Published in final edited form as:

J Theor Biol. 2013 February 7; 318: 140–163. doi:10.1016/j.jtbi.2012.10.023.

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

Amirhossein Manzourolajdad^{a,1}, Yingfeng Wang^b, Timothy I. Shaw^a, and Russell L. Malmberg^{a,c}

^aInstitute of Bioinformatics, University of Georgia, Davison Life Sciences Bldg, Room B118B, 120 Green St, Athens, GA 30602, USA

^bComputer Science and Mathematics Division, Oak Ridge National Lab, Oak Ridge, TN 37831, USA

^cDepartment of Plant Biology, University of Georgia, 2502 Miller Plant Sciences, Athens, GA 30602, USA

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.

Keywords

RNA secondary structure; stochastic context-free grammars; structural entropy; riboswitch; bacterial type B RNase P; structural ambiguity

1. 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 interaction (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; Niranjankumari 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; Niranjankumari 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 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 $X \rightarrow aYb$ 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. 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)certainly 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 (Yockey, 2005; Adami et al., 2000) 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 effectively capture 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)certainly 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. 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 time¹. 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 $\Sigma = \{A, C, G, U\}$. 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:

$$S_0 = \gamma_0 \Rightarrow \gamma_1 \Rightarrow \gamma_2 \Rightarrow \dots \Rightarrow \gamma_{n_\pi} = y \quad (1)$$

where $\gamma_i \in (NU\Sigma)^*$, and $\gamma_i \Rightarrow \gamma_{i+1}$ satisfies that $\gamma_i = \omega X \beta$ and $\gamma_{i+1} = \omega \alpha \beta$ for some $\omega \in \Sigma^*$, $\beta \in (NU\Sigma)^*$ and rule $X \rightarrow \alpha$ nonterminal in the grammar, i.e., the occurrence of 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).

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

We denote the derivation (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) illustrates the correspondence between derivations and secondary structures with CFG, where an example grammar with only four types of generic rules is used:

$$X \rightarrow aYbZ, \quad X \rightarrow aYb, \quad X \rightarrow aY, \quad X \rightarrow a \quad (2)$$

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.

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 (1) under a given SCFG Model (G, Θ) is defined as

$$p(\pi, y | G, \Theta) \equiv p(\pi, y) = P(S_0 \Rightarrow_{\pi}^* y) = \prod_{i=1}^{n_{\pi}} P(R_{\pi}^i) \quad (3)$$

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

3. Structural entropy over SCFG ensembles

As noted previously, Shannon's entropy measures the (un)certainly 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:

$$\log P(S_0 \Rightarrow^* y) - \frac{1}{P(S_0 \Rightarrow^* y)} \sum_{\pi \in \Pi(y)} P(S_0 \Rightarrow_{\pi}^* y) \log P(S_0 \Rightarrow_{\pi}^* y) \quad (4)$$

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 is used, here (Durbin, 1998). (See Appendix A)

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 $\langle X \rightarrow \gamma, i, j \rangle_{\pi}$ to denote the instance of rule $X \rightarrow \gamma$ 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 $\lambda \in (\mathcal{M} \Sigma)^*$, where $\Sigma = \mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{U}$.

Likewise, we denote an instance of rule $X \rightarrow \gamma$ applied in some structure by $\langle X \rightarrow \gamma, i, j \rangle$. Note that the applications of rule $X \rightarrow \gamma$ in $\langle X \rightarrow \gamma, i, j \rangle \pi$ and in $\langle X \rightarrow \gamma, i, j \rangle$ have the same probability, which is the probability $\mathcal{F}(X \rightarrow \gamma)$ for rule $X \rightarrow \gamma$ given in the SCFG.

The term $\sum_{\pi \in \Pi(y)} P(S_0 \Rightarrow_{\pi}^* y) \log P(S_0 \Rightarrow_{\pi}^* y)$ in (4) becomes

$$\begin{aligned} \sum_{\pi \in \Pi(y)} \prod_{\langle Y \rightarrow \delta, k, l \rangle_{\pi}} \mathcal{F}(Y \rightarrow \delta) \log \prod_{\langle X \rightarrow \gamma, i, j \rangle_{\pi}} \mathcal{F}(X \rightarrow \gamma) &= \sum_{\pi \in \Pi(y)} \prod_{\langle Y \rightarrow \delta, k, l \rangle_{\pi}} \mathcal{F}(Y \rightarrow \delta) \prod_{\langle X \rightarrow \gamma, i, j \rangle_{\pi}} \log \mathcal{F}(X \rightarrow \gamma) \\ &= \sum_{\langle X \rightarrow \gamma, i, j \rangle} \log \mathcal{F}(X \rightarrow \gamma) \sum_{\pi \ni \langle X \rightarrow \gamma, i, j \rangle} \prod_{\langle Y \rightarrow \delta, k, l \rangle_{\pi}} \mathcal{F}(Y \rightarrow \delta) \end{aligned} \quad (5)$$

where term

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

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

$$\begin{aligned} \sum_{\pi \ni \langle X \rightarrow \gamma, i, j \rangle} \prod_{\langle Y \rightarrow \delta, k, l \rangle_{\pi}} \mathcal{F}(Y \rightarrow \delta) &= \sum_{\omega, \lambda} P(S_0 \Rightarrow_{\omega}^* \omega X \lambda, \omega \Rightarrow^* y_1 \dots y_i, \lambda \Rightarrow^* y_j \dots y_{n_y}) \times P(X \rightarrow \gamma, \gamma \Rightarrow^* y_{i+1} \dots y_{j-1}) \\ &= \beta(X, i, j, y) \mathcal{F}(X \rightarrow \gamma) P(\gamma \Rightarrow^* y_{i+1} \dots y_{j-1}) \end{aligned}$$

where the *Outside* probability function β is the inclusive definition of the outside probability function (Durbin, 1998). (See Appendix A)

Replacing the corresponding terms in formulae (4) and (5) with the above derivations, the structural entropy of given sequence y is computed as

$$\log \alpha_0 - \frac{1}{\alpha_0} \sum_{i \leq j} \sum_{X \rightarrow \gamma} \beta(X, i, j, y) \mathcal{F}(X \rightarrow \gamma) \log \mathcal{F}(X \rightarrow \gamma) P(\gamma \Rightarrow^* y_{i+1} \dots y_{j-1}) \quad (6)$$

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), for simplicity of presentation (See Appendix A for a derivation generalized to all types of non-stacking² grammar rules). Based on them, term $\Sigma_{X \rightarrow \gamma}$ in (6) can be computed as

²Non-stacking grammar rules here include: $X \rightarrow a$, $X \rightarrow aY$, $X \rightarrow Ya$, $X \rightarrow aYbZ$, $X \rightarrow YaZb$, and $X \rightarrow aYb$, but not $X \rightarrow e$.

$$\beta(X, i, j, y) \left[f(X, aYbZ) \sum_{i+2 < k < j-1} \alpha(Y, i+2, k-1, y) \alpha(Z, k+1, j-1, y) \right. \\ \left. + f(X, aYb) \alpha(Y, i+2, j-2, y) + f(X, aY) \alpha(Y, i+2, j-1) + \delta(i+2=j) f(X, a) \right] \quad (7)$$

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

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

Equation (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 (7) must be modified according to the left-most derivation criterion to avoid redundant enumeration of derivation trees. Appendix A contains the algorithmic details of a derivation generalized to ambiguous grammars. In dealing with structurally ambiguous grammars, we refer to (7) as the redundant derivation and its modified version in Appendix A 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 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:

$$\text{TP Entropy}(y) = \sum_{i < j} H(I_{i,j}|y) \quad (8)$$

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 Appendix I.1).

$$H(\Pi|y) \leq \text{TP Entropy}(y) \quad (9)$$

4. Evaluating the Structural Entropy of ncRNAs

The structural³ 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

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

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 folding 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 (un)stability, alternative/multiple folds, and formation of pseudoknots could all affect the structural entropy, since they are directly related to the folding distribution valid on the sequence. There is no reason to believe that the entropy of sequences with such different structural features will have a similar behavior.

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.

4.1. 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.

5. Methodology

5.1. 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 (dataset) 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 dataset. Please refer to (Dowell and Eddy, 2004) for details about grammar rules and training datasets.

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:

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

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

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-art thermodynamic folding models.

5.2. Data Collection

Having a reliable annotated secondary structure for the RNA sequence is very 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 dataset includes sequences of one stable secondary structure, such as miR-NAs, 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 contains 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 include: 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 regarding average length and average identity is included in Appendix B.

5.2.1. 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 C.2 and C.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.

5.3. Measuring the Significance of Folding Entropy

Various primary and secondary structure randomizations have been performed to both evaluate 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 we a priori 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 single-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 structural entropy of the ncRNA 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 implementation for each sequence, separately. The Altschul-Erikson 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 then empirically obtained by comparing the structural entropy of the ncRNA 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 Appendix G 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 dinucleotide 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 were used in the single-nucleotide randomization test and 60 sequences were used for the di-nucleotide randomization test)⁴

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 Appendix E. 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. Appendix D 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 arbitrary 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

⁴RNase 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 the 60 sequences used for di-nucleotide shuffling test. Number of sequences with available results slightly vary for various models.

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.

5.4. 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:

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

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 (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.

6. 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 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 {60nt – 600nt}⁵. This is also true for redundant derivation of structural entropy for the structurally ambiguous 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 1 shows

⁵Simulation results suggest higher-order relationship between sequence length and its structural entropy for length values higher than 600nt (data not shown).

consistently high entropy for rfamv5-trained parameter sets compared to those for benchmark and mixed80 regardless of choice of grammar.

6.1. 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 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 distribution⁶ (see H.15). 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 H.16); i.e., p-values of some sequences belonging to a ncRNA class do not seem to be a function of 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 H.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 H.17 is a plot between structural 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, structural 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 corresponding sequence.

The p-value distribution of folding entropy of sequences belonging to the same class of ncRNA has been observed to be linear in most cases and under most models (data only shown for the RUN (benchmark) model, figure 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

⁶Even 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 H.7 and H.8 for sum of pair-wise Kolmogorov-Smirnov (KS) distance corresponding to RNA families for each model)

under a given the grammar model (Table H.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 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 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-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 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 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 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) (See Appendix E for details about clusters and sequences). Results for the IVO (benchmark) grammar are moved to F.12. 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 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 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 F.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 F.12 and table F.4)

6.2. 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 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 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 7 and Table F.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. Figures 8 and 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 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 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 H.18 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 10 and 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 H. 19, H.20, H.21, and H.22). 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.

7. Discussion

7.1. Discerning Structural Ambiguity

Shannon's entropy of SCFG ensembles on a given sequence was derived and implemented for both structurally ambiguous and unambiguous grammars. Appendix A 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.

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.

7.2. 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 1).

7.3. The Significance of Folding Entropy of ncRNAs

A qualitative inspection of figures 3 and 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 C.3 to average entropy p-values in 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:

7.3.1. 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 H.14 and H.15). 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 C.3 and H.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.

7.3.2. Impact of Class of ncRNA—The average structural entropy p-values obtained from single- and dinucleotide randomization tests show a somewhat consistent ranking across most models (See figure 3 and tables H.5 and H.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 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 6 for average p-values of riboswitch sub-families under various folding models).

7.4. 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 7 and table F.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 8 and 9). The low folding entropy results of the miRNA are also in line with results of the base-pairing entropy test presented here, 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.

7.5. 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 Appendix I.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 Appendix I.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 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 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 can 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 H.18). 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 3).
 - b. *Scenario 2:* High entropy can 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 H.18 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 H.18 to 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 H.9 and figure H.17). 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.

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 a sequence. 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 7 and F.12 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 10 and 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.

7.6. 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 3 and 4, and tables H.5 and H.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 (9). Conclusions derived from the di-nucleotide randomization test are, however, subject to flaws of this test, as previously discussed.

7.7. SCFG Modeling of Non-coding RNA Sequences

7.7.1. 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 1). Our conclusion about grammar design is that structural ambiguity can play a major role in characterizing ncRNA sequences.

7.7.2. 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, 8 and 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.

8. 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.

Acknowledgments

Funding: This work was supported in part by research grants NIH BISTI R01GM072080-01A1, NIH ARRA Administrative Supplement, and NSF IIS Award No: 0916250. Most of the work was completed when the first authors were associated with the University of Georgia RNA Informatics Lab to which the research grants were awarded. The conceptual framework of this research was first conceived by Dr. Liming Cai of the RNA Informatics Lab who subsequently provided many technical derivations that have lead to the theoretical results in this work. We also thank Robin D. Dowell for providing us with source code for model training through email correspondence. Also, Yann Ponty greatly helped us regarding proper use of the GenRGenS software package, through email correspondence. We also would like to acknowledge Ming Zhang.

Appendix A. 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:

$$\omega \rightarrow \gamma c_j; \gamma \rightarrow c_i \delta \quad (\text{A.1})$$

Where c_i and c_j are substructures: $c_i, c_j \in \{a, aYb\}$, and λ, ω, γ 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: $\alpha(\gamma, i, j, y)$.

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: $\beta(\gamma, i, j, y)$.

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 (7) with the following modifications for the expression containing the inside and outside probability function coefficients:

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

If $X \neq Z$:

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

If $X = Z$:

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

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

If $X \neq Y$:

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

If $X = Y$:

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

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 exclusive 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.

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

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

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

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

With initialization:

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

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

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

Appendix B. Data Collection

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

Table B.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

Accession	Name	Type	No. Seed	Average Length	Average % Identity
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

Accession	Name	Type	No. Seed	Average Length	Average % Identity
RF00010	bact. type A	RNase P	306	380	62

Appendix C. Sensitivity and Specificity (PPV) of models to Annotated Secondary Structure of ncRNAs

Table C.2

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 C.3

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

Mixed80-trained	BJK Sen.	BJK Sp.	RUN Sen.	RUN Sp.	IVO Sen.	IVO Sp.
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

Appendix D. Generating random structures

In order to generate random sequences with structure, we used the Gen-RGenS 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.

Appendix E. 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).

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

Table F.4

Structural entropy p-values of riboswitch and miRNA sequences against random sequences with structure under various folding models (See Appendix E 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 F.12.

Structural entropy p-values of short ncRNA sequences against random sequences with structure under the IVO model (See Appendix E for information about clusters and sequences). Benchmark training set was used. Other training sets yield similar results (See F.4).

Appendix G. P-value Stability Test for Dinucleotide Shuffling

Figure G.13.

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 dataset. 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.

Appendix H. Structural Entropy Empirical P-values for single nucleotide composition and dinucleotide shuffling tests

Figure H.14.

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.

Figure H.15.

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.

Table H.5

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 H.6

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

ncRNA Family	RND10 (Di.)	RND10 (Sing.)
snoRNA	0.476	0.24

Table H.7

Kolmogorov-Smirnov test statistic distance of dinucleotide-shuffled basepairing (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 H.8

Kolmogorov-Smirnov test statistic distance of dinucleotide-shuffled basepairing (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 Appendix A).

	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

Table H.9

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

	AA	CA	GA	UA
miRNA	-0.33	0.29	-0.18	-0.07
RNase P (bact. Type B)	0.21	-0.24	0.3	0.41

Figure H.16.

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.

Figure H.17.

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 H.18.

Structural entropy p-values of Bacterial type B RNase P 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.

Figure H.19.

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 Appendix E for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See Appendix D for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

Figure H.20.

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 Appendix E for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See Appendix D for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

Figure H.21.

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 Appendix E for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See Appendix D for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

Figure H.22.

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 Appendix E for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See Appendix D for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

Appendix I. Supplementary

Appendix I.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, 2006, Theorem 2.6.6, pg. 30).

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

Substituting for pairing uncertainties gives

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

Appendix I.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, 2006, pg. 409) states that the Shannon folding entropy of such a distribution is higher than any other folding distribution so long as the sizes of the possible folds are kept equal.

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

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*⁷ nucleotides with uniform alphabet distribution will have higher entropy than any other sequences of the same length and alphabet. Let's call this random sequence space Y^{**} .

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

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*⁸ instance of the random sequence y^{**} under 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.

Appendix I.2.1. 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:

$$\exists f(\cdot), X_1, X_2: H(f(X_1)) < H(f(X_2)) \quad \text{where } H(X_1) > H(X_2) \quad (\text{I.1})$$

⁷iid is the shorthand for independent and identically distributed random variables

⁸The typical set is a set of sequences whose probability is close to one (Cover and Thomas, 2006, pg. 62).

Consider the following scenario:

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

$$p_{x_1}(0) = 1/2 \quad p_{x_1}(1) = 1/2$$

$$p_{x_2}(0) = 3/5 \quad p_{x_2}(1) = 2/5$$

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

$$H(X_1) = - (1/2) \log(1/2) - (1/2) \log(1/2) = 1$$

$$H(X_2) = - (3/5) \log(3/5) - (2/5) \log(2/5) = 0.971$$

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

$$\begin{aligned} H(f(X_1)) &= - ((1/2)(2/3) + (1/2)(1/4)) \log((1/2)(2/3) + (1/2)(1/4)) \\ &\quad - ((1/2)(1/3) + (1/2)(3/4)) \log((1/2)(1/3) + (1/2)(3/4)) = 0.995 \end{aligned}$$

$$\begin{aligned} H(f(X_2)) &= - ((3/5)(2/3) + (2/5)(1/4)) \log((3/5)(2/3) + (2/5)(1/4)) \\ &\quad - ((3/5)(1/3) + (2/5)(3/4)) \log((3/5)(1/3) + (2/5)(3/4)) = 1 \end{aligned}$$

Hence, function $f(.)$, X_1 , and X_2 satisfy I.1.

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:

$$H(X_1^L) = L$$

$$H(X_2^L) = 0.917L$$

$$H(F(X_1^L)) = 0.995L$$

$$H(F(X_2^L)) = L$$

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.

Appendix I.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:

GGGGGGGGGGCGCGCGCGCGCCCCCCCC

With a single hairpin loop:

(((((((((.....))))))))))

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:

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

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:

GGGGGGGGGGCGCGCGCGCCCCCGCGCG; (.....); 0
 GGGCCCCCGGGCGCGGGGGGGCGCGCGCG; ((.....)); 0.453339
 GGGGGCGGGCGCGCGCCCCCGGGGGCGCG; (((.....)); 0.58333
 GGGGGGGGGCGCGCGGGGGGGGGGGGG; (((((((.....)))))); 0.619343
 GGGGGGGGGCGCGCGGGGGGGGGGGGG; (((((((.....)))))); 0.631615

References

- Adami C, Ofria C, Collier TC. Evolution of biological complexity. *PNAS*. 2000; 97:4463–4468. [PubMed: 10781045]
- Altschul SF, Erickson BW. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol*. 1985; 2:526–38. [PubMed: 3870875]
- Barash D, Sikorski J, Perry EB, Nevo E, Nudler E. Adaptive mutations in RNA-based regulatory mechanisms: computational and experimental investigations. *Israel Journal Of Ecology and Evolution*. 2006:52.
- Barrandon C, Spiluttini B, Bensaude O. Non-coding RNAs regulating the transcriptional machiner. *Biol. Cell*. 2008; 100:8395.
- Batey RT, Rambo RP, Doudna JA. Tertiary motifs in RNA structure and folding. *Angew Chem Int Ed Engl*. 1999; 38:2326–43. [PubMed: 10458781]
- Bernauer J, Huang X, Sim AY, Levitt M. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*. 2011; 17:1066–1075. [PubMed: 21521828]
- Bocobza S, Adato A, Mandel T, Shapira M, Nudler E, Aharoni A. Riboswitch-dependent gene regulation and its evolution in the plant kingdom. *Genes Dev*. 2007; 21:2874–9. [PubMed: 18006684]
- Brannvall M, Mattsson JG, Svard SG, Kirsebom LA. RNase p RNA structure and cleavage reflect the primary structure of tRNA genes. *J Mol Biol*. 1998; 283:771–83. [PubMed: 9790839]
- Cech TR, Damberger SH, Gutell RR. Representation of the secondary and tertiary structure of group I introns. *Nat Struct Biol*. 1994; 1:273–80. [PubMed: 7545072]
- Chan CY, Ding Y. Boltzmann ensemble features of RNA secondary structures: a comparative analysis of biological RNA sequences and random shuffles. *J Math Biol*. 2008; 56:93–105. [PubMed: 17909813]
- Clote P, Ferre F, Kranakis E, Krizanc D. Structural rna has lower folding energy than random rna of the same dinucleotide frequency. *RNA*. 2005; 11:578–91. [PubMed: 15840812]
- Cover, TM.; Thomas, JA. *Elements of Information Theory*. 2nd edition. Wiley-Interscience; 2006.
- D’Haeseleer P. What are DNA sequence motifs? *Nature Biotechnology*. 2006; 24:423–5.
- Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*. 2003; 31:7280–301. [PubMed: 14654704]
- Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*. 2006; 22:e90–8. [PubMed: 16873527]
- Dowell RD, Eddy SR. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*. 2004; 5:71. [PubMed: 15180907]
- Du X, Wang ED. Tertiary structure base pairs between d- and tpsicloops of escherichia coli tRNA(leu) play important roles in both aminoacylation and editing. *Nucleic Acids Res*. 2003; 31:2865–72. [PubMed: 12771213]
- Durbin, R. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press; 1998.
- Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A. Rfam: updates to the RNA families database. *Nucleic Acids Res*. 2009; 37:D136–40. [PubMed: 18953034]
- Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*. 2005; 33:2433–9. [PubMed: 15860779]
- Gilbert SD, Love CE, Edwards AL, Batey RT. Mutational analysis of the purine riboswitch aptamer domain. *Biochemistry*. 2007:46.
- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. 2005; 33:D121–4. [PubMed: 15608160]
- Guerrier-Takada C, Altman S. A physical assay for and kinetic analysis of the interactions between m1 RNA and tRNA precursor substrates. *Biochemistry*. 1993; 32:7152–61. [PubMed: 7688247]

- Hall MN, Gabay J, Debarbouille M, Schwartz M. A role for mRNA secondary structure in the control of translation initiation. *Nature*. 1982; 295:616–8. [PubMed: 6799842]
- Hofacker IL. Vienna RNA secondary structure server. *Nucleic Acids Res*. 2003; 31:3429–31. [PubMed: 12824340]
- Huynen M, Gutell R, Konings D. Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol*. 1997; 267:1104–12. [PubMed: 9150399]
- Kazantsev AV, Pace NR. Bacterial RNase p: a new view of an ancient enzyme. *Nat Rev Microbiol*. 2006; 4:729–40. [PubMed: 16980936]
- Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*. 1999; 15:446–54. [PubMed: 10383470]
- Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucl Acids Res*. 2003; 31:3423–8. [PubMed: 12824339]
- Lu Y, Turner RJ, Switzer RL. Function of RNA secondary structures in transcriptional attenuation of the bacillus subtilis pyr operon. *PNAS*. 1996; 93:14462–7. [PubMed: 8962074]
- Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*. 2004; 10:1178–1190. [PubMed: 15272118]
- McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*. 1990; 29:1105–19. [PubMed: 1695107]
- Miklos I, Meyer IM, Nagy B. Moments of the boltzmann distribution for RNA secondary structures. *Bull Math Biol*. 2005; 67:103147.
- Morris, KV. *RNA and the Regulation of Gene Expression: A Hidden Layer of Complexity*. Caister Academic Press; 2008.
- Morris, KV. *Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection*. Caister Academic Press; 2012.
- Niranjanakumari S, Stams T, Crary SM, Christianson DW, Fierke CA. Protein component of the ribozyme ribonuclease p alters substrate recognition by directly contacting precursor tRNA. *Proc Natl Acad Sci U S A*. 1998; 95:15212–7. [PubMed: 9860948]
- Ponty Y, Termier M, Denise A. GenRgenS: software for generating random genomic sequences and structures. *Bioinformatics*. 2006; 22:1534–5. [PubMed: 16574695]
- Quarta G, Kim N, Izzo JA, Schlick T. Analysis of riboswitch structure and function by an energy landscape framework. *J Mol Biol*. 2009; 393:993–1003. [PubMed: 19733179]
- Repoila F, Darfeuille F. Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biol. Cell*. 2009; 101:117–131. [PubMed: 19076068]
- Scarabino D, Crisari A, Lorenzini S, Williams K, Tocchini-Valentini GP. trna prefers to kiss. *EMBO J*. 1999; 18:4571–8. [PubMed: 10449422]
- Schneider T, Stephens R. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*. 1990; 18:6097–100. [PubMed: 2172928]
- Shannon C. A mathematical theory of communication. *Bell System Technical Journal*. 1948; 27:379–423.
- Shaw TI, Manzour A, Wang Y, Malmberg RL, Cai L. Analyzing modular RNA structure reveals low global structural entropy in microRNA sequences. *J Bioinform Comput Biol*. 2011; 9:283–98. [PubMed: 21523933]
- Simmonds P, Karakasiliotis I, Bailey D, Chaudhry Y, Evans DJ, Goodfellow IG. Bioinformatic and functional analysis of RNA secondary structure elements among different genera of human and animal caliciviruses. *Nucleic Acids Res*. 2008; 36:2530–46. [PubMed: 18319285]
- Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. Non-coding RNAs: regulators of diseases. *Journal of Pathology*. 2010; 220:126–13. [PubMed: 19882673]
- Tinoco IJ, Bustamante C. How RNA folds. *J Mol Biol*. 1999; 293:271–81. [PubMed: 10550208]
- Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet*. 2004; 20:44–50. [PubMed: 14698618]
- Wang Y, Manzour A, Shareghi P, Shaw TI, Li YW, Malmberg RL, Cai L. Stable stem enabled shannon entropies distinguish non-coding RNAs from random backgrounds. *BMC Bioinformatics*. 2012; 13(Suppl 5):S1.

- Westhof E, Masquida B, Jossinet F. Predicting and modeling RNA architecture. CSH Perspectives 3. 2011
- Yockey, HP. Information theory, evolution, and the origin of life. Cambridge University Press; 2005.
- Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31:3406–15. [PubMed: 12824337]
- Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981; 9:133–48. [PubMed: 6163133]

\$watermark-text

\$watermark-text

\$watermark-text

Highlights

- Shannon's Entropy of RNA secondary-structure was derived and implemented under SCFG.
- MicroRNA has a low structural entropy, confirming their more deterministic fold.
- Signs of significantly high structural diversity observed for certain classes of RNA.
- Both unusually high and low structural diversity may be signals indicative of RNA function.

Figure 1.

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.

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.

Figure 3.

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).

Figure 4.

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 5. 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 6.

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 B.1 for information about sequences.

Figure 7. Structural entropy p-values of short ncRNAs sequences against random sequences with structure (See Appendix E 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 F.4).

Figure 8. 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.

Figure 9. 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.

Figure 10.

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 Appendix E for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See Appendix D for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.

Figure 11.

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 Appendix E for details about sequences and clusters.). P-values calculated empirically by comparing with random sequences with structure (See Appendix D for details about generating random structures for each model). 2-order polynomial trendline of p-values are shown for each grammar model.