Published in final edited form as:

Bull Math Biol. ; 82(10): 133. doi:10.1007/s11538-020-00804-0.

On the Problem of Reconstructing a Mixture of RNA Structures

Torin Greenwood¹, Christine E. Heitsch²

¹North Dakota State University, Fargo, USA

²Georgia Institute of Technology, Atlanta, USA

Abstract

A growing number of RNA sequences are now known to exist in some distribution with two or more different stable structures. Recent algorithms attempt to reconstruct such mixtures using the list of nucleotides in a sequence in conjunction with auxiliary experimental footprinting data. In this paper, we demonstrate some challenges which remain in addressing this problem; in particular we consider the difficulty of reconstructing a mixture of two RNA structures across a spectrum of different relative abundances. Although progress has been made in identifying the stable structures present, it remains nontrivial to predict the relative abundance of each within the experimentally sampled mixture. Because the ratio of structures present can change depending on experimental conditions, it is the footprinting data—and not the sequence—which must encode information on changes in the relative abundance. Here, we use simulated experimental data to demonstrate that there exist RNA sequences and relative abundance combinations which cannot be recovered by current methods. We then prove that this is not a single exception, but rather part of the rule. In particular, we show, using a Nussinov–Jacobson model, that recovering the relative abundances is difficult for a large proportion of RNA structure pairs. Lastly, we use information theory to establish a framework for quantifying how useful auxiliary data is in predicting the relative abundance of a structure. Together, these results demonstrate that aspects of the problem of reconstructing a mixture of RNA structures from experimental data remain open.

Keywords

RNA secondary structure; Thermodynamic optimization; Auxiliary data

1 Introduction

Determining the structural conformations of an RNA sequence reveals functional information. Identifying structures in the laboratory is difficult, so discrete optimization methods are used instead. These methods use the list of nucleotides in a sequence to predict two-dimensional approximations of structures, called secondary structures, that still encode functional information. However, an increasing number of RNA molecules are now known to fold into multiple stable structures (Leonard et al. 2013; Spasic et al. 2018). For example, approximately 20% of eukaryotic RNA folds into multiple structural conformations *in vivo* (Lu et al. 2016). Differing conformations can also yield multiple biological functions for the

✉ Torin Greenwood, torin.greenwood@ndsu.edu.

same RNA sequence, as is the case for riboswitches (Antunes et al. 2018). In response, single-structure prediction methods have been refined to find distributions of RNA structures, as reviewed in Schroeder (2018).

Predicting distributions of structures requires identifying the most common structures in the distribution (the *structural modes*), and the fraction of each within the distribution (the *conformational ratios*). Current prediction methods, described below, intake two pieces of information: the list of nucleotides in the RNA sequence, and auxiliary experimental footprinting data. This typically noisy data contains a signal positively correlated to each nucleotide's unpairedness. For a given RNA sequence, the conformational ratio between structural modes is likely to change depending on a variety of extrinsic factors (Höbartner and Micura 2003). As a result, the auxiliary data must contain information about the conformational ratios that cannot be recovered from the sequence alone.

In this paper, we establish a mathematical framework to evaluate how well current prediction models can identify structural modes and conformational ratios. We focus on bimodal structural distributions because they are the simplest form of multimodal distribution. To be precise, consider an RNA sequence with two structural modes A and B . As described in Sect. 3, if A and B have corresponding data sequences S and T , we model data from a distribution where mode A has conformational ratio p by the linear interpolation, $M(p) = pS + (1 - p)T$. We ask: given the RNA sequence and $M(p)$, can a prediction model recover the conformational ratio p and the structural modes A and B ?

In Sect. 3, we test the ability of several common prediction methods to recover distributions of known RNA sequences. In particular, Rsample is a recent enhancement to the data-directed prediction algorithms designed to identify multimodal conformations (Spasic et al. 2018). Rsample advanced the field by improving structural mode and conformational ratio predictions. We use the Bistable and HIV-1 RRA sequences from the test set in the Rsample publication. Currently, there is not a large collection of experimental data available from distributions of structures for the same sequence under different experimental conditions. Thus, in order to analyze the performance of prediction methods systematically, we use empirically derived distributions of auxiliary data from Sükösd et al. (2013), described below, to simulate data. Then, we check the performance of prediction methods as we vary the conformational ratio p of one of the structural modes. We find the following:

Conclusion 1: For the HIV-1 RRA sequence, current prediction methods cannot recover all possible conformational ratios of the structural modes.

In the remaining sections, we provide mathematical evidence that these challenges are not unique to HIV-1 RRA. We focus on the key role that auxiliary data plays in predicting ratios. Our analysis creates a framework for any type of data that gives a signal of unpairedness for individual nucleotides, including SHAPE data (described below), chemical probing with dimethyl sulfate, or when NMR chemical shifts are used to predict base pairing probabilities. Regardless of the type of data used, two pivotal factors are the way the auxiliary data is incorporated into prediction models, and how much information the data contains about conformational ratios.

To separate the factors, in Sect. 4, we first consider how well prediction models perform when given noiseless auxiliary data. In order to remove the noise from the auxiliary data, we consider a binary model of data that describes whether each nucleotide in a structure is paired or unpaired, without revealing the complementary nucleotide in a pairing. Then, we investigate a Nussinov–Jacobson model of assigning highly simplified energies to structures (Nussinov and Jacobson 1980), which is amenable to combinatorial analysis but still captures critical features of the other models. Despite the Nussinov–Jacobson model’s simplicity, its analysis has advanced our understanding of secondary structure prediction in other contexts, e.g. Clote (2017); Clote et al. (2007). In Theorem 1, we find:

Conclusion 2: As the length of RNA sequences increases, almost all pairs of structures have at least 25% of conformational ratios that cannot reliably be reproduced by a Nussinov–Jacobson energy model.

After having analyzed an energy model directly, we then turn towards analyzing the auxiliary data directly in Sect. 5. Some of the challenges of interpreting auxiliary data in multimodal distributions are revealed in Vieweger and Nesbitt (2018). Motivated by the ability of Rsample to identify structures within a distribution and the “sample and select” algorithm from Quarrier et al. (2010), we assume that the structural modes in a bimodal distribution are known, but that their conformational ratio p is not. We then use tools from information theory to quantify how well the ratio can be recovered from the auxiliary data. We find the following:

Conclusion 3: Let n be the number of positions where the nucleotides in two structural modes are not both paired nor both unpaired. Any unbiased estimator of the conformational ratio p will have a variance that decreases at most linearly in n .

2 Background and Related Work

Before proceeding, we give some background information on auxiliary data and prediction models.

2.1 Nearest Neighbor Thermodynamic Model

Energy models form one common approach to predicting RNA secondary structures. In particular, the nearest neighbor thermodynamic model (NNTM) has hundreds of parameters, corresponding to all the possible features in an RNA structure (Mathews and Turner 2006; Turner and Mathews 2010). Structures are then assigned energy scores by summing the subscores of their components. Then, one popular approach is to find the minimum free energy (MFE) structure by using dynamic programming algorithms [based on the Zuker algorithm (Zuker and Stiegler 1981)]. Because the NNTM is a model, these MFE structures do not always have perfect accuracy. Plausible alternatives can be generated by predicting suboptimal structures (Zuker 1989; Wuchty et al. 1999), and computing base pairing probabilities under a Boltzmann partition function (McCaskill 1990; Hofacker et al. 1994; Mathews 2004).

Alternatively, when analyzing multimodal distributions, we use these suboptimal structures to predict a distribution. Given a fixed RNA sequence, energies can be used to define a Boltzmann distribution of structures: a structure S with energy $\mathcal{E}(S)$ is assigned the probability,

$$\mathbb{P}(S) := \frac{e^{-\mathcal{E}(S)/RT}}{Z}, \quad (1)$$

where Z is a partition function over all possible structures for the sequence, $R = 0.001987$ kcal/(mol K) is the gas constant, and T is the temperature, which we take to have default value 310K. Note that when an RNA sequence is predicted to fold into a single structure, a prediction algorithm typically aims to find the structure S which minimizes the energy, or equivalently maximizes $\mathbb{P}(S)$. To search the space of suboptimal structures, there are efficient algorithms which can sample structures from the distribution (Ding 2003). In Sect. 3, we use pseudoenergy models that incorporate the NNTM and auxiliary data, described below. Then, in Sect. 4, we use a simplified model based on the Nussinov–Jacobson model (Nussinov and Jacobson 1980), which is easier to analyze combinatorially.

2.2 Auxiliary Data and Pseudoenergies

In order to improve the accuracy of predictions, auxiliary laboratory data can be used to reweight predicted distributions of structures by adding additional *pseudoenergy* terms to each structure. As described in Deigan et al. (2009), when auxiliary data is included in single-structure predictions, the accuracy of these methods is high enough to recover important structural information about wide classes of RNA sequences. One type of auxiliary data is chemical probing data, where folded RNA sequences are exposed to chemicals, and the reactivity of each nucleotide is measured. Different chemicals can be used, and each can give a different structural signal (Rice et al. 2014). In Sect. 3, we focus on SHAPE data, named after the method, *selective 2'-hydroxyl acylation analyzed by primer extension*. However, the framework we establish in Sects. 4 and 5 can be used for any auxiliary data that gives information about each nucleotide's unpairedness, including chemical probing with other reagents like dimethyl sulfate, and using NMR chemical shifts to predict base pairing probabilities.

Although SHAPE data is positively correlated with each nucleotide's unpairedness, the signal is noisy due to a variety of experimental and biological factors. There are several methods for converting this signal into pseudoenergies to shift the Boltzmann distribution, and Eddy (2014) compares some common SHAPE methods for single-structure prediction. Below, we use the models from Deigan et al. (2009), Zarringhalam et al. (2012), and Washietl et al. (2012), and we use the implementation in the ViennaRNA software package (Lorenz et al. 2011).

Recently, new data-directed approaches aim to probe or reconstruct multimodal distributions. We focus on Rsample (Spasic et al. 2018), which uses a single SHAPE sequence and an advanced pseudoenergy model. Rsample succeeds in identifying the constituent structures within several known RNA structural distributions. Identifying the correct ratios between structures remains challenging, as we will see in Sect. 3. Other prediction models

include REEFIT (Cordero and Das 2015), which uses additional mutate-and-map information to guide predictions. Finally, Ensemble RNA (Woods et al. 2017) probes the structural space for potential structural modes without being limited by the exponential decay of Boltzmann probabilities. All three programs are described in Schroeder (2018).

2.3 Modeling SHAPE Data

Although there is evidence of the existence of multimodal structural distributions, there is not yet a large collection of data derived from known multimodal distributions. Thus, we model SHAPE data coming from a multimodal distribution with a linear interpolation of SHAPE data from two structures, as described in Sect. 3, and in line with Spasic et al. (2018). To model the SHAPE data from a single known structure, we use the empirically derived distributions developed in Sükösd et al. (2013). Here, the authors used SHAPE data collected in Deigan et al. (2009) from a 16S and a 23S *Escherichia coli* ribosomal sequence to fit distributions for edge-paired, center-paired, and unpaired nucleotides. A nucleotide is center-paired if it and both its neighbors belong to the same helix, or if it is paired and adjacent to a single-nucleotide bulge. Otherwise, a paired nucleotide is edge-paired.

In Sect. 3, we use these empirical distributions to simulate data to illustrate the potential problem with reconstructing conformational ratios. Later, in Sect. 5, we use the empirical distributions again to analyze how much information SHAPE data contains about the ratios. In contrast, in Sect. 4, we use a noiseless binary model of SHAPE data so that we can analyze the prediction models without confounding factors.

2.4 Total Variation Distance

We choose a method to quantify error when predicting conformational ratios, based on the total variation distance. The total variation distance is a common distance measure for probability measures, defined as the maximum difference the two measures could assign to any event. When comparing two Bernoulli distributions with parameters p and $\hat{p}$, the total variation distance is simply $|p - \hat{p}|$.

To apply this distance to distributions of RNA structures, consider predicting a target bimodal distribution comprised of ratio p of structure A and $(1 - p)$ of structure B . To quantify how well a prediction method can reconstruct this distribution, we define $\hat{p}$ to be the probability that a structure in the prediction method's Boltzmann distribution is closer to structure A than it is to structure B (by comparing F-measures). We estimate $\hat{p}$ by taking a sample of 1000 structures, and then we call the absolute difference between this estimate and p the total variation distance. Note that the reduction of the Boltzmann distribution to only two classes of structures (those closer to A and those closer to B) causes this total variation distance to be less than had we allowed some structures to be classified as a mismatch. Below, we use this definition to find lower bounds on the error of common prediction methods in predicting p . This error measure can easily be extended to distributions with more than two modes.

3 Examples

We begin with the synthetic 25-nucleotide Bistable 1 sequence from Höbartner and Micura (2003), which is modeled after riboswitches. In Spasic et al. (2018), Rsample was found to accurately reconstruct a bimodal Bistable distribution when given SHAPE data corresponding to a known experimental setting. However, the conformational ratio is expected to be dependent on external factors, so we will explore the accuracy of the models when the conformational ratio is changed. To do so, we simulate auxiliary experimental data sequences $\{S_i\}$ and $\{T_i\}$ corresponding to the Bistable structures A and B , where S_i corresponds to data for the i th nucleotide of A . Then, we model mixtures of ratio p of structure A and $(1 - p)$ of structure B with the linear interpolation, $M(p) = \{pS_i + (1 - p)T_i\}$. We input the Bistable RNA sequence and $M(p)$ into the prediction models from Deigan et al. (2009), Zarringhalam et al. (2012), Washietl et al. (2012), and Spasic et al. (2018), for each value of p between 0 and 1, incremented by 0.05. We sampled 1000 structures from the Boltzmann distribution generated by each model.

Due to the simplicity of the Bistable sequence, over 92% of the structures in every sample closely matched structure A or B (with F-measure above 90%). Thus, we instead measure how well the conformational ratio is predicted in Fig. 1 by using a total variation distance, as described in Sect. 2.4. Rsample has the best performance overall, with an average total variation distance of under 0.1. Note that every method must get some ratio correct as long as the pseudoenergies they use are a continuous function of the auxiliary data. Thus, even methods which perform poorly overall will have a small range of p values where they are successful. Also, the total variation distance is typically small for values of p near 0 and 1, confirming that the SHAPE data is able to direct single-conformation predictions.

To gain insight into why the total variation distance is sometimes large, we look at the breakdown of the samples into structural modes A and B in Fig. 1. For all but one of the methods, we see that as the weighting of structure A increases in the auxiliary data, the predicted conformational ratio increases in the sample. But, the total variation distance is high for the Deigan et al. and Zarringhalam et al. methods because structure A dominates over structure B for most p values.

Unfortunately, predictions are worse for other sequences. In Fig. 2, we ran the same experiment with the HIV-1 RRA sequence from the test set in the Rsample publication (Spasic et al. 2018). Here, Rsample consistently overweights one of the two known HIV-1 RRA conformations, as the other models did for the Bistable sequence. However, Rsample still performs the best out of the methods we tested, with an average total variation distance of 0.23. Because the HIV-1 RRA sequence is more complicated, some samples contained structures whose sets of helices did not match either known conformation, as shown in green in Fig. 2.

In this paper, we will show mathematically why it is difficult to use pseudoenergy models to reconstruct conformational ratios. In Sect. 4, we will identify conditions in pseudoenergy predictions that can lead to the skewed samples in Fig. 1, and in turn, high total variation distances. Then, we show that under the Nussinov–Jacobson energy model and a simple

random model for RNA secondary structures, a large set of RNA structures have many conformational ratios that are hard to reconstruct. Finally, in Sect. 5, we use information theory to show how the ability to reconstruct a conformational ratio depends on the constituent structures in a distribution.

4 Pseudoenergy Models with Noiseless Auxiliary Data

Here, we look at a simple energy model for RNA structures to illustrate the challenges in reconstructing conformational weightings common to all energy models. The Nussinov–Jacobson energy model, based on the algorithm from Nussinov and Jacobson (1980) and analyzed in Clote (2017); Clote et al. (2007), assigns energies to structures by totaling the number of base pairs in the structure. Thus, the structure with the most base pairs is considered optimal. With this in mind, for any structure A , define $[\#bp](A)$ to be the number of paired nucleotides in A . Then, we define an extension to the Nussinov–Jacobson model that incorporates auxiliary data: the energy of structure A given auxiliary data sequence $M = \{M_i\}$ is:

$$\mathcal{E}_{\text{NJ}}(A \mid M) = -[\#bp](A) + \sum_{i=1}^n C|x_i - M_i|, \quad (2)$$

where $C > 0$ is a model parameter, and x_i is 1 if the i th nucleotide of A is unpaired, or 0 if it is paired. The auxiliary data pseudoenergy term is motivated by the Zarringhalam et al. model from Zarringhalam et al. (2012). Here, the auxiliary data is rescaled to the range $[0, 1]$, so that $|x_i - M_i|$ is an agreement score between the structure and the data. Details on our data model are below.

Although the Nussinov–Jacobson model is much simpler than the models used to predict secondary structures, it will still illustrate the problems inherent to other pseudoenergy models. In particular, the parameter C determines how much the auxiliary data should be weighted relative to the underlying energies assigned to each structure. The Deigan et al. model, Zarringhalam et al. model, and Rsample have similar parameters. We consider extreme values of C : when C is small, one structure may dominate over the other structure in a bimodal mixture because the pseudoenergy is not strong enough to overcome differences in thermodynamic energy. On the other hand, if C is very large, the pseudoenergy is large enough that the Boltzmann distribution is collapsed towards the single structure most compatible with the data.

In order to analyze the Nussinov–Jacobson model, we make two simplifications. First, we remove all noise from the auxiliary data. That is, the auxiliary data sequence corresponding to a structure A will be a sequence of zeros and ones, where a zero represents that the nucleotide is paired, and a one represents that the nucleotide is unpaired. Then, we will consider selecting approximate structures at random as follows: for each nucleotide in a sequence, we will choose whether it is paired or unpaired independently of all other nucleotides with probability $q \in (0, 1)$. Choosing $q \approx 0.6$ corresponds to the approximate percentage of paired nucleotides in RNA sequences, where choosing $q = 0.5$ would give all structures equal probability. We will show the following:

Theorem 1 Fix $q \in (0, 1)$ and fix $C > 0$ in the Nussinov–Jacobson model. Consider RNA structures where each nucleotide is paired independently with probability q . Let P_n be the probability that for two randomly chosen structures of length n , at least 25% of the possible conformational ratios cannot be reconstructed with the Nussinov–Jacobson model with total variation distance less than 0.25. Then, $\lim_{n \rightarrow \infty} P_n = 1$.

This indicates that regardless of the parameterization of the Nussinov–Jacobson model, almost all structures of long enough length have a large set of ratios that cannot be reconstructed even when the auxiliary data is noiseless. In Fig. 3, we will show that even for small values of n , P_n is large. Although a similar analysis becomes much more complicated for other pseudoenergy models, there are examples of structures for the Deigan et al. and Zarringhalam et al. models where conformational ratios cannot be recovered regardless of the parameters in the models.

To quantify the problems in the Nussinov–Jacobson model when C is large or small, we turn back to the simulated experiments from Sect. 3. We saw in Fig. 1 that for Rsample, as p increases, the predicted conformational ratio in the sample increases nearly linearly, which is the target behavior we would hope to see for all multimodal distributions. However, in the HIV-1 RRA samples in Fig. 2, the red structure consistently dominates over the blue structure for the majority of p values. We aim to characterize this problem mathematically.

Consider a general RNA sequence with two stable structures A and B . In this section, we model a noiseless data sequence $\{S_i\}$ by defining $S_i = 0$ if the i th nucleotide in A is paired, and 1 if it is unpaired. We similarly define a binary auxiliary data sequence $\{T_i\}$ for B . Then, as before, we let $M(p) = \{pS_i + (1 - p)T_i\}$ be an auxiliary data sequence corresponding to a mixture of ratio p of structure A and ratio $(1 - p)$ of structure B . We define $[\#bp](A)$ to be the number of nucleotides paired in A as above, and also $[\#bp](A - B)$ to be the number of nucleotides paired in A that are not paired in B . Finally, we let $[\#bp](A \oplus B) = [\#bp](A - B) + [\#bp](B - A)$.

We define the *crossover point* for structures A and B to be the value of p where two structures appear with equal probability in the predicted Boltzmann distribution. In an accurate prediction, the crossover point should be near $p = 0.5$, but for Rsample in the HIV-1 RRA example, it is above $p = 0.7$. This means that there is a weighting of the HIV-1 RRA structures where the total variation distance between the correct and predicted distributions is at least 0.2.

Next, we see that the range of values where both structures appear in Fig. 2 is small: for Rsample, one of the two structures was barely present in all proportions $p \geq 0.3$. Thus, we define the *crossover window* for structures A and B to be the range of p values for which both structures appear with probability at least 0.1. We would expect the crossover window to be of width approximately 0.8 (corresponding to the range $0.1 \leq p \leq 0.9$), but it is less than 0.5 in this example. Below, we find that a short crossover window leads to high total variation distances. One advantage of using crossover windows to bound total variation distances is that the ratio of probabilities of structures A and B is independent of the partition function in the Boltzmann distribution.

We will see that when C is relatively small, this can lead to uncentered crossover points, because the data pseudoenergies are not strong enough to overcome the energy differences between competing structures. On the other hand, when C is large, a small change in the data can lead to large changes in energy, which will result in short crossover windows. Now, we find two simple results about crossover points and crossover windows under the Nussinov–Jacobson model:

Lemma 2 *Consider an RNA sequence with two structures A and B that have corresponding binary auxiliary data sequences $\{S_i\}$ and $\{T_i\}$. In the Nussinov–Jacobson SHAPE-directed energy model, the crossover point for structures A and B is given by*

$$p^* = \frac{1}{2} + \frac{1}{2C} - \frac{[\#bp](A-B)}{C[\#bp](A\Delta B)}.$$

Proof Each nucleotide counted in $[\#bp](A-B)$ contributes $C(1-p)$ to the energy of A because the nucleotide is paired in A (so $x_i=0$), and the auxiliary data value at this nucleotide is $(1-p)$, because B has an unpaired nucleotide here. Similarly, each nucleotide counted in $[\#bp](B-A)$ also contributes $C(1-p)$ because $x_i=1$ and the auxiliary data value is p . Thus,

$$\mathcal{E}_{\text{NJ}}(A \mid M(p)) = -[\#bp](A) + [\#bp](A-B)C(1-p) + [\#bp](B-A)C(1-p).$$

A similar equation holds for B , and the crossover point p^* satisfies the equation where the energies for A and B are equal. This gives:

$$p^* = \frac{1}{2} + \frac{[\#bp](B) - [\#bp](A)}{2C([\#bp](A-B) + [\#bp](B-A))}.$$

The set-theoretic relation $[\#bp](A) + [\#bp](B-A) - [\#bp](A-B) = [\#bp](B)$ can be used to simplify the equation for p^* to the one given in the lemma. $\square$

Lemma 3 *Consider an RNA sequence with two structures A and B that have corresponding binary auxiliary data sequences $\{S_i\}$ and $\{T_i\}$. In the Nussinov–Jacobson SHAPE-directed energy model, the crossover window for structures A and B has length at most*

$$\frac{RT \ln 9}{C[\#bp](A\Delta B)}.$$

Proof In order to have at least 10% of each structure appearing in the SHAPE-directed distribution, a necessary (but insufficient) pair of conditions is:

$$\begin{aligned} \mathbb{P}(A \mid M(p)) &\geq 0.1 \geq \frac{1}{9} \mathbb{P}(B \mid M(p)), \\ \mathbb{P}(B \mid M(p)) &\geq \frac{1}{9} \mathbb{P}(A \mid M(p)). \end{aligned} \tag{3}$$

The advantage of these conditions is that by including probabilities on both sides of the inequality, the partition function Z can be cancelled from the inequality entirely. Inserting the expressions for $\mathcal{E}_{NJ}(A | M(p))$ and $\mathcal{E}_{NJ}(B | M(p))$ from the proof of Theorem 2 into Eq. (3) and taking natural logarithms yields the equivalent set of inequalities,

$$[\#bp](B) - [\#bp](A) - RT \ln 9 \leq C(2p - 1)[\#bp](AB) \leq [\#bp](B) - [\#bp](A) + RT \ln 9.$$

In order to find an upper bound for the length of the crossover window, we must find how slowly the expression $C(2p - 1)[\#bp](AB)$ can increase from the lower limit to the upper limit, as a function of p . Recognizing that the difference between the lower and upper limits is $2RT \ln 9$, and letting ℓ be the length of the crossover window, this is equivalent to:

$$2C[\#bp](AB)\ell \leq 2RT \ln 9,$$

which can be rearranged to finish the proof. $\square$

Note that as C becomes very large, the width of the crossover window tends to zero and the crossover point tends to 0.5. As a result, the predicted distributions become almost entirely structure A for $p > 0.5$, and almost entirely structure B for $p < 0.5$. With Lemmas 2 and 3, we are ready to show that the Nussinov–Jacobson model fails to reconstruct conformational ratios for most randomly chosen pairs of structures.

Proof of Theorem 1 We will break into cases, depending on whether the parameter $C \ll n^{-3/4}$ or $C > n^{-3/4}$. If C is large, then small changes in experimental data lead to big changes in predicted conformational ratios, causing the crossover window to be narrow. On the other hand, if C is small, the data is not weighted heavily, causing uncentered crossover points. As we will see, both of these conditions will lead to large maximum total variation distances.

Fix any small $\epsilon > 0$. We will show that for n sufficiently large, $P_n = 1 - \epsilon$, where P_n is the probability that a randomly selected pair of structures of length n has at least 25% of the possible conformational ratios that cannot be reconstructed with total variation distance less than 0.25.

Case 1: $C \ll n^{-3/4}$. In this case, we will argue that the crossover point is uncentered with high probability, which will lead to large total variation distances with high probability. In particular, if the crossover point $p^* \neq 1$ for structures A and B , then the total variation distance must be at least 0.25 for all $p \in [0.75, 1]$. (Note that $p^* > 1$ is not physically possible, and corresponds to a situation where the auxiliary data is not weighted enough in the model to ever favor structure A over structure B .) By Lemma 2, $p^* \neq 1$ is equivalent to:

$$[\#bp](A - B) \leq \left(\frac{1}{2C} - \frac{1}{2}\right)C[\#bp](AB).$$

Then, let i represent the number of nucleotides where the two structures are both paired or both unpaired, so that $[\#bp](A - B) = n - i$. If we condition on knowing i in advance, then we can see that $p^* \neq 1$ is equivalent to:

$$[\# \text{bp}](A - B) \leq \left(\frac{1}{2} - \frac{C}{2}\right)(n - i).$$

Note that for all $0 \leq i \leq n$, the right-hand side is nonpositive and nonincreasing in C . Thus, this inequality is most restrictive on $[\# \text{bp}](A - B)$ when C is the largest possible value, which we have assumed is when $C = n^{-3/4}$. Therefore,

$$\begin{aligned} \mathbb{P}(p^* \geq 1) &\geq \sum_{i=0}^n \mathbb{P}([\# \text{bp}](A \Delta B) = n - i) \\ &\cdot \mathbb{P}\left([\# \text{bp}](A - B) \leq \left(\frac{1}{2} - \frac{n^{-3/4}}{2}\right)(n - i)\right). \end{aligned}$$

If we let $j = [\# \text{bp}](A - B)$, we can use our model where a nucleotide is paired or unpaired independently with probability q to rewrite the sum:

$$\mathbb{P}(p^* \geq 1) = \sum_{i=0}^n \binom{n}{i} (q^2 + (1 - q)^2)^i (q(1 - q))^{n-i} \sum_{j=0}^{\left\lfloor \left(\frac{1}{2} - \frac{n^{-3/4}}{2}\right)(n-i) \right\rfloor} \binom{n-i}{j}. \quad (4)$$

We will show that the inner sum is at least $2^{n-i-1} \left(1 - \frac{\epsilon}{2}\right)$ for all $i \leq n$, for n sufficiently large. Applying this inequality will allow us to approximate the remaining sum using a normal distribution. In order to show this, we will interpret the inner sum as a scaled cumulative distribution function for a binomial distribution. Let $k = n - i$. We have:

$$\sum_{j=0}^{\left\lfloor \left(\frac{1}{2} - \frac{n^{-3/4}}{2}\right)(n-i) \right\rfloor} \binom{n-i}{j} = 2^k \sum_{j=0}^{\left\lfloor \left(\frac{1}{2} - \frac{n^{-3/4}}{2}\right)k \right\rfloor} \binom{k}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{k-j}. \quad (5)$$

Now, let $\{X_m\}$ be a sequence of independent and identically distributed Bernoulli random variables with parameter $1/2$. Then,

$$\begin{aligned} \sum_{j=0}^{\left\lfloor \left(\frac{1}{2} - \frac{n^{-3/4}}{2}\right)k \right\rfloor} \binom{k}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{k-j} &= \mathbb{P}\left(X_1 + \dots + X_k \leq \left(\frac{1}{2} - \frac{n^{-3/4}}{2}\right)k\right) \\ &= \mathbb{P}\left(\frac{X_1 + \dots + X_k - k/2}{\sqrt{k}/2} \leq -n^{-3/4}\sqrt{k}\right). \end{aligned}$$

Here, the Central Limit Theorem implies that this renormalized sum of Bernoulli random variables approaches a standard normal distribution (where we used that the mean and standard deviation of each X_m are both $1/2$). Also, because the Bernoulli distribution has moments of all orders, we can apply the Berry–Esseen theorem to get a bound on how quickly this probability approaches the approximation given by the normal distribution [see, for example, Theorem 3.4.17 in Durrett (2019)]. Let B be the constant from the Berry–Esseen theorem that depends only on the parameter of the Bernoulli random variables, $1/2$,

and let Φ be the cumulative distribution function for a normal random variable with mean 0 and standard deviation 1. Then, we have:

$$\mathbb{P}\left(\frac{X_1 + \dots + X_k - k/2}{\sqrt{k}/2} \leq -n^{-3/4}\sqrt{k}\right) \geq \Phi\left(-n^{-3/4}\sqrt{k}\right) - \frac{B}{\sqrt{k}}.$$

Since $k = n - i \rightarrow n$, as n approaches infinity, we see that the input to Φ approaches 0 as n grows large. Additionally, if we assume $k > C_\epsilon$ for some large constant C_ϵ that depends on ϵ but is independent of n , then we can assume:

$$\Phi\left(-n^{-3/4}\sqrt{k}\right) - \frac{B}{\sqrt{k}} \geq \Phi(0)\left(1 - \frac{\epsilon}{2}\right) = \frac{1}{2}\left(1 - \frac{\epsilon}{2}\right).$$

Tracing back to Eq. (5), we have shown for n sufficiently large and for $k > C_\epsilon$,

$$\sum_{j=0}^{\left\lfloor \left(\frac{1}{2} - \frac{n^{-3/4}}{2}\right)k \right\rfloor} \binom{k}{j} \geq 2^{k-1} \left(1 - \frac{\epsilon}{2}\right). \quad (6)$$

Recall that $k = n - i$, so that $k > C_\epsilon$ is equivalent to $i < n - C_\epsilon$. Thus, substituting Eq. (6) into Eq. (4) when $i < n - C_\epsilon$, and dropping the terms where $i \geq n - C_\epsilon$, gives us:

$$\begin{aligned} \mathbb{P}(p^* \geq 1) &\geq \sum_{i=0}^{\lfloor n - C_\epsilon - 1 \rfloor} \binom{n}{i} (q^2 + (1-q)^2)^i (q(1-q))^{n-i} 2^{n-i-1} \left(1 - \frac{\epsilon}{2}\right) \\ &= \frac{1}{2} \left(1 - \frac{\epsilon}{2}\right) \sum_{i=0}^{\lfloor n - C_\epsilon - 1 \rfloor} \binom{n}{i} (q^2 + (1-q)^2)^i (2q(1-q))^{n-i} \\ &= \frac{1}{2} \left(1 - \frac{\epsilon}{2}\right) \left(1 - \sum_{i=\lfloor n - C_\epsilon \rfloor}^n \binom{n}{i} (q^2 + (1-q)^2)^i (2q(1-q))^{n-i}\right), \end{aligned}$$

where the last line is true because the summation again represents a binomial distribution, this time with parameter $q^2 + (1-q)^2$. Because C_ϵ is independent of n , the last sum approaches 0 as n becomes large. For this reason, for n sufficiently large,

$$\mathbb{P}(p^* \geq 1) \geq \frac{1}{2}(1 - \epsilon).$$

Finally, doubling through symmetry, we see that if $C \leq C_n^*$, then for all n sufficiently large,

$$\mathbb{P}(p^* \leq 0 \text{ or } p^* \geq 1) \geq 1 - \epsilon,$$

proving in this case, $P_n \rightarrow 1 - \epsilon$. Case 2: Now, assume that $C > n^{-3/4}$. Our goal will be to show that the crossover window for any two random structures has width at most 0.05 with probability at least $1 - \epsilon$. Note that when the crossover window has width at most 0.05, then this forces the total variation distance to be at least 0.25 for at least 25% of the possible

conformational ratios p : in the best-case scenario, the crossover window would be centered around $p = 0.5$, in which case the total variation distance for the ranges $0.35 \leq p \leq 0.475$ and $0.525 \leq p \leq 0.65$ is at least 0.25.

Now, let Δ be the width of the crossover window for randomly selected structures A and B . Then, we know from Lemma 3 that

$$\ell \leq \frac{RT \ln 9}{C[\# \text{ bp}](A\Delta B)},$$

and so we will set

$$\frac{RT \ln 9}{C[\# \text{ bp}](A\Delta B)} \leq 0.05.$$

This is equivalent to

$$[\# \text{ bp}](A\Delta B) \geq \frac{RT \ln 9}{C \cdot 0.05}.$$

Using that $C > n^{3/4}$, we have:

$$\begin{aligned} \mathbb{P}(\ell \leq 0.05) &\geq \mathbb{P}\left([\# \text{ bp}](A\Delta B) \geq \frac{RT \ln 9}{0.05n^{-3/4}}\right) \\ &\geq \mathbb{P}\left([\# \text{ bp}](A\Delta B) \geq [20RT \ln 9]n^{3/4}\right) \\ &= 1 - \mathbb{P}\left([\# \text{ bp}](A\Delta B) < [20RT \ln 9]n^{3/4}\right) \\ &= 1 - \sum_{i=0}^{\lfloor [20RT \ln 9]n^{3/4} \rfloor} \binom{n}{i} (2q(1-q))^i (q^2 + (1-q)^2)^{n-i}. \end{aligned}$$

Now, we repeat the procedure from Case 1 to analyze the remaining sum: let $\{Y_m\}$ be a sequence of independent Bernoulli random variables with parameter $2q(1-q)$, and standard deviation σ_q . Then,

$$\begin{aligned} &\sum_{i=0}^{\lfloor [20RT \ln 9]n^{3/4} \rfloor} \binom{n}{i} (2q(1-q))^i (q^2 + (1-q)^2)^{n-i} = \mathbb{P}(Y_1 + \dots + Y_n \leq [20RT \ln 9]n^{3/4}) \\ &= \mathbb{P}\left(\frac{Y_1 + \dots + Y_n - n(2q(1-q))}{\sigma_q \sqrt{n}} \leq \frac{[20RT \ln 9]n^{3/4} - n(2q(1-q))}{\sigma_q \sqrt{n}}\right) \leq \Phi\left(\frac{[20RT \ln 9]n^{1/4} - \sqrt{n}(2q(1-q))}{\sigma_q}\right) \\ &\quad + \frac{B_q}{\sqrt{n}}, \end{aligned}$$

where the last line is again due to the Berry–Esseen theorem, and the constant B_q is a potentially larger constant than B in Case 1 that depends only on the parameter of the Bernoulli random variables, $\{Y_m\}$. Here, we see that the input to Φ approaches negative infinity as n approaches infinity. Additionally, the $B_q/\sqrt{n}$ term is arbitrarily small as n increases. Thus, we can conclude that for n sufficiently large,

$$\sum_{i=0}^{\lfloor [20RT \ln 9]n^{3/4} \rfloor} \binom{n}{i} (2q(1-q))^i (q^2 + (1-q)^2)^{n-i} < \epsilon,$$

from which we can conclude that for all $C > n^{-3/4}$, for n sufficiently large,

$$\mathbb{P}(\ell \leq 0.05) \geq 1 - \epsilon.$$

Again, this implies that when $C > n^{-3/4}$, $P_n \rightarrow 1 - \epsilon$. Thus, between Case 1 and Case 2, we found sufficient conditions for small and large C where the total variation distance was above 0.25 for at least 25% of conformational ratios with high probability, which completes the proof. $\square$

Theorem 1 tells us that the Nussinov model is unsuccessful at recreating conformational ratios for long RNA sequences. We can also use the same proof structure to compute the probability that conformational ratios cannot be constructed for short RNA sequences. The challenging part of such an analysis is picking the optimal cutoff C_n^* between the two cases. As the cutoff increases, the probability in Case 1 increases while the probability in Case 2 decreases. However, the probabilities in Case 2 are a step function of the cutoff, indicating that there are only finitely many cutoff values to consider for each value of n to optimize the bound. Figure 3 illustrates a lower bound on the probability that 20% or more of conformational ratios cannot be constructed with a total variation distance less than 0.25 for a sequence of length n . Here, the probability of a position being paired was set to $q = 0.6$ (corresponding to approximately the percentage of pairings in RNA structures). For all sequences above length 80, we see that there is at least a 30% chance that a range of conformational ratios cannot be reconstructed.

5 Information Content in Auxiliary Data

We have found evidence that pseudoenergy-based prediction models often cannot reconstruct the weightings in a bimodal distribution of structures, even if the data is noiseless. But, how much weighting information is contained in the data itself? In Li and Reidys (2019), the authors use information theory to predict RNA structures. Here, we quantify how well structural weightings can be recovered by using Fisher information.

In examining the information in the data, we separate the prediction of structures from the prediction of weightings. This is motivated by the “sample and select” algorithm from Quarrier et al. (2010), where a sample of putative structures is generated from the undirected Boltzmann distribution, and then the structure which agrees most closely with the auxiliary data is selected. For multimodal distributions, this method would need to be refined so that a set of putative structural modes is selected, and then the auxiliary data is used to predict the conformational ratios between these structures.

Here, we assume that the structures in a bimodal distribution are already known in advance. This represents the best-case scenario, where the only unknown in a distribution of

structures is the conformational ratio. While this may sound like a strong assumption, signals from the component structures of a distribution often appear in the undirected Boltzmann distribution. In particular, RNA Profiling (Rogers and Heitsch 2014), revealed signals from both structures for the Bistable and HIV RRA sequences. Additionally, as we have seen in Sect. 3 above, Rsample (Spasic et al. 2018), is often able to identify the correct structures within a distribution, regardless of whether the conformational ratio is correct.

Thus, given a sequence of auxiliary data corresponding to a mixture of two known structures A and B , we aim to reliably predict the conformational ratio p of structure A in the distribution. We will show first through example, and then rigorously via Fisher information, that the ability to recover p depends on the number of locations where structures A and B differ. To the best of our knowledge, this is the first analysis of how the ability to predict p depends on the structures in a distribution. Although we focus on SHAPE data below, the framework below applies to any type of auxiliary data that gives information about unpairedness for individual nucleotides.

5.1 Example

Consider the HIV-1 RRA sequence from Sect. 3. We use the same simulated sequences $\{T_i\}$ and $\{S_i\}$ for each of the two known HIV-1 RRA structures. Then, as in Sect. 3, we mix these sequences together with various proportions, p , to get a new SHAPE sequence $M(p) = \{pS_i + (1-p)T_i\}$.

Now, with only the numerical values of the sequence $M(p)$ and the two HIV-1 RRA structures, we check whether we can recover the value, p . We use only the values from $M(p)$ corresponding to where structures A and B differ because the SHAPE values come from different distributions at these points, and are expected to carry the most information about p . Each of these values is drawn from a convolution of two different known distributions, with an unknown weighting. A common way to predict the weighting is to use a maximum likelihood estimator, which finds the value $\hat{p}$ which maximizes the probability of producing the observed auxiliary data sequence. In summary, we simulated sequences $\{T_i\}$ and $\{S_i\}$, mixed them in different ratios p , and used the maximum likelihood estimator $\hat{p}$ to estimate p from $M(p)$. The results are shown in Fig. 4.

Here, the estimator successfully recovered the p values with a maximum error of 0.11, which is an improvement over the pseudoenergy models in Fig. 2. However, if we repeat the same procedure on the 16S rRNA Four way Junction from the test set in Spasic et al. (2018), then the maximum error is 0.24. This can be explained by enumerating the structural differences between the modes: the two HIV-1 RRA structures differ in 31 positions, while the two 16S rRNA structures differ in only 16 positions.

5.2 Fisher Information

To assess how accurate an estimator $\hat{p}$ can recover p , we study the variance of $\hat{p}$. The more the structures in a distribution differ, the more information the auxiliary data is expected to carry about p , which should lead to a lower variance of $\hat{p}$. This means that the variance of $\hat{p}$

is not directly a function of the length of the RNA sequences, but is instead a function of the number of locations where the structures differ.

To quantify the relationship between the variance of an estimator and the number of structural differences between modes, we use Fisher information. The Fisher information of a random variable, defined below, describes how much information the random variable contains about a parameter. By using Fisher information and the model of SHAPE data from Sükösd et al. (2013), we will be able to conclude statements like the following:

Lemma 4 *Consider a sequence of n simulated SHAPE values derived from an equal mixture of two structures, where at each position, one structure is center-paired and the other structure is unpaired. Let $\hat{p}$ be any unbiased estimator of the true conformational ratio, $p = 0.5$. Then,*

$$\text{Var}(\hat{p}) \geq \frac{1}{2.15n}.$$

To prove Lemma 4, consider a single shape value $X = pX_1 + (1 - p)X_2$ corresponding to a nucleotide that is unpaired with SHAPE value X_1 in proportion p of structures, and center-paired with SHAPE value X_2 in proportion $1 - p$ of structures. We choose to compare these distributions because they are the most different, and represent the scenario where p should be recovered most easily. We describe how to modify the approach to include edge-paired nucleotides below.

The Fisher information is a measure of how much information a single random observation of X gives about a fixed parameter p on average. Some values of X give less information about p than others. For example, if X is a middling value, it is unclear whether X is the average of a paired and unpaired signal, or whether X is a lower-than-usual unpaired signal.

Formally, let $g(x, p)$ be the probability density function of the random variable X with parameter p . For a fixed value of x but varied values of p , if $g(x, p)$ has a sharp maximum, then a maximum likelihood estimator should be able to recover p more easily. This is quantified by the Fisher information:

$$I(p) := \mathbb{E} \left[\left(\frac{\partial}{\partial p} \log g(X, p) \right)^2 \right] = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial p} \log g(x, p) \right)^2 g(x, p) dx. \quad (7)$$

The Cramér–Rao bound illustrates the utility of Fisher information: for any unbiased estimator $\hat{p}$ of a parameter p ,

$$\text{Var} \hat{p} \geq \frac{1}{I(p)}.$$

Additionally, when collecting independent samples, the Fisher information is additive. Therefore, if we have a sample of n SHAPE values corresponding to nucleotides that are unpaired in proportion p of structures, and center-paired in proportion $1 - p$ of structures, then for any unbiased estimator $\hat{p}$ of p ,

$$\text{Var } \hat{p} \geq \frac{1}{nI(p)}.$$

Lemma 4 follows from calculating the Fisher information for $p = 0.5$. Typically, the maximum likelihood estimator for a parameter is not unbiased, but under mild regularity conditions, it asymptotically approaches an unbiased estimator with minimal variance, given by $\frac{1}{nI(p)}$ as n grows large (see, for example, Theorem 9.18 in Wasserman (2004)).

The Fisher information could also be used to describe the variance of an estimator when the true value is not $p = 1/2$, by keeping track of more details about the structural differences in the modes. If there are k positions where structure A is unpaired and structure B is paired, and ℓ positions where structure A is center-paired and structure B is unpaired, and if structure A has conformational ratio p , then

$$\text{Var } \hat{p} \geq \frac{1}{kI(p) + \ell I(1-p)}.$$

This could be extended to include edge-paired nucleotides by considering more terms.

Now, we compute the Fisher information. Since $g(x, p)$ is the density function for a weighted sum, it is the weighted convolution of the paired and unpaired SHAPE distributions:

$$g(x, p) = \int_0^{y+0.025(1-p)} \frac{1}{p(1-p)} f\left(\frac{y}{p}\right) f\left(\frac{x-y}{1-p}\right) dy.$$

Here, f and f represent the unpaired and center-paired SHAPE probability distribution functions from Sükösd et al. (2013), respectively. $f(x)$ is a generalized extreme value distribution with parameters $\xi = 0.762581$, $\sigma = 0.0492536$, and $\mu = 0.0395857$. $f(x)$ is an exponential distribution with parameter $\lambda = 1.46797$. The domain of integration is the support of the integrand.

To evaluate the Fisher information, we note:

$$\frac{\partial}{\partial p} \ln g(x, p) = \frac{1}{g(x, p)} \cdot \frac{\partial}{\partial p} g(x, p).$$

Because f and f are continuous and have continuous derivatives over their support, and because the bounds of the convolution $g(x, p)$ are continuous with respect to p , we can use the Leibniz rule to evaluate the derivative of $g(x, p)$:

$$\begin{aligned} \frac{\partial}{\partial p} g(x, p) &= \frac{d}{dp} \left(\frac{1}{p(1-p)} \right) \int_0^{y+0.025(1-p)} f\left(\frac{y}{p}\right) f\left(\frac{x-y}{1-p}\right) dy \\ &+ \frac{1}{p(1-p)} \int_0^{y+0.025(1-p)} \frac{\partial}{\partial p} \left(f\left(\frac{y}{p}\right) f\left(\frac{x-y}{1-p}\right) \right) dy. \end{aligned}$$

In its general form, the Leibniz rule also involves terms where the integrand is evaluated at the bounds of the integral, but in this case, these terms are zero. The second term above can be split into two integrals after applying the product rule, allowing us to write $\frac{\partial}{\partial p} g(x, p)$ as the sum of three integrals. Plugging this expression into (7) and using an iterated numerical integration in MATLAB, we evaluated the Fisher information, and the results are shown in Fig. 5.

$I(0.5) = 2.15$, yielding the bound in Lemma 4. This can be leveraged to determine how many structural differences two structural modes need to have in order to accurately recover p values. If we would like the standard deviation of $\hat{p}$ to be less than 0.1, for example, the number of differences $n \geq 47$. In other words, it is only possible to recover p with this level of certainty if our two structures have at least 47 positions where they differ.

6 Conclusion

Recent advances to pseudoenergy models such as Rsample (Spasic et al. 2018) are able to identify multiple structural modes within a distribution. However, it is still challenging to reliably reconstruct the conformational ratios of the modes. Using the Nussinov–Jacobson model, we found that many long RNA sequences tend to have very different structures with similar energies. When this happens, a short crossover window leads pseudoenergy models to inaccurate predictions. Future work could explore the impact that the base composition of a sequence has on reconstructing mixtures of structures.

An alternative approach is to separate predictions of structural modes from their conformational ratios. When the structural modes are identified in advance, more accurate conformational ratios can be found by using estimators from statistics. The variance of the estimator depends on how many differences there are between the structural modes. In practice, when predicting an unknown conformational weighting between structural modes, the more the modes differ, the higher the confidence in the prediction. Fisher information can be used to give an estimated confidence level when using a maximum likelihood estimator to recover the conformational ratio.

Acknowledgements

The authors thank the anonymous referees for valuable comments improving the exposition of the paper and quality of the results. This work was supported by funds from the National Institutes of Health (R01GM126554 to CEH) and the National Science Foundation (DMS1344199 to CEH).

References

- Clote P (2017) An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov–Jacobson energy model. *J Comput Biol* 12(1):83–101
- Clote P, Kranakis E, Krizanc D, Stacho L (2007) Asymptotic expected number of base pairs in optimal secondary structure for random RNA using the Nussinov–Jacobson energy model. *Discrete Appl Math* 155(6):759–787
- Cordero P, Das R (2015) Rich RNA structure landscapes revealed by mutate-and-map analysis. *PLoS Comput Biol* 11(11):e1004473 [PubMed: 26566145]

- Deborah A, Jorge Natasha AN, Caffarena Ernesto R, Fabio P (2018) Using RNA sequence and structure for the prediction of riboswitch aptamer. A comprehensive review of available software and tools. *Front Genet* 8:231 [PubMed: 29403526]
- Deigan KE, Li TW, Mathews DH, Weeks KM (2009) Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci* 106(1):97–102 [PubMed: 19109441]
- Ding Y (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* 31(24):7280–7301 [PubMed: 14654704]
- Durrett R (2019) *Probability: theory and examples*, 5th edn. Cambridge University Press, Cambridge
- Eddy SR (2014) Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu Rev Biophys* 43(1):433–456 [PubMed: 24895857]
- Höbartner C, Micura R (2003) Bistable secondary structures of small RNAs and their structural probing by comparative imino proton NMR spectroscopy. *J Mol Biol* 325(3):421–431 [PubMed: 12498793]
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M (1994) Fast folding and comparison of RNA secondary structures. *Chem Mon* 125(2):167–188
- Leonard CW, Hajdin CE, Karabiber F, Mathews DH, Favorov OV, Dokholyan NV, Weeks KM (2013) Principles for understanding the accuracy of SHAPE-directed RNA structure modeling. *Biochemistry* 52(4):588–595 [PubMed: 23316814]
- Li TJX, Reidys CM (2019) On an enhancement of RNA probing data using information theory. *arXiv preprint*
- Lorenz R, Bernhart SH, Höner zu Siederdisen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA package 2.0. *Algorithms Mol Biol* 6:26 [PubMed: 22115189]
- Lu Z, Cliff Zhang Q, Lee B, Flynn RA, Smith MA, Robinson JT, Davidovich C, Gooding AR, Goodrich KJ, Mattick JS (2016) RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* 165(5):1267–1279 [PubMed: 27180905]
- Mathews DH (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* 10(8):1178–1190 [PubMed: 15272118]
- Mathews DH, Turner DH (2006) Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol* 16(3):270–278 [PubMed: 16713706]
- McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29(6–7):1105–1119 [PubMed: 1695107]
- Nussinov R, Jacobson AB (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci USA* 77(11):6309–6313 [PubMed: 6161375]
- Quarrier S, Martin JS, Davis-Neulander L, Beauregard A, Laederach A (2010) Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA* 16(6):1108–1117 [PubMed: 20413617]
- Rice GM, Leonard CW, Weeks KM (2014) RNA secondary structure modeling at consistent high accuracy using differential shape. *RNA* 20(6):846–854 [PubMed: 24742934]
- Rogers E, Heitsch CE (2014) Profiling small RNA reveals multimodal substructural signals in a Boltzmann ensemble. *Nucleic Acids Res* 42(22):171
- Schroeder SJ (2018) Challenges and approaches to predicting RNA with multiple functional structures. *RNA* 24(12):1615–1624 [PubMed: 30143552]
- Spasic A, Assmann SM, Bevilacqua PC, Mathews DH (2018) Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res* 46(1):314–323 [PubMed: 29177466]
- Sükösd Z, Swenson MS, Kjems J, Heitsch CE (2013) Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Res* 41(5):2807–2816 [PubMed: 23325843]
- Turner DH, Mathews DH (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res* 38(1):280–282
- Vieweger M, Nesbitt DJ (2018) Synergistic SHAPE/single-molecule deconvolution of RNA conformation under physiological conditions. *Biophys J* 114(8):1762–1775 [PubMed: 29694857]

- Washietl S, Hofacker IL, Stadler PF, Kellis M (2012) RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res* 40(10):4261–4272 [PubMed: 22287623]
- Wasserman L (2004) *All of statistics: a concise course in statistical inference*, 1st edn. Springer, Berlin
- Woods CT, Lackey L, Williams B, Dokholyan NV, Gotz D, Laederach A (2017) Comparative visualization of the RNA suboptimal conformational ensemble in vivo. *Biophys J* 113(2):290–301 [PubMed: 28625696]
- Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49(2):145–165 [PubMed: 10070264]
- Zarrinhalam K, Meyer MM, Dotu I, Chuang JH, Clote P (2012) Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS ONE* 7(10):1–13
- Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. *Science* 244(4900):48–52 [PubMed: 2468181]
- Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9(1):133–148 [PubMed: 6163133]

Fig. 1.

Each prediction method generated samples of 1000 Bistable RNA structures for each mixed auxiliary data sequence $M(p)$, with p values ranging from 0 to 1. The upper left graph illustrates the error in the conformational ratio, measured using a total variation distance. The bar graphs show the breakdown of samples into the two corresponding structural modes

Fig. 2.

We repeated the process from Fig. 1 for two HIV- 1 RRA structures

Fig. 3.

The plot above gives a lower bound for the probability that a randomly selected pair of structures of length n has at least 20% of the possible conformational ratios that cannot accurately be reconstructed by the Nussinov–Jacobson model

Fig. 4.

The maximum error for the 16S rRNA Four way Junction is higher than the maximum error for the HIV RRA sequence. This is because there are fewer structural differences between the 16S structures than the HIV-1 RRA structures

Fig. 5.

For each value of p , the Fisher information $I(p)$ was calculated by using numerical integration in MATLAB. The vertical axis is on a logarithmic scale