Base-pair Ambiguity and the Kinetics of RNA Folding

Guangyao Zhou[1*], Jackson Loper[2], Matthew T. Harrison[1], Stuart Geman[1]

1 Division of Applied Mathematics, Brown University, Providence, RI, USA
2 Data Science Institute, Columbia University, New York, NY, USA

* guangyao_zhou@brown.edu

Introduction

Discoveries in recent decades have established a wide range of biological roles served by RNA molecules, in addition to their better-known role as carriers of the coded messages that direct ribosomes to construct specific proteins. Non-coding RNA molecules participate in gene regulation, DNA and RNA repair, splicing and self-splicing, catalysis, protein synthesis, and intracellular transportation [1, 2]. To understand the mechanisms of these actions, emphasis has to be placed on the native structures, both secondary and tertiary. Our focus here will be on secondary structure, often a vital intermediary, and a useful abstraction for understanding the functions of non-coding RNA molecules [3].

Because of the time-consuming nature of experimental determination of RNA structures, a considerable amount of work has been put into computational (as opposed to experimental) approaches. For secondary structure prediction, comparative analysis is the gold standard [4]. But accurate comparative analysis typically requires a large number of homologous sequences, which may or may not be available. On the other hand, the prevailing *computational* approaches XXX References??? XXX are based on an important and controversial assumption. Namely, that the structures of non-coding RNA molecules, *in vivo*, are in thermal equilibrium. If this were the case, and if the configuration energies could be accurately specified, then secondary (and in principle, even tertiary) structures could be explored through Monte Carlo sampling of the Gibbs equilibrium distribution. (XXX or is it the Boltzmann equilibrium??? XXX) Alternatively, depending on the nature of the energy landscape, secondary structures could be approximated by identifying the minimum free energy (MFE) configuration [5, 6]. However, even if we put aside the necessity for approximations, the biological relevance of equilibrium configurations has been a source of misgivings at least since 1969, when Levinthal pointed out that the time required to equilibrate might be too long by many orders of magnitude [7]. In light of these observations, and considering the "frustrated" nature of the folding landscape, many have argued that when it comes to structure prediction for macromolecules, kinetic accessibility is more relevant than equilibrium thermodynamics [3, 8, 9].

The primary structures of RNA molecules typically afford many opportunities to form short or medium-length stems[1], most of which do not participate in the native structure. This situation not only makes it hard for the biologist to accurately predict secondary structure, but equally challenges the molecule to avoid these "energetic traps." Once formed, they require a large amount of energy (not to mention time) to be unformed. By what mechanisms might these traps be discouraged, if not actually avoided? Ideally, the only stems available in the primary structure would be those that participate in the secondary structure, but this is obviously too restrictive for all but

[1] By which we will mean sequences of G·U ("wobble pairs") and/or Watson-Crick pairs.

the shortest or simplest of the structural RNA molecules. Still, there can be little doubt that the necessity for repeatable and efficient folding produces a selective pressure against the more disruptive ambiguities—e.g. a sequence that can form half of either of two stems, both of which are kinetically accessible and stable, but only one of which is native, or a sequence that is unpaired in the native structure but can nonetheless form a stem that is stereochemically inconsistent with the native structure. By this reasoning, which emphasizes kinetics rather than equilibrium, *per se*, we might expect to find information in the intra-molecular ambiguities about the folding process, or even the native structure.

We will introduce quantitative measures of ambiguity and demonstrate their statistical relationships to native secondary structure and to qualitative distinctions between RNA families. Using a somewhat arbitrary convention, we will refer to sequences of length four (four consecutive nucleotides) as *segments*. Keeping this convention in mind, each line of statistical evidence is based on the same formal definition of the "local ambiguity" of a given segment. This is simply the number of its complementary pairs in the molecule. When we refer to the *location* of a segment, we will mean the location of the first element of the segment, counting from $5'$ to $3'$, and when we refer to the local ambiguity of a location, we will mean the local ambiguity of the segment at that location. Local ambiguity is an intrinsic property of the primary structure. We will be interested in exposing its statistical relationships to any given candidate secondary structure of the same molecule. For this purpose, for any particular secondary structure we distinguish three different kinds of locations:

***Single*:** Locations where all nucleotides in the corresponding segment are unpaired in the secondary structure;

***Double*:** Locations where all nucleotides in the corresponding segment are paired in the secondary structure;

***Transitional*:** Locations where some nucleotides in the corresponding segment are paired and others are unpaired in the secondary structure.

Double and *transitional* locations participate in the candidate secondary structure, while *single* locations do not. As a general rule, local ambiguities measured at any of these locations will increase with the length of the molecule—there are simply more opportunities to find complementary sequences. Instead of using the local ambiguities themselves, we focus on the *differences* between ambiguities in and around stems (*double* and *transitional* locations) from those at unpaired (*single*) locations. In particular, we defined the "T-S ambiguity index" to be the difference between the average local ambiguity at *transitional* locations minus the average at *single* locations. We defined the "D-S ambiguity index" similarly, but adjusted for the evident bias at *double* locations, each of which, after all, has at least one complementary pair somewhere in the molecule. The D-S ambiguity index, then, is the average local ambiguity at *double* locations minus the average over those *single* locations that have at least one ambiguity.

Using only primary and (comparative-analysis) secondary structures, elementary counting statistics, and exact, distribution-free, tests, we will give evidence that both the T-S and D-S ambiguity indexes significantly separate two groups of non-coding RNA molecules: one group consists of RNA families that operate, *in vivo*, as single entities—the Group I and Group II Introns; the other group is made up of RNA families known to be active as protein-RNA complexes (i.e. as ribonculeoproteins)—the transfer-messenger RNAs (tmRNA), the RNAs of signal recognition particles (SRP RNA), the ribonuclease P family (RNase P), and the 16s and 23s ribosomal RNAs (16s and 23s rRNA). In particular, "unbound" RNA molecules, which perform their functions without being part of a larger nucleoprotein complex, have systematically lower T-S and

D-S ambiguity indexes than the “bound” RNA molecules found in ribonucleoproteins. There are many possible explanations—see *Discussion*. Possibly, the unbound molecules are more sensitive to ambiguities and their energy traps, especially those ambiguities that involve the structurally critical *double* and *transitional* regions, then the bound molecules, whose secondary structures may well be influenced by their chemical relationships to proteins. Interestingly, this distinction between the ambiguity indexes of bound and unbound molecules largely disappears when MFE structures are used instead of comparative-analysis structures in defining *double*, *transitional*, and *single* locations. In fact, for most unbound molecules we can classify a candidate secondary structure (was it derived from comparative analysis or by a minimum-free-energy calculation?) just by looking at the difference in the T-S or the D-S index under the two structures.

The *Results* section is organized as follows: we first develop some basic notation and definitions, and then present an exploratory and largely informal statistical analysis. This is followed by formal results comparing ambiguities in unbound versus bound molecules, and then by a comparison of the ambiguities implied by secondary structures derived from comparative analyses to those derived through minimization of free energy. The *Results* section is followed by a *Discussion*, and then by the section on *Materials and Methods*, which, among other things, includes detailed information about the data and its (open) source, as well as links to code that can be used to reproduce our results or for further experimentation.

Results

Basic Notation and Definitions

Consider a non-coding RNA molecule with N nucleotides. Counting from $5'$ to $3'$, we denote the primary structure by

$$p = (p_1, p_2, \cdots, p_N), \text{where } p_i \in \{A, G, C, U\}, i = 1, \cdots, N \tag{1}$$

and the secondary structure by

$$s = \{(j, k) : \text{nucleotides } j \text{ and } k \text{ are paired, } 1 \leq j < k \leq N\} \tag{2}$$

and we define the *segment* at *location* i to be

$$P_i = (P_{i,1}, P_{i,2}, P_{i,3}, P_{i,4}) = (p_i, p_{i+1}, p_{i+2}, p_{i+3}) \tag{3}$$

There is no particular reason for using segments of length four, and in fact all qualitative conclusions are identical with segment lengths three, four, or five, and, for that matter, whether or not we include the G·U wobble pairs. We chose to include them.

Which segments are viable complementary pairs to P_i? The only constraint on location is that an RNA molecule cannot form a loop of two or fewer nucleotides. Let A_i be the set of all segments that are potential pairs of P_i:

$$A_i = \{P_j : 1 \leq j \leq i - 7 \text{ or } i + 7 \leq j \leq N - 3\} \tag{4}$$

We can now define the *local ambiguity function*,

$$a(p) = (a_1(p), \cdots, a_{N-3}(p))$$

which is a vector-valued function of the primary structure p. The vector has one component, $a_i(p)$, for each segment P_i, which is, simply, the number of feasible segments that are complementary to P_i:

$$\begin{aligned} a_i(p) &= \#\{P \in A_i : P \text{ and } P_i \text{ are complementary}\} \\ &= \#\big\{P_j \in A_i : (P_{i,k}, P_{j,5-k}) \in \{(A,U),(U,A),(G,C),(C,G),(G,U),(U,G)\}, \\ &\qquad k = 1, \ldots, 4\big\} \end{aligned} \tag{5}$$

We want to explore the relationship between ambiguity and secondary structure. We can do this conveniently, on a molecule-by-molecule basis, by introducing another vector-valued function, this time depending only on a purported secondary structure. Specifically, the new function assigns a descriptive label to each location (i.e. each nucleotide), determined by whether the segment at the given location is fully paired, partially paired, or fully unpaired.

Formally, given a secondary structure s, as defined in equation (2), and a location $i \in \{1, 2, \ldots, N-3\}$, let $f_i(s)$ be the number of nucleotides in P_i that are paired under s:

$$f_i(s) = \#\{j \in P_i : (j,k) \in s \text{ or } (k,j) \in s, \text{ for some } 1 \leq k \leq N\} \tag{6}$$

Evidently, $0 \leq f_i(s) \leq 4$. The "paired nucleotides function" is then the vector-valued function of secondary structure defined as $f(s) = (f_1(s), \ldots, f_{N-3}(s))$. Finally, we use f to distinguish three types of locations (and hence three types of segments): location i will be labeled

$$\begin{cases} \textit{single} \text{ if } f_i(s) = 0 \\ \textit{double} \text{ if } f_i(s) = 4 \\ \textit{transitional} \text{ if } 0 < f_i(s) < 4 \end{cases} \qquad i = 1, 2, \cdots, N-3 \tag{7}$$

A First Look at the Data: Shuffling Nucleotides

Our goal is to explore connections between ambiguities and basic characteristics of RNA families, as well as the changes in these relationships, if any, when using comparative, as opposed to MFE, secondary structures. For each molecule and each location i, the segment at i has been assigned a "local ambiguity" $a_i(p)$ that depends only on the primary structure, and a label (*single*, *double*, or *transitional*) that depends only on the secondary structure. Since the local ambiguity, by itself, is strongly dependent on the length of the molecule, and possibly on other intrinsic properties, we proposed two *relative* ambiguity indexes: T-S and D-S, each of which depends on both the primary (p) and purported secondary (s) structures. Formally,

$$d_{\text{T-S}}(p,s) = \frac{\sum_{j=0}^{N-3} a_j(p) c_j^{\text{tran}}(s)}{\sum_{j=0}^{N-3} c_j^{\text{tran}}(s)} - \frac{\sum_{j=0}^{N-3} a_j(p) c_j^{\text{single}}(s)}{\sum_{j=0}^{N-3} c_j^{\text{single}}(s)} \tag{8}$$

where we have used c_i^{double} and c_i^{single} for indicating whether location i is *transitional* or *single* respectively. In other words, for each $i = 1, 2, \ldots, N-3$

$$c_i^{\text{tran}} = \begin{cases} 1, & \text{if location } i \text{ is } \textit{transitional} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$c_i^{\text{single}} = \begin{cases} 1, & \text{if location } i \text{ is } \textit{single} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

In short, the T-S ambiguity index is the difference in the averages of the local ambiguities at *transitional* sites and *single* sites. Since, as noted earlier, the local ambiguity at every *double* location, i, is at least one ($a_i \geq 1$), the D-S index involves only those *single* locations, j, for which it is also the case that $a_j \geq 1$:

$$d_{\text{D-S}}(p,s) = \frac{\sum_{j=0}^{N-3} a_j(p) c_j^{\text{double}}(s)}{\sum_{j=0}^{N-3} c_j^{\text{double}}(s)} - \frac{\sum_{j=0}^{N-3} a_j(p) \hat{c}_j^{\text{single}}(p,s)}{\sum_{j=0}^{N-3} \hat{c}_j^{\text{single}}(p,s)} \tag{11}$$

Here, $\hat{c}^{\text{single}}$, which evidently depends on both primary and secondary structure, indicates those *single* locations with at least one pair: for each $i = 1, 2, \ldots, N-3$

$$\hat{c}_i^{\text{single}} = \begin{cases} 1, & \text{if location } i \text{ is } \textit{single} \text{ and } a_i(p) \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Thinking kinetically, we might expect to find relatively small values of the T-S and D-S indexes ($d_{\text{T-S}}$ and $d_{\text{D-S}}$) in RNA families which fold and operate as individual entities—what we call here the unbound RNA. One reason for believing this is that larger numbers of partial matches for a given sequence in or around a stem would likely interfere with the *nucleation* of the native stem structure, potentially disrupting folding, and nucleation appears to be a critical and perhaps even rate-limiting step in folding. Indeed, the experimental literature [10–13] has long suggested that stem formation in RNA molecules is a two-step process. When forming a stem, there is usually a slow nucleation step, resulting in a few consecutive base pairs at a nucleation point, followed by a fast zipping step. Since the ambiguity indexes are functions of the secondary structure, s, our expectation are made, implicitly, under the assumption that s is correct. For the time being we will focus only on comparative structures, returning later to the questions about MFE structures raised in the *Introduction*.

How are we to gauge $d_{\text{T-S}}$ and $d_{\text{D-S}}$, and how are we to compare their values across different RNA families? Consider the following experiment: for a given RNA molecule we create a "surrogate" which has the same nucleotides, and in fact the same counts of *all* four-tuple segments as the original molecule, but is otherwise ordered randomly. If ACCU appeared eight times in the original molecule, then it appears eight times in the surrogate, and the same can be said of all sequences of four successive nucleotides—the frequency of each of the 4^4 possible segments is preserved in the surrogate. If we also preserve the locations of the *transitional*, *double*, and *single* labels, then we can compute new values for $d_{\text{T-S}}$ and $d_{\text{D-S}}$, say $\tilde{d}_{\text{T-S}}$ and $\tilde{d}_{\text{D-S}}$, from the surrogate. If we produce many surrogate sequences then we will get a distribution of surrogate values $\tilde{d}$, one for each of the two ambiguity indexes, to which we can compare d. We made several experiments of this type—for each of the two indexes, T-S and D-S, within each of the seven RNA families (Group I and Group II Introns, tmRNA, SRP RNA, RNase P, and 16s and 23s rRNA).

To make this precise, consider an RNA molecule with primary structure p and comparative secondary structure s. Construct a segment "histogram function," $\mathcal{H}(p)$, which is the vector of the number of times that each of the 4^4 possible segments appears in p.[2]. Let $\mathcal{E}(p)$ be the set of all permutations of the ordering of nucleotides in p that preserve the frequencies of four-tuples:

$$\mathcal{E}(p) = \{p' : \mathcal{H}(p') = \mathcal{H}(p)\}$$

Clever algorithms exist for efficiently drawing independent samples from the uniform distribution on $\mathcal{E}$—see [14–16]. Let $p^{(1)}, \ldots, p^{(K)}$ be K such samples, and let $d_{\text{T-S}}\big(p^{(1)}, s\big), \ldots, d_{\text{T-S}}\big(p^{(K)}, s\big)$ and $d_{\text{D-S}}\big(p^{(1)}, s\big), \ldots, d_{\text{D-S}}\big(p^{(K)}, s\big)$ be the corresponding T-S and D-S ambiguity indexes. The results are not vary sensitive to K, nor to the

[2] $\mathcal{H}$ is made a vector by fixing an arbitrary ordering of all possible segments.

particular sample, provided that K is large enough. We used K =10,000. Finally, let $\alpha_{\text{T-S}}(p,s) \in [0,1]$ and $\alpha_{\text{D-S}}(p,s) \in [0,1]$ be the left-tail empirical probabilities, under the distributions defined by the ensembles

$$d_{\text{T-S}}(p,s), d_{\text{T-S}}\big(p^{(1)},s\big), \dots, d_{\text{T-S}}\big(p^{(K)},s\big) \tag{13}$$

$$\text{and} \qquad d_{\text{D-S}}(p,s), d_{\text{D-S}}\big(p^{(1)},s\big), \dots, d_{\text{D-S}}\big(p^{(K)},s\big) \tag{14}$$

of choosing an ambiguity index less than or equal to $d_{\text{T-S}}(p,s)$ and $d_{\text{T-S}}(p,s)$, respectively:

$$\alpha_{\text{T-S}}(p,s) = \frac{1 + \#\{k \in \{1,\dots,K\} : d_{\text{T-S}}\big(p^{(k)},s\big) \le d_{\text{T-S}}(p,s)\}}{1+K} \tag{15}$$

$$\alpha_{\text{D-S}}(p,s) = \frac{1 + \#\{k \in \{1,\dots,K\} : d_{\text{D-S}}\big(p^{(k)},s\big) \le d_{\text{D-S}}(p,s)\}}{1+K} \tag{16}$$

In essence, each α score is a self-calibrated ambiguity index.

It is tempting to interpret $\alpha_{\text{T-S}}(p,s)$ as a p-value from a conditional hypothesis test: Given s and $\mathcal{H}$, test the null hypothesis that $d_{\text{T-S}}(p,s)$ is statistically indistinguishable from $d_{\text{T-S}}(p',s)$, where p' is a random sample from $\mathcal{E}$. If the alternative hypothesis were that $d_{\text{T-S}}(p',s)$ is too small to be consistent with the null, then the null is rejected in favor of the alternative with probability $\alpha_{\text{T-S}}(p,s)$. The problem with this interpretation, and the analogous one for $\alpha_{\text{D-S}}(p,s)$, is that this null hypothesis violates the simple observation that given $\mathcal{H}$ there is information in s about p, whereas $p^{(1)},\dots,p^{(K)}$ are independent of s given $\mathcal{H}$. A larger problem is that there is no reason to believe the alternative; we are more interested in *relative* than absolute ambiguity indexes. Thinking of $\alpha_{\text{T-S}}(p,s)$ and $\alpha_{\text{D-S}}(p,s)$ as a calibrated, intra-molecular indexes, we want to know how they vary across RNA families, and whether these variations depend on the differences between comparative and MFE structures.

Nevertheless, $\alpha_{\text{T-S}}(p,s)$ and $\alpha_{\text{D-S}}(p,s)$ are useful statistics for exploratory analysis. Table 1 provides summary data about the α scores for each of the seven RNA families. For each molecule in each family we use the primary structure and the comparative secondary structure, and K =10,000 samples from $\mathcal{E}$, to compute individual T-S and D-Sα scores (Eqs 15 and 16). Keeping in mind that a smaller value of α represents a smaller *calibrated* value of the corresponding ambiguity index, $d(p,s)$, there is evidently a disparity between ambiguity indexes of RNA molecules that form ribonucleoproteins and those of RNA that are already active as individual molecules. As a group, the unbound molecules have systematically lower ambiguity indexes. As already noted, this observation is consistent with, and in fact anticipated by, the kinetic point of view. Shortly, we will support this observation with ROC curves and rigorous hypothesis tests.

Does the MFE structure similarly separate single-entity RNA molecules from those that form ribonucleoproteins? A convenient way to explore this question is to recalculate and recalibrate the ambiguity indexes of each molecule in each of the seven families, but using the MFE in place of the comparative secondary structures. The results are summarized in Table 2. By comparison to the results shown from Table 1, the separation of unbound from bound molecules nearly disappears under the MFE secondary structures. Possibly, the comparative structures, as opposed to the MFE structures, better anticipate the need to avoid energy traps in the folding landscape. Here too we will revisit the data using ROC curves and proper hypothesis tests.

Formal Statistical Analyses

The T-S and D-S ambiguity indexes ($d_{\text{T-S}}(p,s)$ and $d_{\text{D-S}}(p,s)$) are intra-molecular measures of the extent to which segments within single-stranded regions have more

family	number molecules	median length	median $\alpha_{\text{T-S}}$	median $\alpha_{\text{D-S}}$
Group I Introns	116	451	0.432	0.908
Group II Introns	34	990	0.181	0.761
tmRNA	404	363	0.926	0.988
SRP RNA	346	274	0.790	0.967
RNase P	407	330	0.925	0.985
16s rRNA	279	1512	0.938	1.000
23s rRNA	59	2913	1.000	1.000

Table 1. Comparative Secondary Structures: calibrated ambiguity indexes, by RNA family. The number of molecules, the median length (number of nucleotides), and the median α scores for T-S and D-S ambiguity indexes (Eqs 15 and 16) for each of the seven RNA families studied. RNA molecules from the first two families are active as single molecules (unbound); the remaining five are bound in ribonucleoproteins. Unbound RNA molecules have lower ambiguity indexes.

family	number molecules	median length	median $\alpha_{\text{T-S}}$	median $\alpha_{\text{D-S}}$
Group I Introns	116	451	0.833	0.994
Group II Introns	34	990	0.841	0.997
tmRNA	404	363	0.867	0.984
SRP RNA	346	274	0.803	0.953
RNase P	407	330	0.955	0.998
16s rRNA	279	1512	0.982	1.000
23s rRNA	59	2913	1.000	1.000

Table 2. MFE Secondary Structures: calibrated ambiguity indexes, by RNA family. Identical to Table 1, except that the ambiguity indexes and their calibrations are calculated using the MFE secondary structures rather than those derived from comparison analyses. There is little evidence in the MFE secondary structures for lower ambiguity indexes among unbound RNA molecules.

Watson-Crick and wobble pairings than segments that lie in and around sub-sequences that are double-stranded. As such, they depend on both p and any purported secondary structure, s. Based on calibrated versions of these indexes ($\alpha_{\text{T-S}}(p,s)$ and $\alpha_{\text{D-S}}(p,s)$) and employing the comparative secondary structure for s we found support for the idea that non-coding RNA molecules which are active as individual entities are more likely to have small ambiguity indexes than RNA molecules destined to operate as part of ribonculeoproteins. Furthermore, the difference appears to be sensitive to the approach used for identifying secondary structure—there is little, if any, evidence in the MFE secondary structures for lower ambiguities among unbound molecules.

These qualitative observations can be used to formulate precise statistical hypothesis tests. Many tests come to mind, but perhaps the simplest and most transparent are based on nothing more than the molecule-by-molecule signs of the ambiguity indexes. Whereas ignoring the actual values of the indexes is inefficient in terms of information, an possibly also in the strict statistical sense, tests based on signs require very few assumptions and are, therefore, more robust to model misspecification. All of the p-values that we will report are based on the hypergeometric distribution, which arises as follows.

We are given a population of M molecules, each with a binary outcome measure $B_m \in \{-1,+1\}$, $m = 1,\ldots,M$. There are two subpopulations of interest: the first M_1

molecules, making up population 1, and the next M_2 molecules, which make up population 2. (So $M_1 + M_2 = M$.) We observe n_1 plus values in population 1 and n_2 in population 2

$$n_1 = \#\big\{m \in \{1, 2, \ldots, M_1\} : B_m = +1\big\} \tag{17}$$

$$n_2 = \#\big\{m \in \{M_1 + 1, M_1 + 2, \ldots, M\} : B_m = +1\big\} \tag{18}$$

We suspect that population 1 has less than its share of plus ones, meaning that the $n_1 + n_2$ population of plus ones was not randomly distributed among the M molecules. To be precise, let N be the number of plus ones that appear from a draw, without replacement, of M_1 samples from $B_1, \ldots, B_M$. Under the null hypothesis, H_o, n_1 is a sample from the hypergeometric distribution on N:

$$\mathbb{P}\{N = n\} = \frac{\binom{M_1}{n}\binom{M_2}{n+1+n_2-n}}{\binom{M}{n_1+n_2}} \quad \max\{0, n_1 + n_2 - M_2\} \leq n \leq \min\{n_1 + n_2, M_1\} \tag{19}$$

The alternative hypothesis, H_a, is that n_1 is too small to be consistent with H_o, leading to a left-tail test with p-value $\mathbb{P}\{N \leq n_1\}$ (which can be computed directly or using a statistical package, e.g. *hygecdf* in MatLab).

It is by now well recognized that p-values should never be the end of the story. One reason is that *any* departure from the null hypothesis in the direction of the alternative, no matter how small, is doomed to be statistically significant, with arbitrarily small p-value, once the sample size is sufficiently large. In addition to reporting p-values, we will also display estimated ROC curves, summarizing performance of a related classification problem. There are eight such problems: Under each of the comparative and MFE secondary structures, how well can we use $d_{\text{T-S}}(p, s)$ or $d_{\text{D-S}}(p, s)$ to classify a given molecule as bound or unbound? And, given $d_{\text{T-S}}(p, s)$ or $d_{\text{D-S}}(p, s)$ for a molecule that is known to be bound or unbound, how well can we classify the secondary structure, s, as coming from a comparative analysis versus an MFE analysis?

Single-entity RNA Molecules versus Protein-RNA Complexes

Consider an RNA molecule, m, selected from one of the seven families in our data set, with primary structure p and secondary structure s, computed by comparative analysis. Given only the T-S ambiguity index of m (i.e. given only $d_{\text{T-S}}(p, s)$), how accurately could we classify m as being unbound (i.e. from the Group I or Group II Introns) as opposed to bound (i.e. from one of the five families tmRNA, SRP RNA, RNase P, 16s rRNA or 23s rRNA)? The foregoing exploratory analysis suggests constructing a classifier that declares a molecule to be 'unbound' when $d_{\text{T-S}}(p, s)$ is small, e.g. $d_{\text{T-S}}(p, s) < t$, where the threshold t governs the familiar trade off between rates of "true positives" (an unbound m is declared 'unbound') and "false positives" (a bound m is declared 'unbound'). Large values of t favor low rates of false positives at the price of low rates of true positives, whereas small values of t favor high rates of true-positives at the price of high rates of false positives. Since for each molecule m we have both the correct classification (unbound or bound) and the statistic d, we can estimate the ROC performance of our threshold classifier by plotting the empirical values of the pair, (# false positives,# true positives), for each value of t. The ROC curve for the two-category (unbound versus bound) classifier based on thresholding $d_{\text{T-S}}(p, s) < t$ is shown in the upper-left panel of Figure 1. Also shown is the estimated area under the curve (AUC=0.81), which has a convenient and intuitive interpretation, as it is equal to the probability that for two randomly selected molecules, m from the unbound population and m' from the bound population, the T-S ambiguity index of m will be smaller than the T-S ambiguity index of m'.

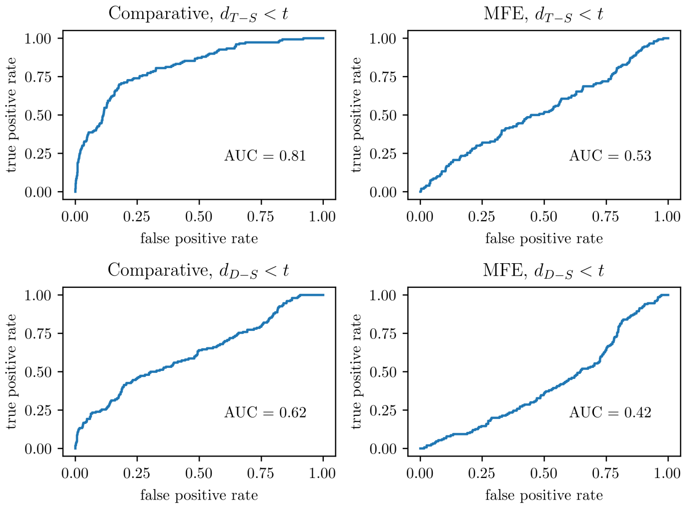

Fig 1. Bound or Unbound? ROC performance of classifiers based on thresholding T-S and D-S ambiguity indexes. Small values are taken as evidence for molecules that are active as single entities (unbound), as opposed to parts of ribonucleoproteins (bound). Classifiers in the left two panels use comparative secondary structures to compute ambiguity indexes; those on the right use (approximate) minimum free energies. In each of the four experiments, a conditional p-value was also calculated, based only on the signs of the indexes and the null hypothesis that positive indexes are distributed randomly among molecules of all types as opposed to the alternative that positive indexes are more typically found among families of bound RNA. Under the null hypothesis, the test statistic is hypergeometric—see Eq 19. *Upper Left:* $p = 1.2 \times 10^{-34}$; *Lower Left:* $p = 7.3 \times 10^{-8}$; *Upper Right:* $p = 0.02$; *Lower Right:* $p = 0.92$.

As mentioned earlier, we can also associate a traditional p-value to the problem of separating unbound from bound molecules, based on the T-S ambiguity indexes. We consider only the signs (positive or negative) of these indexes, and then test whether there are fewer than expected positive indexes among the unbound, as opposed to the bound, population. This amounts to computing $\mathbb{P}\{N \leq n_1\}$ from the hypergeometric distribution—Eq (19). The relevant statistics can be found in Table 3, under the column labels **#mol's** and $\#d_{\textbf{T-S}} > 0$. Specifically, $M_1 = 116 + 34 = 150$ (number of unbound molecules), $M_2 = 404 + 346 + 407 + 279 + 59 = 1495$ (number of bound molecules), $n_1 = 50 + 8 = 58$ (number of positive T-S indexes among unbound molecules) and $n_2 = 368 + 269 + 379 + 210 + 53 = 1279$. The resulting p-value, $1.2 \cdot 10^{-34}$, is essentially zero, meaning that the positive T-S indexes are not distributed proportional to the sizes of the unbound and bound populations, which is by now obvious, in any case. To repeat our caution, small p-values conflate sample size with effect size, and for that reason we have chosen additional ways, using permutations as well as classifications, to look at the data.

The comparative secondary structure of an RNA molecule, when combined with its primary structure, can be used to construct a measure—the T-S ambiguity index—which distinguishes unbound from bound molecules with good accuracy. Can the same can be said for the D-S index? Yes, albeit with lower accuracy. To demonstrate, we followed the identical procedure, except that we assigned the index $d_{\text{D-S}}(p, s)$ rather than $d_{\text{T-S}}(p, s)$ to each molecule. The ROC curve (with area 0.62) is shown in the lower-left panel of Figure 1. The hypergeometric test, based on the

family	#mol's	#$d_{\mathbf{T\text{-}S}} > 0$	#$d_{\mathbf{D\text{-}S}} > 0$	#$d_{\tilde{\mathbf{T}}\text{-}\tilde{\mathbf{S}}} > 0$	#$d_{\tilde{\mathbf{D}}\text{-}\tilde{\mathbf{S}}} > 0$
Group I Introns	116	50	100	94	114
Group II Introns	34	8	21	27	33
tmRNA	404	368	396	358	396
SRP RNA	346	269	314	264	302
RNase P	407	379	393	377	404
16s rRNA	279	210	251	254	278
23s rRNA	59	53	56	54	58

Table 3. Numbers of Positive Ambiguity Indexes, by family. #mol's: number of molecules; **#$d_{\mathbf{T\text{-}S}} > 0$** and **#$d_{\mathbf{D\text{-}S}} > 0$**: numbers of positive T-S and D-S ambiguity indexes, secondary structures computed by *comparative analysis*; **#$d_{\tilde{\mathbf{T}}\text{-}\tilde{\mathbf{S}}} > 0$** and **#$d_{\tilde{\mathbf{D}}\text{-}\tilde{\mathbf{S}}} > 0$**: numbers of positive T-S and D-S ambiguity indexes, secondary structures computed by *minimum free energy*.

thresholded (signed) values of $d_{\text{D-S}}$ (positive counts, for each RNA family, can be found in Table 3) has p-value $7.3 \cdot 10^{-8}$. Finally, the right-hand panels in Figure 1 mirror the left-hand panels, except that all secondary structures were computed by (approximately) maximizing free energy rather than comparative analysis. The classification results are substantially less convincing and the p-values substantially higher.

Comparative Analysis versus Minimum Free Energy

As we have just seen, ambiguity indexes based on MFE secondary structures, as opposed to comparative secondary structures, do not make the same stark distinction between single and bound RNA molecules. To explore this a little further, we can turn the analyses of the previous paragraphs around and ask to what extent knowledge of the group of a molecule (unbound or bound) and its ambiguity index (e.g. the T-S ambiguity index) is sufficient to predict the source of the secondary structure—comparative or free energy?

Interestingly, there is a sharp difference in predicting the source of secondary structure in unbound molecules as opposed to bound molecules. Consider the top two ROC curves in Figure 2. In each of the two experiments a classifier was constructed by thresholding the T-S ambiguity index, declaring the secondary structure to be 'comparative' when $d_{\text{T-S}}(p, s) < t$ and 'MFE' otherwise, since, as we have seen, smaller values of the ambiguity indexes are generally associated with the comparative secondary structure. The ROC curves are then swept out by varying t. The difference between the two panels is in the population used for the classification experiments—unbound molecules in the upper-left and bound molecules in the upper-right. In unbound molecules, small values of the T-S ambiguity index are a good indication that a secondary structure was derived from comparative rather than free-energy analysis, but not in bound molecules. The corresponding hypothesis tests seek evidence against the null hypotheses that in a given group (unbound or bound) the set of positive T-S ambiguity indexes ($d_{\text{T-S}}(p, s) > 0$) are equally distributed between the comparative and free-energy derived indexes, and in favor of the alternatives that the T-S ambiguity indexes are less typically positive for the comparative secondary structures. The necessary data can be found in Table 3. The results are consistent with the classification experiments: the hypergeometric p-value is is $5.4 \cdot 10^{-14}$ for the unbound population and 0.07 for the bound population.

The same experiments, with the same conclusions, were also performed using the D-S ambiguity index, as shown in the bottom two panels of Figure 2, for which the corresponding hypergeometric p-values are $3.8 \cdot 10^{-7}$ (unbound population) and 0.01

(bound population).

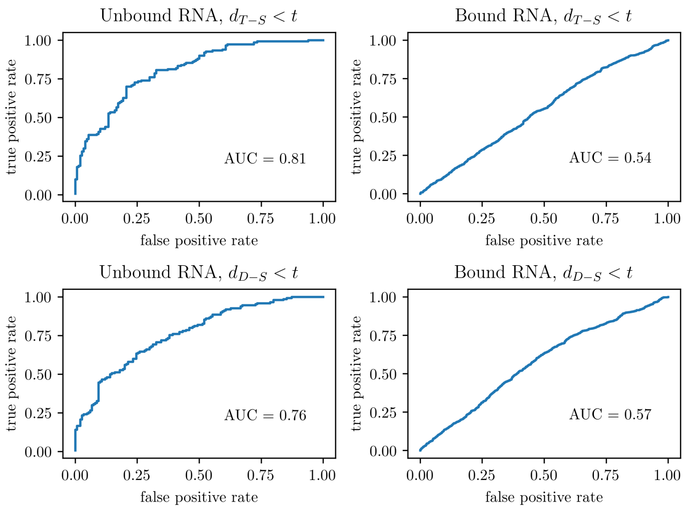

Fig 2. Comparative or MFE? As in Figure 1, each panel depicts the ROC performance of a classifier based on thresholding the T-S (top two panels) or D-S (bottom two panels) ambiguity indexes. Here, small values are taken as evidence for comparative as opposed to MFE secondary structure. Either index, T-S or D-S, can be used to construct a good classifier of the origin of a secondary structure for the unbound molecules in our data set (left two panels) but not for the bound molecules (right two panels). Conditional p-values were also calculated, using the hypergeometric distribution and based only on the signs of the indexes. In each case and the null hypothesis is that comparative secondary structures are as likely to lead to positive ambiguity indexes as are MFE structures, whereas the alternative is that positive ambiguity indexes are more typical when derived from MFE structures: *Upper Left:* $p = 5.4 \times 10^{-14}$; *Upper Right:* $p = 0.07$; *Lower Left:* $p = 3.8 \times 10^{-7}$; *Lower Right:* $p = 0.01$.

Discussion

Materials and Methods

Datasets

In this paper, we obtained comparative analysis secondary structures data for seven different families of RNA molecules from the RNA STRAND database [17], a curated collection of RNA secondary structures. These families include: Group I Introns and Group II Introns [18], tmRNAs and SRP RNAs [19], the Ribonuclease P RNAs [20], and 16s rRNAs and 23s rRNAs [18]. Table 4 contains information about the numbers and lengths (measured in nucleotides) of the RNA molecules in each of the seven groups studied. Note that we excluded families like tRNAs, 5s rRNAs and hammerhead ribozymes since most of the molecules in these families are too short to be interesting for our purpose. Also, since we are focusing on comparative analysis secondary structures, we excluded any secondary structures derived using x-ray crystallography or NMR.

family	number	min length	max length	median
Group I Introns	116	210	2630	451
Group II Introns	34	619	2729	990
tmRNA	404	102	437	363
SRP RNA	346	66	533	274
RNase P	407	189	486	330
16s rRNA	279	612	2394	1512
23s rRNA	59	953	4381	2913

Table 4. Data Summary. The seven families of RNA used in the experiments. Table includes the number of molecules in each family, as well as basic statistics about the numbers of nucleotides in the primary sequence of each of the molecules. Data was downloaded from the RNA STRAND database.

Minimum Free Energy Methods

Exact dynamic programming algorithms based on carefully measured thermodynamic parameters are still the prevalent methods for RNA secondary structures prediction. There exist a large number of software packages for the energy minization [21–27]. In this paper, we used the ViennaRNA package [21] to obtain the MFE secondary structures for our statistical analysis.

Reproducing the Results

The results presented in this paper can be easily reproduced. Follow the intrustions on `https://github.com/StannisZhou/rna_statistics`. Here we make a few comments regarding some implementation details.

- In the process of obtaining the data, we used the *bpseq* format, and excluded structures derived from x-ray crystallography or NMR structures, as well as structures for duplicate sequences. Concretely, this means picking a particular type, and select *No* for *Validated by NMR or X-Ray* and *Non-redundant sequences only* for *Duplicates* on the search page of the RNA STRAND database. A copy of the data we used is included in the GitHub repository, but the readers should feel free to try out the analysis on other data.

- When processing the data, we ignored molecules for which we have nucleotides other than *A, G, C, U*, and molecules for which we don't have any base pairs.

- When comparing the local ambiguities in different regions of the RNA molecules, we ignored molecules for which we have empty regions (i.e. at least one of *single, double* and *transitional* is empty), as well as molecules where all local ambiguities in *single* or *double* are 0.

- For shuffling the molecules, we used an efficient and flexible implementation of the Euler algorithm [14–16], called uShuffle [28], which is conveniently available as a python package.

Acknowledgments

Appendix

References

1. Morris KV, Mattick JS. The rise of regulatory RNA. Nat Rev Genet. 2014;15(6):423–437.

2. Kung JTY, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. Genetics. 2013;193(3):651–669.

3. Higgs PG. RNA secondary structure: physical and computational aspects. Q Rev Biophys. 2000;33(3):199–253.

4. Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res. 1992;20(21):5785–5795.

5. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999;288(5):911–940.

6. Zuker M, Mathews DH, Turner DH. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In: Barciszewski J, Clark BFC, editors. RNA Biochemistry and Biotechnology. NATO Science Series. Springer Netherlands; 1999. p. 11–43. Available from: http://link.springer.com/chapter/10.1007/978-94-011-4485-8_2.

7. Levinthal C. How to fold graciously. Mossbauer spectroscopy in biological systems. 1969;67:22–24.

8. Flamm C, Hofacker IL. Beyond energy minimization: approaches to the kinetic folding of RNA. Monatsh Chem. 2008;139(4):447–457.

9. Baker D, Agard DA. Kinetics versus thermodynamics in protein folding. Biochemistry. 1994;33(24):7505–7509.

10. Pörschke D. Model calculations on the kinetics of oligonucleotide double helix coil transitions. Evidence for a fast chain sliding reaction. Biophys Chem. 1974;2(2):83–96.

11. Pörschke D. A direct measurement of the unzippering rate of a nucleic acid double helix. Biophys Chem. 1974;2(2):97–101.

12. Pörschke D. Elementary steps of base recognition and helix-coil transitions in nucleic acids. Mol Biol Biochem Biophys. 1977;24:191–218.

13. Mohan S, Hsiao C, VanDeusen H, Gallagher R, Krohn E, Kalahar B, et al. Mechanism of RNA Double Helix-Propagation at Atomic Resolution. J Phys Chem B. 2009;113(9):2614–2623.

14. Kandel D, Matias Y, Unger R, Winkler P. Shuffling biological sequences. Discrete Appl Math. 1996;71(1):171–185.

15. Fitch WM. Random sequences. J Mol Biol. 1983;163(2):171–176.

16. 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(6):526–538.

17. Andronescu M, Bereg V, Hoos HH, Condon A. RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database. BMC Bioinformatics. 2008;9(1):340.

18. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics. 2002;3:2.

19. Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, Wower IK, et al. The tmRDB and SRPDB resources. Nucleic Acids Res. 2006;34(Database issue):D163–8.

20. Brown JW. The Ribonuclease P Database. Nucleic Acids Res. 1999;27(1):314.

21. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6(1):26.

22. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol. 2008;453:3–31.

23. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010;11:129.

24. Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, et al. NUPACK: Analysis and design of nucleic acid systems. J Comput Chem. 2011;32(1):170–173.

25. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics. 2009;25(4):465–473.

26. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. 2003;31(24):7280–7301.

27. Reeder J, Giegerich R. RNA secondary structure analysis using the RNAshapes package. Curr Protoc Bioinformatics. 2009;Chapter 12:Unit12.8.

28. Jiang M, Anderson J, Gillespie J, Mayne M. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics. 2008;9:192.