Base-pair Ambiguity and the Kinetics of RNA Folding: a Hypothesis-Driven Statistical Analysis

Guangyao Zhou[1*], Jackson Loper[2], 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

1 Abstract

Non-coding RNA molecules contribute to cellular function through diverse roles, including genome regulation, DNA and RNA repair, RNA splicing, catalysis, protein synthesis, and intracellular transportation [1, 2]. The mechanisms of these actions can only be fully understood in terms of native secondary and tertiary structures. When provided with a sufficient number of homologous sequences, the gold standard for secondary structure prediction continues to be comparative analysis [3]. Alternatively, the prevailing *computational* approach to secondary structure is through the Gibbs (thermal) equilibrium, by Monte Carlo sampling or approximating the minimum free energy (MFE) configuration [4, 5]. Aside from the necessary approximations, an enduring debate concerns the biological relevance of equilibrium configurations [6–9]. Here we adopt a kinetic perspective and argue that the existence of reliable folding on biologically relevant time scales suggests an *intra-molecular* statistical relationship between secondary and *primary* structures: as compared with other locations, nucleotide sequences in and around secondary-structure stems will have fewer Watson-Crick matches that are inconsistent with the native structure. An "ambiguity index", one for each pair of molecule and presumed secondary structure, measures the prevalence of false matches and hence the tendency to form metastable structures incompatible with native structures. The ambiguity index statistically separates an ensemble of RNA molecules that operate as single entities (Group I and II Introns) from an ensemble that operates as protein-RNA complexes (SRP and tmRNAs), and ensembles of secondary structures determined by comparative analysis from ones based on thermal equilibrium. We find lower average ambiguity in single-entity RNA's than protein-RNA complexes, and, among single-entity RNA's, lower ambiguity with comparative analyses than equilibrium analyses. Both comparisons are supported by exact and highly significant hypothesis tests. These experiments, motivated by a hypothesized mechanism of folding, and the first of their kind, are consistent with folding to metastable but not necessarily equilibrium structures.

2 Author summary

Recent discoveries indicate that, in addition to being a messenger between DNA and protein, RNA molecules assume a wide range of biological functions. For biological macromolecules, structure is function. Experimental determination of RNA structures is still time-consuming, and computational approaches are of great importance. The prevailing computational approach tries to find the structure with the minimum energy,

yet the relevance of this minimum energy structure as the native structure is still hotly debated. In this paper, we adopt a kinetic perspective, and argue that more emphasis should be placed on the folding process when trying to develop computational methods for RNA structure prediction. We present some statistical analyses using the primary and secondary data (sequence and base-pairs data) of RNA molecules, based on the concept of "local ambiguity", i.e. the molecule's tendency to "make a mistake" at a certain location when forming secondary structures. Our results show the deficiencies of the minimum energy approach, and demonstrate the importance of considering the kinetics as well as protein-RNA interactions in developing computational approaches for RNA secondary structure prediction.

3 Introduction

RNA has long been at the center of molecular biology. However, discoveries in recent decades suggest that RNA molecules take on a wide range of biological roles, in addition to functioning as a messenger between DNA and protein. These non-coding RNA molecules contribute to cellular function through diverse roles, including genome regulation, DNA and RNA repair, RNA splicing, catalysis, protein synthesis, and intracellular transportation [1, 2]. To understand the mechanisms of these actions, emphasis has to be placed on the native secondary and tertiary structures of these RNA molecules. Despite the recent increase in our knowledge of RNA tertiary structure, RNA secondary structure is still of considerable importance, and is a useful abstraction in understanding the functions of non-coding RNA molecules [7].

Because of the time-consuming nature of experimental determination of RNA structures, a considerable amount of work has been put into computational prediction of RNA structures. For secondary structure prediction, when provided with a sufficient number of homologous sequences, the gold standard continues to be comparative analysis [3]. Alternatively, the prevailing *computational* approach to secondary structure is through the Gibbs (thermal) equilibrium, by Monte Carlo sampling or approximating the minimum free energy (MFE) configuration [4, 5]. Aside from the necessary approximations, an enduring debate concerns the biological relevance of equilibrium configurations [6]. People have long argued that, when it comes to structure prediction for macromolecules, we need to consider the kinetics in addition to the thermodynamics [7–9].

In this paper, we adopt a kinetic perspective and argue that the existence of reliable folding on biologically relevant time scales suggests an *intra-molecular* statistical relationship between secondary and *primary* structures. Our basic intuition is that, adopting the kinetic perspective, it should be harder for the molecule to make mistakes at locations that participate in the secondary structure. Otherwise the molecule would tend to get stuck in incorrect metastable states, and won't be able to fold correctly on a biologically relevant time scale.

In addition to this basic intuition, experimental literature [10–13] has long suggested that the stem-formation in RNA molecules is a two-step process. When forming a stem, we usually have a nucleation step, where we form a few consecutive base pairs at a nucleation point, followed by a fast zipping step. It seems especially intuitive that it should be harder for the molecule to make mistakes at these nucleation points, which are among the locations that participate in the secondary structure.

To present statistical evidence supporting the above hypotheses, we introduce the idea of the local ambiguity, with the goal of quantifying the possibility for the molecule to "make a mistake" at a particular location in the process of forming secondary structures. In our definition, for a particular location, we identify a nucleotide segment at this location, go through all the viable pairing candidate segments of this segment,

and use the number of candidate segments that are complementary to this segment as the measure of local ambiguity at this location.

Because of our definition of the local ambiguity, when we talk about a location within a molecule, we are really talking about the segment that corresponds to this location. Referring to the secondary structure, it's clear that we have three different kinds of locations:

***Single*:** Locations where all nucleotides in the segments are unpaired in the secondary structure

***Double*:** Locations where all nucleotides in the segments are paired in the secondary structure

***Transitional*:** Locations where some nucleotides in the segments are paired, while others are unpaired in the secondary structure

where *double* and *transitional* locations participate in the secondary structure, while *single* locations don't.

The goal of this paper is to verify our basic intuition, by looking at the differences in terms of local ambiguity between locations that participate in the secondary structure and the locations that don't, and to establish an intra-molecular statistical relationship between secondary and primary structures. This is achieved by some exploratory analysis as well as two sets of exact and highly significant hypothesis tests, which unveil fascinating results on the local ambiguities in different regions of the RNA molecules, the possible mechanistic differences in the structure formation of single-entity RNAs and protein-RNA complexes, as well as the subtle differences between the comparative analysis approach and the minimum free energy approach for RNA secondary structure prediction.

These experiments, motivated by a hypothesized mechanism of folding, and the first of their kind, are simple in the sense that they involve only RNA primary and secondary structure data (i.e. nucleotide sequences and base pairs) and elementary counting statistics, yet they yield significant insight into the folding of non-coding RNA molecules, and are consistent with folding to metastable but not necessarily equilibrium structures.

The rest of the paper is organized as follows: in the next section, we are going to make some basic notations and definitions, before presenting some exploratory analysis, as well as the two sets of exact hypothesis tests. Then we are going to move on to the final conclusion, before detailing vairous materials and methods used in the paper.

4 Results and Discussions

4.1 Basic Notations and Definitions

For a given RNA molecule, we are going to consider its primary and secondary structures data. Assume the length of the molecule is N, we denote the primary structure data 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 data by

$$s = \{(j, k) : 1 \leq j < k \leq N, \text{a base pair exists between the } j\text{th and the } k\text{th nucleotides}\} \tag{2}$$

With the above notations, we can make the definition of the local ambiguity precise. In this paper, we are going to consider segments of length 4. Assume the length of an

RNA molecule is N, then we are going to consider $N-3$ locations in this molecule. The segment at location i is given by

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

When trying to determine the local ambiguity of a location, we need to take into account that it's usually considered impossible for RNA to form a hairpin loop that contains less than 3 nucleotides. As a result, we define the set of viable pairing candidate segments for location i to be

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

In this paper, we are only going to consider Watson-Crick base pairs. As a result, two segments P_i and P_j are said to be complementary if

$$\forall 1 \leq k \leq 4, (P_{i,k}, P_{j,5-k}) \in \{(A,U), (U,A), (G,C), (C,G)\} \tag{5}$$

Using the above definitions, we define the local ambiguity function

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

by

$$a_i(p) = |\{P \in A_i : P \text{ and } P_i \text{ are complementary}\}|, i = 1, \cdots, N-3 \tag{6}$$

where $|\cdot|$ gives us the cardinality of the set, and $a_i(p)$ gives us the local ambiguity of the molecule with primary structure p at location i.

Finally, to formally state the idea of labelling different locations using secondary structure data, consider a molecule of length N with secondary structure s. Define the double-stranded indicator function $b(s) = (b_1(s), \cdots, b_N(s))$ to be

$$b_i(s) = \begin{cases} 1, & \text{if } \exists j \in \{1, \cdots, i-1, i+1, \cdots, N\}, \text{ s.t. } (i,j) \in s \text{ or } (j,i) \in s \\ 0, & \text{otherwise} \end{cases}, i = 1, \cdots, N \tag{7}$$

where $b_i(s)$ indicates whether the ith nucleotide is paired or not in the secondary structure s. Further define the paired nucleotides function

$$f(s) = (f_1(s), \cdots, f_{N-3}(s))$$

to be

$$f_i(s) = \sum_{j=i}^{i+3} b_i(s), i = 1, \cdots, N-3, \tag{8}$$

where $f_i(s)$ is the number of paired nucleotides in the segment at location i. Then we have

$$\text{location } i \text{ is } \begin{cases} \textit{single} \text{ if } f_i(s) = 0 \\ \textit{double} \text{ if } f_i(s) = 4 \\ \textit{transitional} \text{ if } 0 < f_i(s) < 4 \end{cases}, i = 1, 2, \cdots, N-3 \tag{9}$$

4.2 Some Exploratory Analysis

In this section, we are going to present some experimental results where we try to verify our basic intuition by comparing the local ambiguities in different kinds of regions. A natural thing to do is to use permutation tests. Here, the data points are the local ambiguities we get from the primary structure data, and the labels are the three different kinds of regions we get from the secondary structure data.

When comparing the local ambiguities in two different regions, a naive permutation test would directly permute the local ambiguities while keeping the labels intact. There are some obvious issues with this naive approach. The most important issue is that, by employing this naive approach, we are essentially ignoring the ordering/neighborhood information in the local ambiguities data. From our definition of the local ambiguities, it's clear that local ambiguities at nearby locations are correlated, and we should take this correlation information into account when permutating the labels.

Because of these reasons, we employ a different method when permuting the local ambiguities. Instead of directly (and naively) permuting the local ambiguities, we are going to first permute the primary structure data while maintaining the frequencies of segments of length 4, using what we call a Markov shuffling method, and then re-evaluate the local ambiguities at each location. Since the frequencies of segments of length 4 are maintained, by permuting the primary structure data and re-evaluating the local ambiguities, we essentially achieve a permutation of the local ambiguities which takes into account the correlation information (it's exactly a permutation of the local ambiguities when we take all possible segments to be viable pairing candidates, but even if we restrict ourselves to some set of viable pairing candidates, it's still very close to an exact permutation of the local ambiguities).

Note that, by employing the Markov shuffling method, while we solve the problem of taking into account the correlation information, we don't really have a clearly enunciated hypothesis. This stems from the fact that we don't have a very good interpretation of the sequences we get when permuting the primary structure data using the Markov shuffling method. In particular, we can't interpret these sequences as primary structures for surrogate molecues, as these sequences are not consistent with secondary structure that goes with the original primary structure. If we can have an efficient way of permuting the local ambiguities while maintaining the secondary structures, or if we can determine the secondary structures of the permuted sequences reliably, we would solve this problem. Unfortunately, neither of these is realistic, and consequently, the results in this section can only serve as some exploratory analysis. Nevertheless, it turns out this analysis yields great insight into the statistical behaviour of the local ambiguities in different kinds of regions for different kinds of RNA molecules, and provides us with the motivation for the definition of the "ambiguity index" and the design of the formal hypothesis tests.

More details of the Markov shuffling method would be given in the materials and methods section. For now, it suffices to assume that for any k, we have a kth order Markov shuffling algorithm $\mathcal{M}_k$, which takes a sequence p as input, and gives a randomly shuffled sequence $\mathcal{M}_k(p)$, where the frequencies of segments of length k are preserved, as output.

4.2.1 The Problem of Bias

For the exploratory analysis, the natural thing to do is to make pariwise comparisons among the three different kinds of regions. However, caution is needed here because of certain inherent bias in the definition of local ambiguity.

For the *double* region, by definition, the segment at a *double* location would have at least one complementary segment within the molecule. As a result, it's easy to imagine the *double* region being consistently more ambiguous than the *single* and *transitional* regions.

Following a similar reasoning, it's not hard to see that there's a bias of the opposite direction in *single* regions. If it's possible for a particular stem to extend, it would almost certainly make the extension. As a result, the two segments within the *single* regions at the opposite end (either the two inner ends or the two outer ends) of the

same stem won't be complementary to each other, thus lowering the ambiguity of the *single* regions. This is a very small bias, but it's a bias nonetheless.

The above discussion indicates that, for the purpose of verifying our intuition, the only meaningful comparison would be to see if the *transitional* region is less ambiguous than the *single* region. This motivates us to define a central statistic in our work, which we call the "ambiguity index". For an RNA molecule of length N, with primary structure p and secondary structure s, define

$$c_i^{\text{tran}} = \begin{cases} 1, & \text{if location } i \text{ is } \textit{transitional} \\ 0, & \text{otherwise} \end{cases}, i = 1, \cdots, N-3 \tag{10}$$

$$c_i^{\text{single}} = \begin{cases} 1, & \text{if location } i \text{ is } \textit{single} \\ 0, & \text{otherwise} \end{cases}, i = 1, \cdots, N-3 \tag{11}$$

The "ambiguity index" is then given by

$$d(p,s) = \frac{\sum_{j=0}^{N-1} a_j(p) c_j^{\text{tran}}(s)}{\sum_{j=0}^{N-1} c_j^{\text{tran}}(s)} - \frac{\sum_{j=0}^{N-1} a_j(p) c_j^{\text{single}}(s)}{\sum_{j=0}^{N-1} c_j^{\text{single}}(s)} \tag{12}$$

In the following sections, we are going to present some experimental results from permutation-tests-like analysis based on the Markov shuffling method. We present two groups of results, one where we do the permutations on an individual molecule level, and the other where we do the permutations on a group level. We only include the results for the comparison of *transitional* and *single* regions, but for the sake of completeness, we also carry out experiments for the comparison of *single* and *double* regions, and of *transitional* and *double* regions, and include the results in the support information.

4.2.2 Permutating Individual Molecules

The first experiment we are going to do involves permuting individual molecules. The basic idea is, we are going to use the Markov shuffling method to permute the primary structure data of individual molecules, and re-evaluate the local ambiguities at different locations. We then compare the mean local ambiguities in different kinds of regions, and derive a pseudo-p-value for each molecule. We are going to calculate some simple summary statistics for different groups of RNAs.

Formally, assume we have a group of RNA molecules of length $N^{(m)}$, with primary structures $p^{(m)} = p^{(m,0)}$ and secondary structures $s^{(m)}$, $m = 1, \cdots, M$. For each molecule, we are going to generate K random Markov shuffles

$$p^{(m,1)}, \cdots, p^{(m,K)}, m = 1, \cdots, M \tag{13}$$

using the Markov shuffling algorithm $\mathcal{M}_r$.

We want to test the hypothesis that the mean local ambiguity for *transitional* is lower than that for *single*. The obvious statistic we are going to use is

$$d(p^{(m,k)}, s^{(m)}), k = 0, 1, \cdots, K \tag{14}$$

Our pseudo-null-hypothesis would be that there's no difference in terms of mean local ambiguity between *transitional* and *single*. This is a one-sided pseudo-hypothesis-test, and the pseudo-p-value is given by

$$q_{\text{ind}}(p^{(m)}, s^{(m)}, K) = \frac{1 + \sum_{k=1}^{K} \chi_{\{d(p^{(m,k)}, s^{(m)}) \leq d(p^{(m,0)}, s^{(m)})\}}}{1+K} \tag{15}$$

In this section, given a significance level α, we are going to report a simple summary statistic for a group of M RNA molecules

$$\frac{\sum_{m=1}^{M} \chi_{\{q_{\text{ind}}(p^{(m)},s^{(m)},K)<\alpha\}}}{M} \tag{16}$$

which is the percentage of pseudo-hypothesis-tests that are significant at level α.

4.2.3 Permutating a Group of Molecules

We can do a similar experiment, where, instead of permuting individual molecules, we are going to permute a group of molecules. The basic idea is, given a group of molecules, for each permutation, we are going to apply the Markov shuffling algorithm to each individual molecules, and count the number of "permuted molecules" where mean ambiguity for *transitional* is less than that for *single*. This count would serve as our statistic in the pseudo-hypothesis-test.

Formally, again assume we have a group of RNA molecules of length $N^{(m)}$, with primary structures $p^{(m)} = p^{(m,0)}$ and secondary structures $s^{(m)}$, $m = 1, \cdots, M$. Define the set of primary-secondary-structure-tuples to be

$$\mathcal{P} = \mathcal{P}^{(0)} = \{(p^{(m,0)}, s^{(m)}), m = 1, \cdots, M\} \tag{17}$$

We are going to generate K random Markov shuffles of this group

$$\mathcal{P}^{(k)} = \{(p^{(m,k)}, s^{(m)}), m = 1, \cdots, M\}, k = 1, \cdots, K \tag{18}$$

We again want to test the hypothesis that the mean local ambiguity for *transitional* is lower than that for *single*. The statistic we are going to use this time would be

$$g(\mathcal{P}^{(k)}) = \sum_{m=1}^{M} \chi_{\{d(p^{(m,k)},s^{(m)})<0\}}, k = 0, 1, \cdots, K \tag{19}$$

Our pseudo-null-hypothesis would be that there's no difference in terms of mean local ambiguity between *transitional* and *single*. This is again a one-sided pseudo-hypothesis-test, and the pseudo-p-value is given by

$$q_{\text{grp}}(\mathcal{P}, K) = \frac{1 + \sum_{k=1}^{K} \chi_{\{g(\mathcal{P}^{(k)}) \geq g(\mathcal{P}^{(0)})\}}}{1 + K} \tag{20}$$

4.2.4 Hyper-parameters and Experimental Results

We ran the experiments for 4 different groups of RNA molecules (Group I Introns, Group II Introns, SRP RNAs and tmRNAs). In our experiments, we used hyper-parameters

$$K = 10^4, \alpha = 0.05 \tag{21}$$

For secondary structures, we used both the *comparative analysis* structures and the *minimum free energy* structures.

The M values, the percentages of significant (at level α) pseudo-hypothesis tests (denoted "Percentage") when permuting individual molecules, and the pseudo-p-values of the pseudo-hypothesis-tests (denoted "p-value") when permuting a group of molecues for the 4 groups of RNA molecules are reported. The results are reported in Table 1.

		M	Percentage	p-value
Comparative Analysis	Group I Introns	116	0.31896552	0.00009999
	Group II Introns	37	0.67567568	0.00009999
	SRP RNAs	832	0.02163462	1.00000000
	tmRNAs	462	0.14502165	0.98420158
Minimum Free Energy	Group I Introns	116	0.20689655	0.07029297
	Group II Introns	37	0.54054054	0.00009999
	SRP RNAs	832	0.03004808	1.00000000
	tmRNAs	462	0.15800866	0.70602940

Table 1. Exploratory Analysis Results

4.3 Discussions on the Exploratory Analysis Results

4.3.1 Observations

For the p-values from permuting the molecules on a group level, the observation is that, when we look at the comparison *transitional* v.s. *single*, we see different results. For the *comparative analysis* structures, the comparison is significant for *Group I Introns* and *Group II Introns*, but not significant for *SRP RNAs* and *tmRNAs*, while for the *minimum free energy* structures, the comparison is only significant for *Group II Introns*, although it comes very close to being significant for the *Group I Introns*.

We also included the percentages of significant (at $\alpha = 0.05$ level) pseudo-hypothesis tests when we permute individual molecules. These results should be used only as a reference. As we can see from the results, in general, the percentages are negatively correlated with the p-values.

Note that we included the M values in the results, and these M values don't agree with the numbers of molecules in different groups from the dataset we are using. This is because of some implementation details (mainly that we ignore those molecules where we don't have unique Markov shuffles), which would be explained in more detail in the materials and methods section.

4.3.2 Motivation for Formal Hypothesis Tests

For the comparison of *transitional* and *single* regions, the different results we get on different groups of RNA molecules and different kinds of secondary structures are highly interesting, and suggest the possibility of using this comparison to distinguish different kinds of RNA molecules and the two different kinds of secondary structures.

The first interesting thing that's worth exploring further is the different results on different kinds of RNA molecules. It seems the 4 groups of RNA molecules are natually divided into two larger groups, with one group containing the *Group I Introns* and *Group II Introns*, and the other group containing the *SRP RNAs* and *tmRNAs*. Further inspection of these two larger groups reveals the obvious difference between them: *Group I Introns* and *Group II Introns* operate as *single-entity RNAs*, while *SRP RNAs* and *tmRNAs* belong to certain *protein-RNA complexes*. This difference suggest we might be able to statistically separate these two larger groups of RNAs using the "ambiguity index".

The second interesting thing that's worth exploring is the difference between *comparative analysis* structures and *minimum free energy* structures. Our basic intuition in this paper is that kinetics should play a much more important role in the computational prediction of RNA secondary structures, and it might not be biologically plausible to try to find the *minimum free energy* structures. To this end, we want to look at whether there's any qualitative difference between the gold standard of

secondary structure prediction, the *comparative analysis* methods, and the *minimum free energy* method. The different degrees of significance (as shown by the p-values), as well as the different percentages of significant pseudo-hypothesis tests at the individual molecule level suggest we might be able to use the "ambiguity index" to statistically distinguish these two secondary structure prediction methods.

These two things motivate two corresponding sets of formal hypothesis tests, which we present next.

4.4 Formal Hypothesis Tests

4.4.1 Single-entity RNAs v.s. Protein-RNA complexes

The first formal hypothesis test we are going to do is to see if we can use the "ambiguity index" to statistically separate the RNA molecules that operate as *single-entity RNAs* from those that belong to *protein-RNA complexes*. The statistical tool here is still permutation tests.

Formally, the null hypothesis we are going to test would be:

$\mathbb{H}_0$: There's no difference in terms of "ambiguity index" between *single-entity RNAs* and *protein-RNA complexes*

The results from the exploratory analysis suggest that the ambiguity indices of *single-entity RNAs* tend to be smaller than those in *protein-RNA complexes*. Motivated by this, we are going to conduct a one-sided hypothesis test, in which we consider the simplest possible aspect of the "ambiguity index", the sign of the "ambiguity index". The alternative hypothesis is thus given by

$\mathbb{H}_1$: The ambiguity indices are more often negative for *single-entity RNAs* than for *protein-RNA complexes*

Assume we have M_1 *single-entity RNAs* with primary structures $p^{(1,m)}$, secondary structures $s^{(1,m)}$, and M_2 RNAs as part of *protein-RNA complexes* with primary structures $p^{(2,m)}$, secondary structures $s^{(2,m)}$. The test statistic we are going to use is

$$\frac{\sum_{m=1}^{M_1} \chi_{\{d(p^{(1,m)},s^{(1,m)})<0\}}}{M_1} - \frac{\sum_{m=1}^{M_2} \chi_{\{d(p^{(2,m)},s^{(2,m)})<0\}}}{M_2} \tag{22}$$

It's not hard to see that, employing a permutation test where we permute the labels of *single-entity RNAs* and *protein-RNA complexes*, we can calculate the p-value exactly. Define

$$n_t = \sum_{m=1}^{M_t} \chi_{\{d(p^{(t,m)},s^{(t,m)})<0\}}, t = 1, 2 \tag{23}$$

We have the one-sided p-value is given by

$$\sum_{n=n_1}^{\min\{n_1+n_2,M_1\}} \frac{\binom{n_1+n_2}{n}\binom{M-n_1-n_2}{M_1-n}}{\binom{M}{M_1}} \tag{24}$$

We ran the experiments using the 4 groups of RNA molecules. In our terminology, the *Group I Introns* and *Group II Introns* are *single-entity RNAs*, while the *SRP RNAs* and *tmRNAs* belong to *protein-RNA complexes*. As a result, we have $M = 1539, M_1 = 153$ and $M_2 = 1386$. For the secondary structures, we again use both the *comparative analysis* structures and the *minimum free energy* structures. The values of n_1, M_1, n_2 and M_2, as well as the p-values for the two different kinds of secondary structures are summarized in Table 2.

	n_1	M_1	n_2	M_2	p-value
Comparative Analysis	114	153	536	1369	4.143753e-17
Minimum Free Energy	96	153	511	1369	1.484907e-09

Table 2. *Single-entity RNAs* v.s. *Protein-RNA complexes*

4.4.2 Comparative analysis v.s. Minimum Free Energy Methods

From the results in the last section, we can see that the "ambiguity index" can statistically seperate the *single-entity RNAs* from *protein-RNA complexes*, but the separation works for both the *comparative analysis* structures, and the *minimum free energy* structures. To zero in on the differences in terms of the "ambiguity index" between these methods, we conduct a further hypothesis test within the group of *single-entity RNAs* and the group of *protein-RNA complexes*.

Our null hypothesis is

$\mathbb{H}_0$: There's no difference in terms of "ambiguity index" between *comparative analysis* structures and *minimum free energy* structures within the group of RNAs.

We again conduct a one-sided hypothesis test, with the alternative hypothesis

$\mathbb{H}_1$: The "ambiguity indices" are lower for *comparative analysis* structures than for *minimum free energy* structures within the group of RNAs

Assume we have a group of M RNAs, with primary structures $p^{(m)}$, *comparative analysis* secondary structures $s^{(m)}_{\text{comp}}$, and *minimum free energy* secondary structures $s^{(m)}_{\text{MFE}}$, $m = 1, \cdots, M$. The test statistic we are going to use is

$$n_0 = \sum_{m=1}^{M} \chi_{\{d(p^{(m)}, s^{(m)}_{\text{comp}}) < d(p^{(m)}, s^{(m)}_{\text{MFE}})\}} \tag{25}$$

Employing a permutation test where we permute the labels of *comparative analysis* structures and *minimum free energy* structures, the test statistic follows a Binomial distribution $B(M, 0.5)$, and the exact p-value is given by $\sum_{n=n_0}^{M} \frac{\binom{M}{n}}{2^n}$.

We ran the experiments for both the group of *single-entity RNAs*, and the group of *protein-RNA complexes*. The M values, the n_0 statistics and the p-values for the two groups of RNAs are reported in Table 3.

	n_0	M	p-value
Single-entity RNAs	98	153	0.000318
Protein-RNAs complexes	690	1369	0.393482

Table 3. *Comparative Analysis* v.s. *Minimum Free Energy* Methods

4.5 Discussions on the Formal Hypothesis Tests Results

From the above results, we can see that we verified a lot of the suspicions we had in the discussions of the exploratory analysis results. We demonstrated that we can indeed statistically seperate *single-entity RNAs* from *protein-RNA complexes*, using the "ambiguity index". We can also statistically seperate the *comparative analysis* structures from the *minimum free energy* structures for *single-entity RNAs*, using the "ambiguity index", but we can't achieve this when we look at *protein-RNA complexes*.

These results suggest a potentially significant difference between the structure formation mechanisms on *single-entity RNAs* and *protein-RNA complexes*, and a qualitative difference between the *comparative analysis* methods and the *minimum free energy* methods when applied to *single-entity RNAs*.

4.6 Conclusion

First, from the experiments and the results in this paper, "local ambiguity" clearly emerges as a useful concept in the statistical analysis of RNA primary and secondary structure data and the folding process of RNA molecules.

Based on "local ambiguity", an "ambiguity index", one for each pair of molecule and presumed secondary structure, measures the prevalence of false matches and hence the tendency to form metastable structures incompatible with native structures. The ambiguity index statistically separates an ensemble of RNA molecules that operate as single entities (Group I and II Introns) from an ensemble that operates as protein-RNA complexes (SRP and tmRNAs), and ensembles of secondary structures determined by comparative analysis from ones based on thermal equilibrium. We find lower average ambiguity in single-entity RNA's than protein-RNA complexes, and, among single-entity RNA's, lower ambiguity with comparative analyses than equilibrium analyses. These results demonstrate possible mechanistic differences in the structure formation of single-entity RNAs, where kinetics play a more important role, and protein-RNA complexes, where the protein-RNA interactions have significant impacts on the folding process, and a qualitative difference between the comparative analysis approach and the minimum free energy approach in the case where kinetics are important (in this paper, for single-entity RNAs).

These empirical evidence points to the importance of carefully considering the impacts of kinetics and protein-RNA interactions on the folding process of non-coding RNA molecules, and argues against the naive application of thermal equilibrium based approaches for RNA secondary structure prediction.

5 Materials and methods

5.1 Datasets

In this paper, we used secondary structures data from comparative analysis for 4 different kinds of RNA molecules: Group I Introns and Group II Introns, from the comparative RNA website [14], SRP RNAs from SRPDB (Signal Recognition Particle Database) [15], and tmRNAs from tmRDB (tmRNA database) [15]. Refer to the corresponding papers and the websites for more details on how the comparative analysis was done.

To get the Group I Introns and Group II Introns comparative analysis secondary structures data, go to Section 3C (Mass Data Retrieval, `http://www.rna.icmb.utexas.edu/DAT/3C/Structure/index.php`) of the website, and download all the secondary structures data (in bpseq format) for Group I Introns and Group II Introns.

To get the SRP RNAs comparative analysis secondary structures data, go to the SRP RNA page (`https://rth.dk/resources/rnp/SRPDB/srprna.html`) of the SRPDB website, and select get all SRP RNA secondary structures.

To get the tmRNAs comparative analysis secondary structures data, go to the tmRNA page (`http://www.ag.auburn.edu/mirror/tmRDB/rna/tmrna.html`) of the tmRDB website, and select get all ct files.

5.2 Markov shuffling

Randomly shuffled sequences are routinely used in sequence analysis to evaluate the statistical significance of a biological sequence. In this paper, we are using what we call a Markov shuffling method, which is based on the Euler Algorithm [16–18]. The basic idea is to generate uniform random k-let-preserving sequences. In this paper, we used an efficient and flexible implementation of the Euler algorithm, called uShuffle [19]. Refer to the papers for more details on the Euler algorithm and the implementation details.

5.3 Minimum Free Energy Methods for Secondary Structure Prediction

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 process [20–26]. In this paper, we used the ViennaRNA package [20] to obtain the minimum free energy secondary structures for our statistical analysis. Refer to the paper and their website (`https://www.tbi.univie.ac.at/RNA/`) for more details on this package.

5.4 Reproducing the Results

The main results in this paper are summarized in Tables 1, 2, 3, 4 and 5. These experimental results are easily reproducible. To reproduce these 5 tables, follow the instructions at `https://github.com/StannisZhou/rna_statistics`.

Note that the above git repository already contains all the comparative analysis structures data for the 4 different kinds of RNA molecules considered in this paper. However, the readers should feel free to download the data from the original sources, and even adapt the code to apply the same analysis procedures to other kinds of RNA molecules.

When trying to reproduce the results, the readers should also note that the Markov shuffling procedure is quite computationally intensive, and might take a while to finish when trying to get a large number of shuffles.

5.5 Implementation Details

We also need to make a few comments regarding some implementation details. The main thing involved is cleaning up the data, and making sure we have good power in our hypothesis tests.

- 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 doing the Markov shuffling, we ignored molecules for which we don't have unique shuffles. These are mostly relatively short molecules.
- 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).

Because of these small details, the number of molecules used in the analysis (as reported in Table 4, Table 5, Table 2 and Table 3) is smaller than the actual number of molecules we have in the datasets.

6 Acknowledgments

7 Appendix

7.1 Complete Exploratory Analysis Results

For the sake of completeness, in this section, we present detailed comparisons of the local ambiguity among the three different kinds of regions. The basic setup is the same, but instead of making only the comparison between *transitional* and *single*, we make three different comparisons:

- *double* v.s. *single*
- *transitional* v.s. *single* (Denoted *tran* v.s. *single*)
- *transitional* v.s. *double* (Denoted *tran* v.s. *double*)

For these comparisons, again remember that here we are conducting one-sided pseudo-hypothesis tests, and the order matters. We are operating under the assumption that the first kind of region is less ambiguous than the second kind of region.

The M values, the percentages of significant (at level α) pseudo-hypothesis tests (denoted "Percentage") when permuting individual molecules, and the pseudo-p-values of the pseudo-hypothesis-tests (denoted "p-value") when permuting a group of molecues for the 4 groups of RNA molecules are reported. The results for *comparative analysis* structures are reported in Table 4, and the results for *Minimum Free Energy* structures are reported in Table 5.

		M	Percentage	p-value
Group I Introns	*double* v.s. *single*	116	0.05172414	1.00000000
	tran v.s. *single*	116	0.31896552	0.00009999
	tran v.s. *double*	116	0.59482759	0.00009999
Group II Introns	*double* v.s. *single*	37	0.48648649	0.05609439
	tran v.s. *single*	37	0.67567568	0.00009999
	tran v.s. *double*	37	0.35135135	0.00089991
SRP RNAs	*double* v.s. *single*	832	0.00721154	1.00000000
	tran v.s. *single*	832	0.02163462	1.00000000
	tran v.s. *double*	832	0.56490385	0.00009999
tmRNAs	*double* v.s. *single*	462	0.02380952	1.00000000
	tran v.s. *single*	462	0.14502165	0.98420158
	tran v.s. *double*	462	0.66883117	0.00009999

Table 4. Exploratory Analysis Results for *Comparative Analysis* Method

For the p-values from permuting the molecules on a group level, we've already talked aobut the different results for the comparison between *transitional* and *single*. Besides that, the main observation is that, when it involves the *double* regions, the results are very consistent. We have, at the 0.05 level, the *transitional* regions are consistently less ambiguous than the *double* regions, while the *double* regions are never significantly less ambiguous than the *single* regions. The only case where *double* comes close to being significantly less ambiguous than *single* is for the *Group II Introns* with the *comparative analysis* structures. But even that is not significant at a 0.05 level. This demonstrates the effects of the inherent bias for local ambiguity.

The percentages of significant (at $\alpha = 0.05$ level) pseudo-hypothesis tests when we permute individual molecules are also included, but again should be used only as a

		M	Percentage	p-value
Group I Introns	*double* v.s. *single*	116	0.08620690	1.00000000
	tran v.s. *single*	116	0.20689655	0.07029297
	tran v.s. *double*	116	0.77586207	0.00009999
Group II Introns	*double* v.s. *single*	37	0.21621622	0.89061094
	tran v.s. *single*	37	0.54054054	0.00009999
	tran v.s. *double*	37	0.70270270	0.00009999
SRP RNAs	*double* v.s. *single*	832	0.00120192	1.00000000
	tran v.s. *single*	832	0.03004808	1.00000000
	tran v.s. *double*	832	0.68149038	0.00009999
tmRNAs	*double* v.s. *single*	462	0.02597403	1.00000000
	tran v.s. *single*	462	0.15800866	0.70602940
	tran v.s. *double*	462	0.71861472	0.00009999

Table 5. Exploratory Analysis Results for *Minimum Free Energy* Method

reference. As we can see from the results, in general, the percentages are negatively correlated with the p-values.

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. 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.

4. 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.

5. 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`.

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

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

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. 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.

15. 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.

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

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

18. 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.

19. 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.

20. 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.

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

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

23. 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.

24. 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.

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

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