	0.798
36	4298873	4298972	forward	gltP	-5007	0.26	12452	rpiB	0.797
37	585173	585272	forward	appY	-1471	0.29	9602	cusA,cusB,cusC,cusF	0.796
38	1397665	1397764	forward	ynaJ	-1970	0.32	5052	abgR	0.795
39	4492546	4492645	forward	lptF,lptG	-6074	0.29	52	idnK	0.795
40	655892	655991	reverse	crcB	-837	0.32	749	dcuC	0.795
41	266028	266127	reverse	yafZ,ykfA	-331	0.28	299	yafW[2]	0.795
42	2033559	2033658	reverse	yedV,yedW	-1210	0.32	2099	yedJ,yedR	0.794
43	3963904	3964003	reverse	aslA	-18422	0.32	299	rhlB	0.794
44	1463517	1463616	reverse	insC-2,insCD-2,insD-2	-2379	0.33	11899	paaZ	0.794
45	2650116	2650215	forward	xseA	-16443	0.31	352	sseA	0.793
46	582654	582753	forward	tfaX	-322	0.3	202	appY	0.793
47	4076645	4076744	reverse	fdhE,fdoG,fdoH,fdoI	-1628	0.29	5099	yihS,yihT,yihU	0.793
48	2033409	2033508	reverse	yedV,yedW	-1360	0.32	1949	yedJ,yedR	0.793
49	1762598	1762697	forward	lpp	-6868	0.27	4452	ydiK	0.792
50	4501881	4501980	forward	yjgZ	-2220	0.32	152	yjhB,yjhC	0.792
51	1676251	1676350	forward	tqsA	-3231	0.33	152	ydgH	0.791
52	2166386	2166485	forward	cyaR	-1113	0.32	302	yegS	0.789
53	2429322	2429421	forward	folX,yfcH	-8650	0.28	6602	flk	0.789
54	4083889	4083988	forward	yiiF	-5848	0.29	102	fdhD	0.789
55	2491327	2491426	reverse	yfdX	-413	0.32	99	frc	0.789
56	1165025	1165124	reverse	comR	-2349	0.31	4299	fhuE	0.789
57	1933126	1933225	forward	purT	-2994	0.3	1502	yebK	0.787
58	1204407	1204506	reverse	stfE,tfaE	-3284	0.25	2299	ymfK	0.787
59	1493112	1493211	forward	trg	-929	0.32	152	ydcJ	0.786
60	1714050	1714149	forward	gstA	-995	0.31	3802	slyB	0.786
61	1397615	1397714	forward	ynaJ	-1920	0.32	5102	abgR	0.784
62	5034	5133	forward	thrA,thrB,thrC,thrL	35	0.27	152	yaaX	0.784
63	4547775	4547874	reverse	gntP	-152	0.25	10299	nanC,nanM	0.784
64	2257385	2257484	forward	yeiL	-3300	0.3	4452	setB	0.783
65	660791	660890	forward	tatE	-2369	0.29	13402	ybeL	0.781
66	3945051	3945150	reverse	hdfR	-51	0.29	5749	hsrA,yieP	0.781
67	2238382	2238481	forward	preA,preT	-3811	0.33	3502	yeiG	0.779
68	925014	925113	reverse	serW	-44	0.31	3249	cspD	0.779
69	13587	13686	reverse	hokC,mokC	-3115	0.3	1849	yaaI	0.779
70	2734984	2735083	reverse	aroF,tyrA	-1937	0.33	999	rluD,yfiH	0.778
71	2880686	2880785	forward	iap	-4997	0.32	9502	queD	0.777

72	1078128	1078227	forward	rutR	-3976	0.29	352	putP	0.776
73	187962	188061	forward	cdaR	-4293	0.32	1702	rpsB,tff,tsf	0.776
74	497037	497136	reverse	aes	-1152	0.25	7049	priC,ybaM	0.776
75	3181662	3181761	forward	zupT	-268	0.3	1152	yqiC	0.775
76	593123	593222	forward	appY	-9421	0.29	1652	cusA,cusB,cusC,cusF	0.775
77	1250189	1250288	forward	ycgY	-5317	0.33	52	dhaR	0.775
78	3597882	3597981	reverse	rpoH	-21	0.32	249	livJ	0.775
79	117883	117982	forward	guaC	-3347	0.33	802	ampD,ampE	0.774
80	1073265	1073364	forward	ymdF	-5739	0.31	152	rutR	0.774
81	3416188	3416287	reverse	alaU,ileU,rrfD,rrfF,rrlD,rrsD,thrV	-5208	0.28	4749	envR	0.774
82	4238098	4238197	forward	yjbE,yjbF,yjbG,yjbH	-296	0.33	202	psiE	0.773
83	3313859	3313958	forward	psrO	-4390	0.32	2752	argG	0.773
84	238253	238352	reverse	yafU	-444	0.29	2299	rnhA	0.773
85	2627711	2627810	reverse	guaA,guaB	-1220	0.31	799	yfgF	0.773
86	4156263	4156362	forward	argB,argC,argH	32	0.3	202	oxyR	0.772
87	58274	58373	forward	djlA	-46	0.32	152	yabP,yabQ	0.772
88	3108528	3108627	reverse	yghD,yghE	-35	0.33	1399	speC	0.772
89	2750731	2750830	reverse	ratA,ratB	-1250	0.31	2049	grpE	0.772
90	454057	454156	forward	bolA	5	0.32	252	tig	0.771
91	573621	573720	forward	ybcQ	-10	0.26	2952	essD,rrrD,rzpD	0.771
92	905496	905595	forward	amiD,ybjQ	-481	0.31	10152	ybjD	0.771
93	2407114	2407213	reverse	yfbS	-379	0.32	2499	lrhA	0.771
94	1733426	1733525	reverse	ydhP	-670	0.31	1349	grxD	0.771
95	604509	604608	forward	pheP	-1902	0.33	2502	hokE	0.77
96	1431895	1431994	forward	lomR_2,stfR,tfaR	-836	0.31	3202	micC	0.77
97	1395289	1395388	forward	insH-4	-124	0.28	52	ynaJ	0.769
98	253317	253416	forward	dinB,yafN,yafO,yafP	-107	0.3	102	prfH,ykfJ	0.768
99	573571	573670	forward	ybcQ	40	0.28	3002	essD,rrrD,rzpD	0.768
100	3986826	3986925	forward	glmZ	-2151	0.32	2302	cyaA	0.768

[1] Table B.12: Complete list of genes in this operon is nuoA,nuoB,nuoC,nuoE,nuoF,nuoG,nuoH, nuoI,nuoJ,nuoK,nuoL,nuoM,nuoN.

[2] Table B.12: Complete list of genes in this operon is yafW,yafX,yafY,ykfB,ykfF,ykfG,ykfH,ykfI.

B.5 Positive-Control-Set Sequence Segments

B.5.1 Training Set

```
 >Alpha Operon:  E. coli, Alteration:  Unique:  Slow/Fast + Complex Regulatory Mechanism.
UGUGCGUUUCCAUUUGAGUAUCCUGAAAACGGGCUUUUCAGCAUGGAACGUACAUAUUAAAUAGUAGGAGUGCAUAGUGGCCCGUAUAGCAGGCAUUAACAUUCCUGA
((((((.(((((........[[[[....[[[[.....)))))))))))........................]]]].....]]]]..........
>3_166_234 Cobalamin riboswitch:  E. coli, Alteration:  Normal.  Expression Platform, Only.
GUCGCAUCUGGUUCUCAUCAUCGCGUAAUAUUGAUGAAACCUGCGGCAUCCUUCUUCUAUUGUGGAUGC
((((((.................................))))))......................
..............((((((...........))))))........((((((............))))))
>Cobalamin riboswitch:  Bradyrhizobium japonicum.  Alteration:  Normal.  Expression Platform, Only.[1]
GUCACACGCGAAGAUGUCGGUCGGGGAGUACAGGCAUUAGCUUCACCGGAGCAAUCGAUUGCUCCGCCGUAAAGCCUCGUUCGCUGUGACGUGCCACUGACGUCAUGCCGAGGUU
((((((.(((((.........(((((..(((.(((....)))....((((((((....))))))))..)))...))))))))))))))))........(((.((.....)).)))
.....(((.(((((((((.(((.....).)).)))))...))))..((((((((....)))))))).)))..(((((((...((.((((((((.......)))))))))))))))))
>Fluoride riboswitch crcB motif.:  Pseudomonas syringae.  Alteration:  Normal.
GAUCGGCGCAUUGGAGAUGGCAUUCCUCCAUUAACAAACCGCUGCGCCCGUAGCAGCUGAUGAUGCCUACAGAAACCUG
...........[[[[[..((((((.]]]]]..........(((((.......)))))....))))))............
>Fluoride rigoswitch:  Thermotoga petrophila.  Alteration:  Normal.
GGGCGAUGAGGCCCGCCCAAACUGCCCUGAAAAGGGCUGAUGGCCUCUACUGGCUUGAUCAGUAGAGGCCA
.((((........))))......(((((....)))))...(((((((((((((.....)))))))))))))
.((((.[[[[[[]))))......(((((....)))))..]]]]]]......................
>FMN riboswitch:  Fusobacterium nucleatum.  Alteration:  Normal.
UCUUCGGGGCAGGGUGAAAUUCCCGACCGGUGGUAUAGUCCACGAAAGUAUUUGCUUUGAUUUGGUGAAAUUCCAAAACCGACAGUAGAGUCUGGAUGAGAGAAGAAAAGAAAUU
..((((((...(((.......))).....((((......))))..((((....))))...(((((.......))))).....(((......))).....................
(((((......(((.......))).....((((......))))..((((....))))...(((((.......))))).....(((......))).......))))).........
UAAGUUUUUUAACUUGUUUUCUACAUUUUAGUAAUCUUACCCGAAUUCUAUAAUUCGG
.......................................))))))............
.......................................(((((((....)))))))
>FMN riboswitch ribB leader:  E. coli.  Alteration:  Normal.
GCUUAUUCUCAGGGCGGGGCGAAAUUCCCCACCGGCGGUAAAUCAACUCAGUUGAAAGCCCGCGAGCGCUUUGGGUGCGAACUCAAAGGACAGCAGAUCCGGUGUAAUUCCGGGG
..((((((((.....((((.......))))....((((....(((((...)))))....)))).....((((((((....))))))))..........((((.......)))).. 
CCGACGGUUAGAGUCCGGAUGGGAGAGAGUAACGAUUCUGUCGGGCAUGGACCCGCUCACGUUAUUUUGGCUAUAUGCCGCCACUCCUAAGACUGCCCUGAUUCUGGUAACCAUA
....(((.......))).......))))))))..............................................................((((((((((((.....(
(((((...)))))))....))))).))))))))...
>c-di-GMP riboswitch GEMM motif:  Geobacter sulfurreducens.  Alteration:  Normal.
CUAAACCAUCCGCGAGGAUGGGGCGGAAAGCCCACAGGGUCUCACGAAGACAGCCGGGUUGCCGAACUAUCACACCACGAUAGGGCGGCGGCCCGGCU
((...((((((....)))))))...))..((((..(.((((((.....)))))..))))))...................................
>Glycine riboswitch:  aptamer 2 + 10 nt downstream:  Fusobacterium nucleatum.  Alteration:  Normal.[2]
CUCUGGAGAGCUUAUCUAAGAGAUAACACCGAAGGAGCAAAGCUAAUUUUAGCCUAAACUCUCAGGUAAAAGGACGGAGAUAAUUGUGC
(((((......(((((.....))))).(((...((((....((((....))))).....))))..)))......)))))..........
>Lysine riboswitch:  Thermotoga maritima.  Alteration:  Normal.
GACCCGACGGAGGCGCGCCCGAGAUGAGUAGGCUGUCCCAUCAGGGGAGGAAUCGGGGACGGCUGAAAGGCGAGGGCGCCGAAGGGUGCAGAGUUCCUCCCGCUCUGCAUGCCUG
.............((((((((((((.(((((((((((..(.[[[[[[...)))))))))))))))))))))))))))))...((((((((((((.]]]]]])))))))))).)))(
(((((((((....((((((((((((.(((((((((((..(.[[[[[[...)))))))))))))))))))))))))))))...((((((((((((.]]]]]])))))))))).)))(
GGGGUAUGGGGAAUACCCAUACCACUGUCACGGAGGUCUCUCCGUGGAGAGCCGAUCGGGUCUGGAAUCAGAAAAAGAUUUCAGACCUGAUUUGGCGUCUCUUCGGGGAGCGAAG
(((((((((((.....)))))))))))...((((((....))))))((((((((((((((((((((((((.......))))))))))))).))))).))))).............
(((((((((((.....))))))))))).((((((((....))))))))....))).))))))).............................((((((((((((.....)))))
AGACGC
......
))))))
>Magnesium riboswitch mgtA: Salmonella enterica serovar Typhimurium.  Alteration:  Normal
CUUACCGGAGGCGACAUGGACCCUGAACCCACCCCUCUCCCGCGAUGGAGAAUUUUCCUUUUCCGGUAAGCCUGCCUCUGCUGUCUUACCGGUGUGUAAGACAGUGACACAAUAA
.............................................................((((((((...((....))...)))))))).....(((((((.(((.......
(((((((((((.....(((.........)))....(((((......)))))........)))))))))))).........(((((((((((......))))))))))).........
CGUCCCUGUUUUUAUUUAAACAUUGCUCAUCGGGCAAGGCUUUGCCGUGCCUGAAGA
.))).))))))))....(((((.((((((...)))))).).)))).............
......(((((......)))))....((.(((((((.(((...))).)))))))).))
>Magnesium riboswitch mgtA: E. coli.  Alteration:  Normal.
CUUACCGGAGGUUAUAUGGAACCUGAUCCCACGCCUCUCCCUCGACGGAGAUUAAAACUUUUCCGGUAAGCCCGUCUUUUCACGGCGUUACCGGAUGCGUAAGGCCGUGA
(((((((((((((....(((......)))......(((((......)))))....)))))..))))))))).........(((((((.((((.......)))).)))))))
...............................................................((((((((..((((......))))..)))))))))..............
>Moco riboswitch:  E. coli.  Alteration:  Normal.
ACACUCUAGCCUCUGCACCUGGGUCAACUGAUACGGUGCUUUGGCCGUGACAAUGCUCGUAAAGAUUGCCACCAGGGCGAAGGAAGAAAUGACUUCGCCUCCCGUAUCUGGAAAG
((((.(..(((...(((((...((((....)))..)))))...))).(((.((((..........)))).)))..((((((((..(....)..))))))))((.......))...)
GUGU
))))
>pH-responsive riboswitch PRE-alx RNA: E. coli.  Alteration:  pH.
AAGUGAGACCUUGCCGGAAGGCGAGGUCUAUGCAUAAAAAGCAGCGGCUGACGUCUUCCGACGUUGGCCGUUUUUUUAUGUGUA
......(((((((((....)))))))))((((((((((((..((((((((((((....))))))))))))))))))))))))))
((((....))))(.((((((((((.((((..(((.......)))..))))..))))))))))).)......................
>preQ1 riboswitch Class II: Streptococcus pneumoniae.  Alteration:  Normal.
GUUGAAUGAAUCAACCCUUGGUGCUUAGCUUCUUUCACCAAGCAUAUUACACGCGGAUAACCGCCAAAGGAGAAAAGAUG
(((((.....))))).(((((((......[[[[[[))))))))..........((((....)))).]]]]].].........
>Purine riboswitch Adenine-sensing add mRNA aptamer domain:  V. vulnificus.  Alteration:  Normal.
CGCUUCAUAUAAUCCUAAUGAUAUGGUUUGGGAGUUUCUACCAAGAGCCUUAAACUCUUGAUUAUGAAGUCUGUCGCUUUAUCCGAAAUUUUAUAAAGAGAAGACUCAUGAAU
............(((((((.......)))))))........((((((.......)))))).((((((.((((.((.((((((..........)))))).)))))))))))))..
((((((((..(((((((..[[.[[]))))))))[.....](((((((]]..]].))))))..))))))))))........................................
>ROSE-1 riboswitch:  Bradyrhizobium japonicum.  Alteration:  Heat.  Not exact match in genome
```

[10] Table B.13: Complete list of genes in this operon is gspC,gspD,gspE,gspF,gspG, gspH,gspI,gspJ,gspK,gspL,gspM,gspO.

[1] Structure partially validated, partially predicted via Vienna Software where not available.

[2] Ten nucleotides added to the structure with no structure.

```
GCCGCGACAAGCGGUCCGGGCGCCCUAGGGGCCCGGCGGAGACGGGCGCCGGAGGUGUCCGACGCCUGCUCGUACCCAUCUUGCUCCUUGGAGGAUUUGGCUAUGAGGA
(((((.....)))))((((((.((....))))))))......((((((.((((....)))).))))))((((((.(((..((.(((....)))))..))).))))))).. 
>36_1_135 ROSE-P2 riboswitch:  Bradyrhizobium.  Alteration:  Heat.  Not exact match with genome
GCCGCUUACGGGCGGUGCGCGGGCGCCAGGUUGGGCUCGGCCAAAGCAACCAGGGCGCCGGACAGGUGUCUUGCACCGAUCGUCUUGCGCUCCUUCGUAUCCAUCUUGCUCUCUG
((((((....))))))(.((((((.((.....)))))))))..........((((((.((((.((((.....))))....)))).))))))(((((((.(((..((.(((....
GAGGACUUGGCUAUGAGGAC
)))))..))).))))))).. 
>ROSE-2 riboswitch:  E. coli.  Alteration:  Heat.
UCUGGCUACCUGACCUGUCCAUUGUGGAAGGUCUUACAUUCUCGCUGAUUUCAGGAGCUAUUGAUUAUGCGUA
..(((((.((((((((.(((.....))))))))))......((.....))....)))))))).............
....((.....(((((.(((.....))))))))...((..(((.(((....))))))....)).....))...
>SAH riboswitch upstream of ahcY: Ralstonia solanacearum.  Alteration:  Normal.  Inferred structural homology.
AAGUUUGCGAUCCGCUAACCGGUCAAGCCGUGUCGCGGAAGGUUGAUGAACCCGCUGAACUCCGGCAGACCCGGAGAAAGGUGAGCGCCCCAUGACU
.....(((..(((((....((((...))))....))))).((((....))))((((...((((((.....))))))...)))).)))..........
>SAM-I riboswitch metFH2:  Thermoanaerobacter tengcongensis.  Alteration:  Normal.
AAUCUCUUAUCAAGAGAGGUGGAGGGACUGGCCCGAUGAAACCCGGCAACCAGCCUUAGGGCAUGGUGCCAAUUCCUGCAGCGGUUUCGCUGAAAGAUGAGAGAUUCUUGUAGUC
................(((...(((.[[[[))).......))).(((..((((((....))).))).)))....]]]](((((....))))).((((.((((((((.....))))
((((((((((......(((...(((.[[[[))).......))).(((..((((((....))).))).)))....]]]](((((....)))))...)))))))))))).........
UCUUCUUUU
)))).))))
.........
>SAM-II riboswitch metA: Agrobacterium tumefaciens.  Alteration:  Normal.  Structural Homology Inferred.
AGUGGUGAUUUGCCGACCGGCUUGCAGCCACUUUAAAGAAGUCGCUAAAGGGUCGAGGAAAAGGGCAAUUUCCUGGGACCGGCCGCGAUUUCGCUGCCGG
(((((((.......[[[[[........))))))).................]]]]]....(((......))).......(((((((((....)))))))))
>SAM-III riboswitch SMK: Streptococcus gordonii.  Alteration:  Normal.
AGUUAUAUAGUCCCGAUAAGAUGGUAGGAAACUUCUAUCAGUUCUUUGUAACUCUAUAACAAUAUUUUUUAAGGGGGGAC
((((((.((...........(((((((.....)))))))...)).))))))).........................
.((((((..[[[[[...(((.(((((((.....)))))))...))).......))))))..............]]]]]
>SAM-IV riboswitch xylE gene:  Streptomyces coelicolor.  Alteration:  Normal.  Structural Homology Inferred.
GGUCAUGAGUGACAGUCAUGAGGCCCCGGCCGACUGUCCGGCAACCCUCCGUCCGUGGCGGGGUGCCCCGGGUGAAGACCAGGUCGUGGACAGCAAGGUCCACGGCAAGCGCGGA
((((....(.(((((((....(([[[[[[[))))))))))))(((..(((.(())..))).)))]]]]]].]..))))..((((((((((......)))))))))).....
CCCCUCGCGGAACCAGGGGUCC
......................
>SAM-V riboswitch 62 metY: Candidatus Pelagibacter ubique.  Alteration:  Normal.
AGGCGCAUUUGAACUGUAUUGUACGCCUUGCAUAAAGCAAAAGUACUAAAAAA
((((..............[[[[[[))))).............]]]].]]....
>SAM-SAH riboswitch metK: Roseobacter sp.  SK209-2-6.  Alteration:  Normal.  Structural Homology Inferred.
CCUGUCACAACGGCUUCCUGGCGUGACGAGGUGACCUCAGUGGAGCAA
((((((((.....[[[[.....))))).))).........]]]]...
>THF riboswitch folT: Eubacterium siraeum.  Alteration:  Normal.[3]
GGACAGAGUAGGUAAACGUGCGUAAAGUGCCUGAGGGACGGGGAGUUGUCCUCAGGACGAACACCGAAAGGUGGCGGUACGUUUACCGCAUCUCGCUGUUCCAUAUAAGAGAUAA
..........(((((((((........((((((((((((......).)))))))))))....(((((..)))))....)))))))))...........((((((((...(.((((
(((((....(((((((((.((.....((((((((((.[[[[.).))))))))))....(((((..)))))))..))))))))))...]]]].)))))).............
UCGCCGUAUUUAUCCCGCAUUAUGCGAUU
(.........))))).)....))))))))
.............................
>TPP riboswitch thiM mRNA: E. coli.  Alteration:  Normal.
GACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUGAUCUGGAUAAUGCCAGCGUAGGGAAGUCACGGACCACCAGGUCAUUGCUUCUUCACGUUAUGGCA
......(((((.(((((.....)))))........)))))......((((..((((......))))..))))((((.....................)))).............
((((..(((((.(((((.....)))))........)))))......((((..((((......))))..))))..))))..((((((...((((..((((((((((.....))).)
GGAGCAAACUAUGCAAGUCGACCUGCUGGGUUC
.................................
))))))............))))...))).)))
>Tryptophan riboswitch trp Operon:  E. coli.  Alteration:  Normal.
UGGUGGCGCACUUCCUGAAACGGGCAGUGUAUUCACCAUGCGUAAAGCAAUCAGAUACCCAGCCCGCCUAAUGAGCGGGCU
(((((...((((.((.......)).))))....)))))(((.....)))...........(((((((.......)))))))
....................(((((...(((.((....(((.....)))....)))))...))))).............
```

B.5.2 Test Set

```
 >ATP riboswitch:  B. subtilis.  ydaO Motif.  Alteration:  Normal.
AAUCGCUUAAUCUGAAAUCAGAGCGGGGGACCCAAUAGAACGGGUUUUUCCCGUAGGGGUGAAUCCUUUUUAGGUAGGGCUAACUCUCAUAUGCCCGAAUCCGUCAGCUAACCUC
((.(((((..((((....))))((((((...(((.....((((((.....)))))))))(((...((((....))).)((((...[[[[[....))))..)))..........))))
GUAAGCGUUCGUGAGAG
)))))))))...]]]]]
>Cobalamin riboswitch:  Salmonella.  Alteration:  Normal.  Expression Platform, Only.
ACUUCGGUGGGAAGUGGGUUGCGAAGACGCGUACAGUCGAAAGACUGAACAUGCGCGUACCUGUAUACCCCUACCACCCUGAACAGGAUCAGGGU
(((((.....)))))((((((((...((((((((((((....)))))....))))))))...)))).))))).....(((((((......)))))))
>Fluoride riboswitch:  Bacillus cereus.  Alteration:  Normal.
UAGGCGAUGGAGUUCGCCAUAAACGCUGCUUAGCUAAUGACUCCUACCAGUAUCACUACUGGUAGGAGUCUAUU
..((((........))))......(((....))).((((((((((((((((....)))))))))))))).)))
..((((...[[[[[)))))......(((....)))....]]]]].............................
>FMN riboswitch ypaA mRNA: B. subtilis.  Alteration:  Normal.
UAUCCUUCGGGGCAGGGUGGAAAUCCCGACCGGCGGUAGUAAAGCACAUUUGCUUUAGAGCCCGUGACCCGUGUGCAUAAGCACGCGGUGGAUUCAGUUUAAGCUGAAGCCGACA
(((((((((...(.(((.......))))....((((...(((((((....)))))))....))))...((((((((....))))))))....(((((......)))))....(((
GUGAAAGUCUGGAUGGGAGAAGGAUGAUGAGCCGCUAUGCAAAAUGUUUAAAAAUGCAUAGUGUUAUUUCCUAUUGCGUAAAAUACCUAAAGCCCCGAAUUUUUUAUAAAUUCGG
(.......))))).....)))))))))).........................................................((((((((((((((.....))))))))
GGCUUU
))))))
>c-di-GMP riboswitch GEMM motif:  Candidatus Desulforudis.  Alteration:  Normal.[4]
ACCCCGAAAGGGCAAACCGGUACGAAAGUCCGGGACGCAAAGCUACGGGUCCUUAAGUUCCAUGGGGAAUAGGACGGCUGAGCCGCUGGGGUUAUUACUUUCGCGGAGCCGCCCU
...........((...((((.((....))))))...)).....................(((....)))((((((((((...(((((((((((....))))))))))))))))))))
```

[3] Partially predicted.

[4] Structural Homology Inferred.

```
(((((.....((...((((.((....))))))...))...(((.((((((......(((....)))...))).))))))...))))))...............(((((.
AUGGGGCGG
)).......
...))))))
>Glycine riboswitch aptamer 2 + alteration of termination:  Bacillus subtilis.  Alteration:  Normal.
CUCUGGAGAGUGUUUGUGCGGAUGCGCAAACCACCAAAGGGGACGUCUUUGCGUAUGCAAAGUAAACUUUCAGGUGCCAGGACAGAGAACCUUCAUUUUACAUGAGGUGUUUCUC
...........((((((((....)))))))).(((...((((...(.(((((....))))))....))))..)))...(((((((((((......................)))))
(((((......((((((((....)))))))).(((...((((...(.(((((....))))))....))))..)))......)))))...((((((.....))))))........
UGUCCU
))))))
......
>Lysine riboswitch:  B. subtilis.  Alteration:  Normal.  Not the genome version.
GGAUAGAGGUGCGAACUUCAAGAGUAUGCCUUUGGAGAAAGAUGGAUUCUGUGAAAAAGGCUGAAAGGGGAGCGUCGCCGAAGCAAAUAAAACCCCAUCGGUAUUAUUUGCUGGC
((((((((.((((.(((((....((.(((((((((((((.[[[[[.)))...))))))))).))...)))))..)))))..(((((((((.(((]]]]]))).)))))))))((( 
CGGGCAUUGAAUAAAUGUCAGGCUGUCAAGAAAUCAUUUUCUUGGAGGGCUAUCC
((((((((.....)))))))))))(((((...........))))).))))))))
>Magnesium riboswitch M-box of MgtE gene:  B. subtilis.  Alteration:  Normal.
CUUCGUUAGGUGAGGCUCCUGUAUGGAGAUACGCUGCUGCCCAAAAAUGUCCAAAGACGCCAAUGGGUCAACAGAAAUCAUCGACAUAAGGUGAUUUUUAAUGCAGCUGGAUGCU
...........(((((..((((((((......((((((((((....((((....))))....))))....))((((((((((.......))))))))))....))))).((((...
((((((.....(((((..((((((((......((((((((((....((((....))))....))))....))((((((((((.......))))))))))....))))).((((...
UGUCCUAUGCCAUACAGUGCUAAAGCUCUACGAUUGAAGGCGCCCGCACGCUUUUUUUGCCGUGCUUCUUUCACCUUCAAUCCCGAAGG
.))))....))))))))).)).....)))...(((((((((.................................)))))))))......
.))))....))))))))).)).....))).))....))))................................(((((......)))))
>Tuco riboswitch.  Geobacter metallireducens.  Alteration:  Normal.
UAGUUUUUUCUCCGAUCCGUCAUACCUACCAGGCGCAGAGCCUCACGGUAUGCGGUCAACGGGUUCCGCUGGAAACGGCGGUGCCUCCCUUUUGGAAAGGAGAACUCUUUA
(((....((((((...((((..((((....((((.....))))...)))).)))).....((((.((((((....)))))).))))((.....))...))))))....)))
>pH-responsive riboswitch:  Serratia marcescens.  Alteration:  pH.
AAGUGAGACCUUGCCGGAAGGCGAGGUUUGCUUGCAGCGUUCAUAGAGCGGCUGGCGUCUUCCGACGUUGGCCGU
......(((((((((....))))))))).................(((((((((((((....)))))))))))))
((((....))))......(((((.((((.((((((.............)))).)))))..))))))))........
>preQ1 riboswitch Class I subtype II: B. Subtilis.  Alteration:  Normal.
AGAGGUUCUAGCUACACCCUCUAUAAAAAACUAAGGACGAGCUGUAUCCUUGGAUACGGC
((((((...........)))))).......(((((((........)))))))......
((((([[[[......)))))......]]]]......(((((((((..)))))))))
>Purine riboswitch Adenine-sensing ydhL mRNA aptamer domain:  B. subtilis.  Alteration:  Normal.
UUGUAUAACCUCAAUAAUAUGGUUUGAGGGUGUCUACCAGGAACCGUAAAAUCCUGAUUACAAAAUUUGUUUAUGACAUUUUUUGUAAUCAGGAUUUU
........(((((((.......)))))))................((((((((((((((((((..((((...))))..))))))))))))))))))
(((((...(((((((.......)))))))........((((((.......))))))..)))))................................
>Purine riboswitch Guanine-sensing xpt mRNA aptamer domain:  B. subtilis.  Alteration:  Normal.
CACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGGGUGAGCAAUGGAACCGCACGUGUACGGUUUUUUGUGAUAUCAGCAUUGC
.....(((...(((((((.......)))))))........((((((.......))))))..)))((((((((((((.(((((........)))))..............))))))
((((((((..(((((((..[[.[[)))))))))[.....](((((((]]..]].)))))))..))))))))........(((((........)))))............((((((((
UUGCUCUUUAUUUGAGCGGGCAAUGCU
))))))......................
((((((.......)))))))))))))))
>ROSE-N1 riboswitch:  Rhizobium.  Alteration:  Heat.
GCCGAUGCCAAUUGGGUCGGCAUGGUCAGGGAGCGCCACGCUUCUUGGCGCUUCCUCGUAUCUAUGUUGCUCUACGGAGGAUGUAGCUAUGAGAAC
((((((.((....))))))))...(((((((((((...)))))))))))).....((((((.(((((((.(((....)))))))))))).)))))))...
>ROSE-2387 riboswitch:  Mesorhizobium loti.  Alteration:  Heat.
GUCGGUCGCCGCAUAAGGGGCCGAUGUGUCAGGGAGCGCCAUGCUUCUUGGCGUUCCCUCGUAUCUAUGUUGCUCCAAGGAGGAUGUAGUUAUG
(((((((.((......))))))))).....((((((((((((.......)))))))))))))((((.(((((((.(((....)))))))))))).)))))
>SAM-I riboswitch apo yitJ S-box:  B. subtilis.  Alteration:  Normal.
UUCUUAUCAAGAGAAGCAGAGGGACUGGCCCGACGAAGCUUCAGCAACCGGUGUAAUGGCGAUCAGCCAUGACCAAGGUGCUAAAUCCAGCAAGCUCGAACAGCUUGGAAGAUAA
.............((((...(((.....)))......)))).(((.((((((...(((((.....))))).)))..)))))).........((((((.....))))))......(((
((((((((....(((((.(((((.[[[[))))...)).)))))((((.((((((...(((((.....))))).)))..)))))))...(]]]](((((((.....)))))))..))))))
GAAGAGACAAAAUCACUGACAAAGUCUUCUUCUUAAGAGGAC
((((((((...............))).)))))))))).......
)))....................((((((((...))))))))
>SAM-III riboswitch metK SMK box:  Enterococcus faecalis.  Alteration:  Normal.
GUUACAAGUUCCCGAAAGGAUUUAGCAAGUAAUUGUCGUUACUUACUAAAGAUGCCUUGUAACCGAAACUAUUUAGGGGGAA
((((((((............(((((.((((((......)))))).)))))......))))))))....................
........(((((...(((.(((((.((((((......)))))).)))))....))).....................)))))
>SAM-IV riboswitch:  Mycobacterium tuberculosis.  Alteration:  Normal.
CUAGGCUUCGAGUCGGUCAUGAGCGCCAGCGUCAAGCCCCGGCUUGCUGGCCGGCAACCCUCCAACCGCGGUGGGGUGCCCCGGGUGAUGACCAGGUUGAGUAGCCAUCGCCGGC
..............((((....(.((((((....(([[[[[[[)).)))))))(((..(((.((......))..))).)))]]]]]].]..))))..((((((((((((......)))
UGCGCGGCAAGCGCGGGUCCGCCAUGACGGGCCC
))))))))......(((((((......)))))))
>SAM-V riboswitch:  Candidatus Pelagibacter ubique.  Alteration:  Normal.  [5]
AAUUAAGCCGGGCAGUUGAACCAUAUUGUGCGCCCUGCAUUUGCUUAAGCACUAAAAAGGAGAAA
......((.((((.((..............)))))).))...(((....)))..............
>SAM-SAH riboswitch:  Oceanibulbus indolifex.  Alteration:  Normal.  [6]
AGAGCAUCACAACGGCUUCCUGACGUGGUGCGUAAUUUUUAUUGGAGCA
...(((((((..((([[[[))))..))))))))............]]]].
>THF riboswitch:  Clostridium kluyveri.  Alteration:  Normal.  [7]
AGCAGAGUAGAACGUUGUGCGUAAAUCAUUGAUUUGCAGUGCCUUCUGAACGGGGAGUUGUCAGAGGGACGAAAAGCCUUUUAGGGCUUACGGUACAAGGUUCGCAUCCCGCUGC
.((((....((((.((((((((((((((...)))))))).((.(((((((((((.....))..)))))))))))...((((((....))))))...))))))).)))).(((..((((((
UGCAUAAGGAGAAGCGGUGUAAUGAUUGU
(...........)))))))..))).))))
>TPP riboswitch thi-box tenA: B. subtilis.  Alteration:  Normal.  [8]
AACCACUAGGGGUGUCCUUCAUAAGGGCUGAGAUAAAAGUGUGACUUUUAGACCCUCAUAACUUGAACAGGUUCAGACCUGCGUAGGGAAGUGGAGCGGUAUUUGUGUUAUUUUA
..(((((.((((((((((.....)))))......((((((.....)))))).)))))).....((((..(((((....)))))..))))..)))))...((((((..(((.......))
CUAUGCCAAUUCCAAACCACUUUUCCUUGCGGGAAAGUGGUUU
```

[5] predicted by pknotsRG©program. (Reeder et al., 2007).

[6] Structural Homology Inferred.

[7] predicted by Vienna.

[8] partially predicted by Vienna.

```
).)))))......(((((((((((((....)))))))))))))
>Tryptophan riboswitch trp Operon:  B. subtilis.  Alteration:  Normal.
GGUAGCAGAGAAUGAGUUUAGUUGAGCUGAGACAUUAUGUUUAUUCUACCCAAAAGAAGUCUUUCUUUUGGGUUUAUUUGUUAUAUAGUAUUUUAUCCUCUCAUGCCAUCUUCUC
(((((..(((((((((......((.(((((((...(((((.......(((((((((((....))))))))))).........))))).............))))).))))....)))
.................................(((((((.((....(((((((((((....))))))))))).....))..))))))).....................(((((
AUUCUCCUUGCCAUAAGGAGUGAGAG
)))))).))))).............. 
(..((((((.....))))))))))))
```

B.5.3 Excluded Set

```
 >ATP Operon:  Salmonella, Aleration:  Unique:  Temperature + Overlap with codons of mgtM.
UGGCAAGUUAACGCACGCUAUUCCUGCGCUGCUUGCCGAACCGGUGGGCAGC
((((((((...((((.........))))..))))))))..............
...........................(((((((((((...)))))))))))
>glms riboswitch:  Thermoanaerobacter tengcongensis.  Alteration:  none.
AGCGCCUGGACUUAAAGCCTTAAGGCUUUAAGUUGACGAGGGCAGGGUUUAUCGAGACAUCGGCGGGUGCCCUGCGGUCUUCCUGCGACCGUUAGAGGACUGGUAAAACCACAGG
..[[[[.[(((((((((((....)))))))))))..[[[...((((((...]]](....)]]]]]...))))))(((((.[[[[[[)))))....((((((((..(((((((((.
CGACUGUGGCAUAGAGCAGUCCGGGCAGGAA
...)))))))))..).))))))).]]]]]].
>GlnA riboswitch:  Synechococcus elongatus.  Alteration:  motif.
CGUUGGCCCAGUUUAUCUGGGUGGAAGUAAGGUCUUUGGCCUGAAGCAACGCGCCUCUCA
(((((.(((((.....)))))..[.....(((((...)))))....)))))]........
>58_1_57 Downstream-peptide motif:  Synechococcus sp.  CC9902.  Alteration:  motif.
CGUUGAGCUUCCAAUCGAAGCUGCAGUCAGACCCAUGCCAAGCAACGGGGGCGUGGG
(((((.(((((.....))))).........[[[[[[[[...)))))]].]]]]]]
>Hammerhead ribozyme Type I: Schistosoma Mansoni.  Alteration:  none.  Tertiary stability.
GCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAAUGCCGGCAUCCUGGAUUCCACUGC
((((...(((((((....)(((((......)))))..(((()))))).)))))).)..))))
>Hammerhead ribozyme Type II: Marine metagenome.  Alteration:  none.  Tertiary stability + pseudoknot.
GCGUGUCGGCCACGGCCCCUUCUGGACCUCGUCCGUGGCCCUGACGAGUAGGGUCCAGAGGGGACGAAACACGC
((((((.(((((((([[[[[[[[[[[[[[...))))))))......(((.]]]]]]]]]]]]]]))).))))))
```

Table B.13: Top Entropy Hits in *E. coli*. Significant hits of the forward and reverse strands of the *E. coli* intergenic regions having high RND entropy (p-Val.<0.500), significantly low (p.Val.<0.050), GC and Uracil compositions within the range of those for known riboswitches Threshold values and their corresponding p-values have been calculated separately for each genome-wide test. 50 nt overlap used for 100 nt scan (100090 segments). 175 nt overlap used for 150 nt scan (66414 segments). Distance from upstream and downstream operons are the distance from the center of the hit to the stop and start codons of upstream and downstream operons, respectively. Probability denotes the multinomial regression likelihood of being a riboswitch under the LMFEGCRND model. Positions are according to gb|U00096.2 version of *E. coli* and not gb|U00096.3 version. Negative values indicate distance to upstream operon. Columns `Upsream/Downstream Operon` show gene ID within the operon.

E. coli	Start	End	Strand	Upstream Operon	Dist. to Upstream	MFE	MFE p. Val.	GC	RND	RND p. Val.	Uracil	Dist. to Downstream	Downstream Operon	Probability
100nt	4083889	4083988	forward	yiiF	-5848	-38.4	0.0267	0.53	58.6367989	0.0365	0.29	102	fdhD	0.789
100nt	187962	188061	forward	cdaR	-4293	-36.4	0.0466	0.53	59.0985985	0.0229	0.32	1702	rpsB,tff,tsf	0.776
100nt	952485	952584	forward	ycaK	-2955	-36.8	0.0419	0.52	58.3203011	0.0494	0.27	3452	ycaP	0.765
100nt	4115038	4115137	forward	uspD,yiiS	-3245	-37	0.0396	0.53	58.3563995	0.0477	0.33	1452	zapB	0.756

E. coli	Start	End	Strand	Upstream Operon	Dist. to Upstream	MFE	MFE p. Val.	GC	RND	RND p. Val.	Uracil	Dist. to Downstream	Downstream Operon	Probability
150nt	2686923	2687072	forward	hmp	-1797	-56.00	-	0.5333	90.7522964	0.0077	0.32000	6822	mltF	0.8671584129
150nt	452721	452870	forward	yajQ	-8244	-60.90	-	0.5200	88.2920990	-	0.23333	897	bolA	0.8664909005
150nt	1100699	1100848	forward	ycdZ	-610	-58.70	-	0.4933	87.8781967	-	0.30666	2397	csgA,csgB,csgC	0.8559710383
150nt	2887386	2887535	forward	iap	-11667	-56.40	-	0.5333	89.1240005	0.0294	0.31333	2772	queD	0.8254097700
150nt	3467187	3467336	forward	gspO[10]	-2866	-56.10	-	0.5200	88.5419006	0.0450	0.29333	8397	slyX	0.8172816634
150nt	2553118	2553267	forward	amiA,hemF	-893	-57.40	-	0.5133	87.6467972	-	0.28666	3597	intZ	0.8125300407
150nt	2660264	2660413	forward	ryfA	-8005	-56.90	-	0.5000	87.1239014	-	0.32000	1122	suhB	0.8031908870
150nt	1766798	1766947	forward	lpp	-11038	-57.00	-	0.4800	85.6548004	-	0.25333	222	ydiK	0.7757616639
150nt	1718374	1718523	forward	slyB	72	-58.70	-	0.4867	85.1240005	-	0.32666	597	ydhI,ydhJ,ydhK	0.7731205821
150nt	4356712	4356861	forward	yjdK,yjdO	-5529	-58.80	-	0.5200	86.0333023	-	0.26000	9897	fxsA	0.7661048174
150nt	149580	149729	forward	yadD	-1631	-57.80	-	0.4600	84.3807983	-	0.30666	12447	hrpB	0.7651519775
150nt	4604476	4604625	forward	yjjZ	-334	-57.20	-	0.4867	85.4507980	-	0.27333	1272	holD,rimI,yjjG	0.7621335387
150nt	3120069	3120218	forward	C0719	-389	-56.40	-	0.5267	87.1032028	-	0.27333	6147	glcC	0.7610746622
150nt	1982024	1982173	forward	uspC	-3740	-56.20	-	0.5333	87.3750000	-	0.26666	2847	ftnB	0.7596676350
150nt	3921878	3922027	forward	cbrB,cbrC	-25167	-56.00	-	0.4933	85.5883026	-	0.30666	3222	asnA	0.7384917736
150nt	790921	791070	forward	aroG	-4934	-57.00	-	0.5333	86.4469986	-	0.28000	2997	acrZ	0.7345629930
150nt	4482291	4482440	forward	yjgN	-3263	-58.30	-	0.5200	85.4577026	-	0.28666	1872	lptF,lptG	0.7340587974
150nt	518357	518506	forward	ybbL,ybbM	-1692	-57.10	-	0.5000	85.1729965	-	0.26000	522	ybbA,ybbP	0.7300664783
150nt	1167546	1167695	forward	ycfJ	-106	-57.70	-	0.5333	85.9982986	-	0.27333	672	bhsA	0.7274026275
150nt	1642496	1642645	forward	cspF	-2326	-56.00	-	0.5333	86.4957962	-	0.25333	1347	ydfV	0.7190257311
150nt	3916103	3916252	forward	cbrB,cbrC	-19392	-60.90	-	0.5200	84.0416031	-	0.30000	8997	asnA	0.7181233764
150nt	3258127	3258276	forward	yhaK,yhaL	-4819	-55.70	-	0.5267	86.1580963	-	0.26000	7197	tdcR	0.7085512877
150nt	3721360	3721509	forward	insK	-1209	-57.70	-	0.5133	84.8648987	-	0.26000	2472	wecH	0.7070741653
150nt	2438211	2438360	forward	flk	-1165	-58.50	-	0.5267	84.9561996	-	0.24000	1497	mnmC	0.7053987384
150nt	1268246	1268395	forward	kdsA,ychA,ychQ	75	-56.90	-	0.4933	84.3839035	-	0.29333	222	rdlA	0.7012539506
150nt	219458	219607	forward	arfB,nlpE,yaeQ	-3400	-59.90	-	0.5267	84.2009964	-	0.27333	3297	gmhB	0.6966923475
150nt	3514668	3514817	forward	frlA,frlB,frlC,frlD,frlR	-11784	-56.50	-	0.5333	85.6490021	-	0.27333	6147	mrcA	0.6895118356
150nt	3313884	3314033	forward	psrO	-4385	-55.60	-	0.5067	84.9284973	-	0.28666	2697	argG	0.6809250712
150nt	649808	649957	forward	ybdR	-7180	-58.10	-	0.5333	84.6990967	-	0.25333	1572	dpiA,dpiB	0.6750817299
150nt	2243989	2244138	forward	yeiG	-1142	-57.10	-	0.5333	84.3582993	-	0.26000	3672	yeiH	0.6381362677
150nt	4253685	4253834	forward	ubiA,ubiC	-1695	-55.80	-	0.5267	84.4711990	-	0.29333	897	dgkA	0.6285824776
150nt	2866503	2866652	forward	ygbN	-1937	-56.80	-	0.5333	84.2586975	-	0.30000	8022	iap	0.6268372536
150nt	3576825	3576974	reverse	yhhW	-69	-55.60	-	0.4800	88.6371994	0.0419	0.30666	154	gntK,gntR,gntU	0.8547886610
150nt	260750	260899	reverse	yafW,yafX,yafY,ykfB,ykfF,ykfG,ykfH,ykfI	-1723	-62.00	-	0.5267	86.8554001	-	0.32000	1504	phoE	0.8323625326
150nt	2195866	2196015	reverse	yehS	-13803	-58.00	-	0.5333	88.6897964	0.0405	0.27333	3754	mrp	0.8320623040
150nt	4176725	4176874	reverse	sroH	-11546	-58.80	-	0.5333	88.1481018	-	0.28000	3754	coaA	0.8252514601
150nt	133211	133360	reverse	speD,speE,yacC	-1498	-56.40	-	0.5133	88.0559998	-	0.28000	2029	yacH	0.8126859665
150nt	44332	44481	reverse	apaG,apaH,lptD,pdxA,rsmA,surA	-5969	-64.00	-	0.5333	85.3160019	-	0.26000	2479	caiA,caiB,caiC,caiD,caiE,caiT	0.8009238243
150nt	2248916	2249065	reverse	nupX	-1922	-55.70	-	0.5333	88.3962021	-	0.28000	1354	yeiE	0.7910096645
150nt	2774331	2774480	reverse	ypjA	-1758	-55.60	-	0.5200	87.4978027	-	0.26666	3229	ypjM_1,ypjM_2	0.7727379203
150nt	2184269	2184418	reverse	yehA,yehB,yehC,yehD	-1054	-56.20	-	0.5333	87.6942978	-	0.30666	529	rcnR	0.7722578049
150nt	1655707	1655856	reverse	mlc,ynfK	-8762	-58.20	-	0.5267	86.6440964	-	0.24000	304	ynfC	0.7716723680
150nt	2168389	2168538	reverse	gatR_2	-951	-56.70	-	0.5133	86.7958984	-	0.31333	304	gatR_1	0.7715415359
150nt	3481959	3482108	reverse	yhfA	-1398	-57.90	-	0.5333	86.7799988	-	0.25333	2854	kefB,kefG,yheV	0.7633723617
150nt	314642	314791	reverse	ykgA	-953	-58.20	-	0.5067	85.1943970	-	0.32000	2254	ykgR	0.7403761148
150nt	2783559	2783708	reverse	ileY	-146	-56.20	-	0.4667	84.4744034	-	0.26666	604	ypjC	0.7329583764
150nt	4050403	4050552	reverse	glnA,glnG,glnL	-1410	-56.70	-	0.5267	86.2409973	-	0.27333	1729	yihA	0.7301918864
150nt	2465055	2465204	reverse	yfdK,yfdL,yfdM,yfdN,yfdO	-3965	-58.50	-	0.5333	85.6035004	-	0.25333	2104	mlaA	0.7242872119
150nt	2809717	2809866	reverse	luxS	-2444	-56.20	-	0.5333	86.4486008	-	0.29333	11779	ygaC	0.7205215693
150nt	3341463	3341612	reverse	elbB,mtgA	-5561	-57.50	-	0.5000	84.6742020	-	0.28000	3454	mlaB,mlaC,mlaD,mlaE,mlaF	0.7152099609
150nt	1950122	1950271	reverse	torY,torZ	-2401	-56.44	-	0.5200	85.7341003	-	0.24666	1654	aspS	0.7131192684
150nt	3059434	3059583	reverse	ygfI	-4786	-56.20	-	0.4933	84.9087982	-	0.27333	1879	yqfE	0.7122740149
150nt	2530006	2530155	reverse	pdxK	-4323	-58.20	-	0.5000	84.2724991	-	0.25333	829	zipA	0.7098694444
150nt	757742	757891	reverse	mngR	-6555	-57.90	-	0.4867	83.9291000	-	0.31333	4129	gltA	0.7092900276
150nt	4452551	4452700	reverse	fbp	-4	-56.90	-	0.5200	85.3691025	-	0.29333	4954	ppa	0.7049999237
150nt	3584997	3585146	reverse	ugpA,ugpB,ugpC,ugpE,ugpQ	-317	-59.30	-	0.5133	84.1650009	-	0.32000	229	ggt	0.7046704292
150nt	3459796	3459945	reverse	bfd,bfr	-4396	-59.60	-	0.5200	84.1896973	-	0.23333	6454	gspA,gspB	0.7009978294
150nt	3395856	3396005	reverse	mreB,mreC,mreD	-474	-56.20	-	0.5000	84.8627014	-	0.25333	1654	yhdP	0.6998521686
150nt	2321601	2321750	reverse	yfaA,yfaP,yfaQ,yfaS_1,yfaS_2,yfaT	-3709	-61.10	-	0.5333	83.9000015	-	0.24000	3829	rcsC	0.6947578788
150nt	4169919	4170068	reverse	coaA	-2101	-56.20	-	0.4933	84.4697037	-	0.26000	8704	trmA	0.6920779943
150nt	4029970	4030119	reverse	mobA,mobB	-8880	-57.80	-	0.5133	84.4968033	-	0.30000	1054	fadA,fadB	0.6919249892
150nt	2366493	2366642	reverse	pmrD	-4722	-56.00	-	0.5200	85.1843033	-	0.28000	2929	ais	0.6792802215
150nt	1850076	1850225	reverse	ydjE	-490	-57.40	-	0.5267	84.6417999	-	0.27333	3454	selD,topB,ydjA	0.6695327163
150nt	3246739	3246888	reverse	yhaJ	-4522	-56.00	-	0.4867	83.8035965	-	0.26000	4054	uxaA,uxaC	0.6671289802
150nt	3883703	3883852	reverse	purP	-9513	-56.60	-	0.5200	84.5989990	-	0.28000	2029	dnaA,dnaN,recF	0.6626442075
150nt	4395241	4395390	reverse	yjfN	-18720	-56.20	-	0.5267	84.9857025	-	0.29333	3229	queG	0.6626062393
150nt	2664571	2664720	reverse	hcaT	-79	-56.20	-	0.5267	84.7899017	-	0.24666	3304	trmJ	0.6529098153
150nt	2975533	2975682	reverse	lysA	-47	-57.20	-	0.5333	84.5955963	-	0.24666	1204	omrB	0.6521243453
150nt	796718	796867	reverse	ybhA	-39	-58.40	-	0.5333	83.8794022	-	0.26666	2929	modE,modF	0.6405209899
150nt	3288088	3288237	reverse	rsmI	-2330	-55.90	-	0.5067	83.9057007	-	0.28000	11479	agaR	0.6363855004
150nt	3211791	3211940	reverse	mug	-1119	-56.10	-	0.5200	84.1829987	-	0.25333	3304	tsaD	0.6315983534
150nt	2161503	2161652	reverse	ogrK	-3744	-55.70	-	0.5333	84.3114014	-	0.29333	10429	yegK,yegL	0.6064385772
150nt	3284413	3284562	reverse	rsmI	-6005	-55.60	-	0.5333	84.2777023	-	0.25333	7804	agaR	0.6025381684