DECONVOLVING RNA BASE PAIRING SIGNALS

TORIN GREENWOOD AND CHRISTINE E. HEITSCH

Abstract. The structure of an RNA sequence encodes information about its biological function. A sequence is typically predicted to fold to a single minimum free energy conformation. But, an increasing number of RNA molecules are now known to fold into multiple stable structures. Discrete optimization methods are commonly used to predict foldings, and adding experimental data as auxiliary information improves prediction accuracy when there is a single dominant conformation. In this paper, we analyze the outputs of existing structural prediction models when they receive auxiliary data derived from a mixture of structures. Under a binary model of auxiliary data, we find that current structural prediction methods typically favor distributions with one dominant structure, and hence cannot guarantee accurate reconstruction of multimodal distributions. Additionally, we analyze empirical distributions of auxiliary data used in current prediction models. We show that even when the structures in a distribution are known in advance, it is difficult to determine the weightings of the structures using auxiliary data. RNA secondary structure and thermodynamic optimization with auxiliary data and method of moments estimators

1. Introduction

The combinatorial arrangement of RNA base pairings encodes functional information. However, it is difficult to determine the structure of an RNA sequence experimentally. Instead, discrete optimization methods are used to predict the most probable folding for an RNA sequence. Because three-dimensional structures can contain complicated bonding relationships between nucleotides, optimization methods instead search for two-dimensional approximations, called secondary structures, that still contain important functional information. A popular class of prediction algorithms (including Zuker and Stiegler (1981) and Ding et al. (2004)) assign energies to every potential structure into which a structure can fold. To do so, the algorithms use the Nearest Neighbor Thermodynamical Model from Mathews et al. (1999) and Mathews and Turner (2006). The energies can be used to define a Boltzmann distribution on all potential structures. Then, dynamic programming algorithms introduced in Zuker and Stiegler (1981) can find the minimum free energy structure, which is the most probable structure in the Boltzmann distribution. Another class of prediction algorithms uses stochastic context-free grammars (SCFGs) to generate structures through recursive rules with probabilities attached to each rule, like the grammars in Eddy and Durbin (1994) and Sakakibara et al. (1994). In this case, the structure with the highest probability of occurring becomes the predicted structure.

Unfortunately, none of these methods have perfect accuracy on their own. To improve these methods, auxiliary experimental data can be incorporated. Through chemical footprinting experiments, every nucleotide in a folded RNA sequence can be assigned a reactivity score. Ideally, the reactivity would be 0 for paired nucleotides and a positive constant A for unpaired nucleotides, but for a variety of experimental reasons, the reactivity signal less clear. Several competing methods of incorporating this noisy data into the thermodynamic prediction model have been proposed, as in Deigan et al. (2009), Zarringhalam et al. (2012), and Quarrier et al. (2010), and the data can also be incorporated into SCFGs. As described in Deigan et al. (2009), when auxiliary data is included, the prediction accuracy of these methods is high enough to recover important structural information about wide classes of RNA sequences.

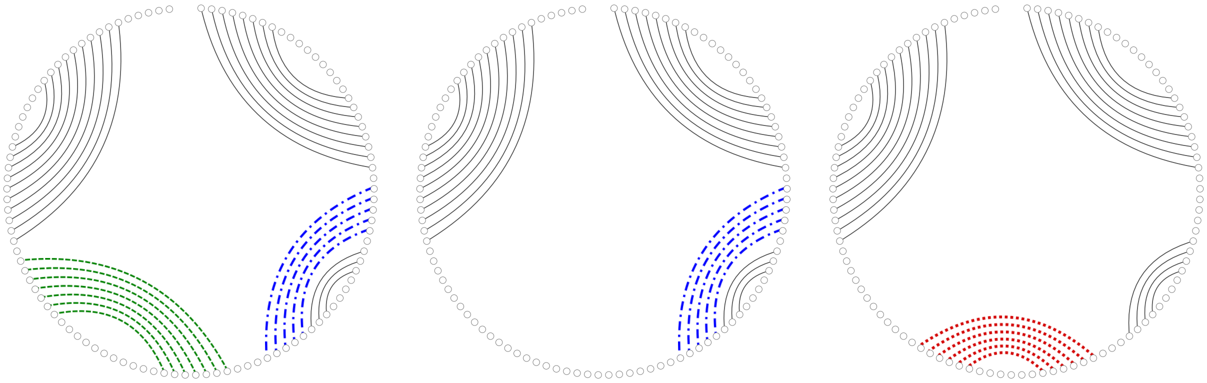

Figure 1. Each diagram represents a potential secondary structure for the same RNA sequence, with dots representing nucleotides and arcs representing bonds. Although all three structures have similarities in gray, the colored regions represent critical differences. In our modelling simulation, current structural prediction methods cannot devolve a 50/50 mixture of the left and right structures. Instead, they inaccurately guess that the ensemble of structures is comprised of nearly 100% the middle structure.

Instead of looking at RNA sequences with one dominant conformation, our goal is to investigate how well existing prediction algorithms are able to predict multimodal distributions when the auxiliary data is derived from a mixture of structural signals. An increasing number of RNA molecules are now known to fold into multiple stable structures, as described in Rogers and Heitsch (2014). In Leonard et al. (2013), the authors uncovered evidence of multimodal structural distributions. We will find shortcomings of existing prediction methods in identifying multimodal distributions, and we will identify improvements that are necessary in the auxiliary biological data in order to accurately predict the weightings of structures in a multimodal structural signal.

2. Challenges in identifying multimodal distributions: an empirical study

In this section, we look at empirical evidence illustrating the challenges in identifying multimodal distributions of RNA secondary structures. Because there is not yet biological data from multimodal distributions of structures, we will simulate biological data in this section by using distributions from Sükösd et al. (2013) that were derived from biological data. The problems in these empirical simulations will guide the intuition behind the more general results we derive later. Background information on RNA folding is saved until Section 4.

Here, we investigate distributions involving two potential secondary structures, R and S, for the same RNA sequence. Each nucleotide in an RNA sequence's secondary structure is either paired to one other nucleotide, or is unpaired, as shown in the examples in Figure 1.

Consider two sequences of auxiliary data corresponding to each of these structures, labelled $\{M_i\}_{i=1}^n$ for structure R and $\{N_i\}_{i=1}^n$ for structure S. To model data coming from a mixture of $100p\%$ of R and $100(1-p)\%$ of S, with $p \in [0,1]$, we use a linear interpolation of the data, $\{pM_i + (1-p)N_i\}_{i=1}^n$. Then, for various values of p, we examine the distribution of structures produced by existing prediction models as they receive interpolated data. We will refer to this as an $S \to R$ interpolation. To gain insight into these distributions, we look at what happens for two specific RNA structures for the RNA sequence VcQrr3. This is a small RNA molecule from the quorum-sensing pathway in the bacterium, *Vibrio cholerae*, studied in Lenz et al. (2004) and Tu and Bassler (2007). Later, in Rogers and Heitsch (2014), the authors found that this sequence can fold into several low-energy conformations, as illustrated below in Figure 3. We will look at the

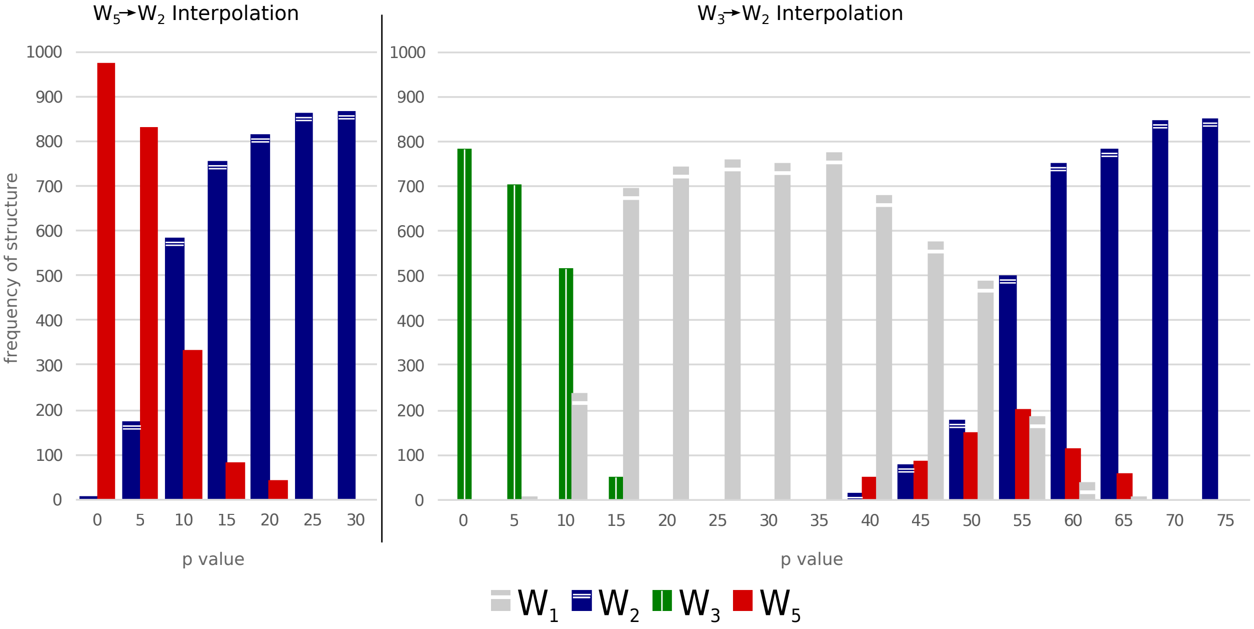

FIGURE 2. Here, we generate a sample of 1000 structures by using the dominant structural prediction method and the auxiliary sequence $pM_1+(1-p)M_2$, for various percentages p from structure W_2.

structures that will be labelled W_3, W_1, and W_2, which are illustrated from left to right in Figure 1. Consider when the corresponding auxiliary sequences M and N are randomly generated using structures W_3 and W_2 and the empirical distributions derived in Sükösd et al. (2013). Using this data for different values of p and the dominant NNTM prediction model described in Section 5.1, we produced 1000 structures for each value of p between 0 and 1, incremented by 0.05. Although the distributions were accurate for $p = 0$ and $p = 1$, a problem occurred for $p = 0.5$: instead of producing a distribution of structures where W_2 and W_3 represented roughly half of the distribution, the algorithm produced a distribution that was nearly 100% comprised of a third structure, labelled W_1. This indicates that the standard model can have trouble identifying multimodal distributions of structures.

To gain more insight into the problems identifying multimodal distributions, we consider two more interpolations, illustrated in Figure 2. Now, we count the frequencies of competing structures in samples generated using mixed auxiliary data. In a $W_5 \to W_2$ interpolation, we see that when 25% of the data comes from structure W_2, the structural prediction model already produces a distribution that is composed nearly entirely of W_2, and W_5 is not represented. With this as motivation, we define the *crossover point* of an $S \to R$ interpolation to be the value of p where R and S occur with equal probability in a distribution. In this example, the crossover point is below $p = 0.1$, and thus is highly uncentered. We ask whether it is possible to reparametrize current models to make crossover points closer to 0.5.

Besides the fact that crossover points may be far from 0.5, another issue is whether the structures that contribute to mixed auxiliary data can be identified reliably. With this in mind, let the *crossover window* be the range of p values for which R and S both occur in at least 20% of the distribution. Crossover windows must span a large range of p values, or the multimodal distribution will not be precisely identifiable. Ideally, a crossover window would have length about 0.6, corresponding to the range $p = 0.2$ to $p = 0.8$ where the mixed data includes at least a 20% contribution from each structure's data. The $W_5 \to W_2$ interpolation indicates that crossover windows can be very short. Also, the $W_3 \to W_2$ interpolation data in Figure 2 shows another problem: a third structure, W_1, has dominated much of the interpolation, and it is no longer evident that the SHAPE data was derived from structures W_2 and W_3. Typically, the greater the number of structural differences between R and S, the greater the number of unexpected structures appearing in an interpolation.

One purpose of this paper is to see that the problems identified in this section are common to both the dominant Deigan structural prediction model and other current structural prediction models. In particular, we will use a binary model of auxiliary data to analyze multimodal distributions, and we will find classes of structures where crossover points are uncentered and crossover windows are narrow. Then, because of the problems with inputting auxiliary data into existing prediction models, we will also investigate distributions for auxiliary data separately. In particular, we look at whether it is possible to recover p when R, S, and the distributions of auxiliary data are known in advance. We will find that it is difficult to recover p with current auxiliary data distributions. However, with a slight modification of the distributions, recovering p becomes viable.

3. Preliminary Results

Before delving into the details of the various models for secondary structure prediction, we examine a simple model. The purpose of this section is to illustrate both the types of results we will find for the other models, as well as the proof techniques. Additionally, we will summarize some of our results about biological auxiliary data.

3.1. **Uncentered crossover points and short crossover windows.**

The empirically-derived auxiliary data distributions from Sükösd et al. (2013) have large variances, and make the analysis of existing prediction models difficult. In order to remove such variability, we generate noiseless auxiliary data signals for structures by assigning a 0 to any paired nucleotide, and assigning 1 to any unpaired nucleotide. For any structure S_j with n nucleotides, let $P_j := \{P_{j,i}\}_{i=1}^n$ be the binary auxiliary sequence corresponding to the structure. Then, we consider mixtures of the binary auxiliary data, defined by $pP_j + (1-p)P_k$, and we call the corresponding interpolation a binary $S_k \to S_j$ interpolation.

Below, we will input these mixed binary auxiliary sequences into several different existing structural prediction models, and analyze the crossover points and crossover windows for each. The particular type of data used in the models is called SHAPE data, described in Section 4.2 below. This data can be incorporated into energy-based models to create new SHAPE-directed energies for each structure. Consider a binary $S_2 \to S_1$ interpolation. Then, regardless of the model, the SHAPE-directed energies depend on only a few features of the structures themselves. First, the models rely on the energies of S_1 and S_2 before the auxiliary data is considered. Second, the models depend on the total number of pairings in each structure. Accordingly, let $[\#\mathrm{bp}](S_1)$ be the number of nucleotides in base pairs in S_1. Finally, the models depend on the number of pairings in S_1 that are not in S_2, and the number pairings in S_2 that are not in S_1. Thus, we let $[\#\mathrm{bp}](S_1 - S_2)$ be the number of paired nucleotides in S_1 that are unpaired in S_2. Also, we let $[\#\mathrm{bp}](S_1 \Delta S_2) = [\#\mathrm{bp}](S_1 - S_2) + [\#\mathrm{bp}](S_2 - S_1)$.

Here, we look at a simple energy model for RNA structures in order to illustrate the types of results and methods used to analyze prediction models. The Nussinov-Jacobson energy model, based on the algorithm from Nussinov and Jacobson (1980) and analyzed in Clote (2005) and 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, we define the energy of structure S given SHAPE sequence M as:

$$\mathcal{E}_{\mathrm{NJ}}(S|M) = -[\#\mathrm{bp}](S) + \sum_{i=1}^{n} C|x_i - M_i|, \tag{1}$$

where $C > 0$ is a model parameter, and x_i is 1 if the ith nucleotide of S is unpaired, or 0 if it is paired. The SHAPE penalty is motivated by the Zarringhalam model from Zarringhalam et al. (2012), analyzed further later. With the Nussinov-Jacobson SHAPE-directed energy model defined, we have the following result about crossover points:

Theorem 1. *In the Nussinov-Jacobson SHAPE-directed energy model, the crossover point in a binary $S_2 \to S_1$ interpolation is given by*

$$p^* = \frac{1}{2} + \frac{1}{2C} - \frac{[\#bp](S_1 - S_2)}{C[\#bp](S_1 \Delta S_2)}.$$

In particular, for all $C \leq 5$ and all structures S_1 and S_2 where $[\#bp](S_1 - S_2) = 0$, we have $p^* \geq 0.6$.

It will not be possible to find an explicit formula for crossover points for the other models. However, we will repeatedly find that crossover points cannot always be centered near $p = 0.5$. The condition that $[\#\text{bp}](S_1 - S_2) = 0$ is the scenario where S_1 has all the pairings in S_2, plus potentially more. In this model, $C = 1$ is always a large enough parameter value to guarantee that the model prefers S_1 when given P_1 and prefers S_2 when given P_2. Thus, a value of $C = 5$ is quite large, and the fact that crossover points can be uncentered for a large class of interpolations when $C \leq 5$ indicates that the model cannot reasonably guarantee centered crossover points.

Theorem 1. Here, we must analyze the penalties of the form $C|x_i - M_i|$ in Equation (1). Each nucleotide counted in $[\#\text{bp}](S_1 - S_2)$ contributes $C(1-p)$ to the energy of S_1 because the nucleotide is paired in S_1 (so $x_i = 0$), and the SHAPE value at this nucleotide is $(1 - p)$, because S_2 has an unpaired nucleotide here. Similarly, each nucleotide counted in $[\#\text{bp}](S_2 - S_1)$ also contributes $C(1 - p)$ because $x_i = 1$ and the SHAPE value is p. Thus,

$$\begin{aligned}&\mathcal{E}_{\text{NJ}}(S_1|pP_1 + (1-p)P_2) = \\ &\qquad - [\#\text{bp}](S_1) + [\#\text{bp}](S_1 - S_2)C(1-p) + [\#\text{bp}](S_2 - S_1)C(1-p).\end{aligned}$$

A similar equation holds for S_2, and the crossover point p^* satisfies the equation where the energies for S_1 and S_2 are equal. This gives:

$$p^* = \frac{1}{2} + \frac{[\#\text{bp}](S_2) - [\#\text{bp}](S_1)}{2C\big[[\#\text{bp}](S_1 - S_2) + [\#\text{bp}](S_2 - S_1)\big]}.$$

The set-theoretic relation $[\#\text{bp}](S_1) + [\#\text{bp}](S_2 - S_1) - [\#\text{bp}](S_1 - S_2) = [\#\text{bp}](S_2)$ can be used to simplify the equation for p^* to the one given in the theorem. Finally, when $[\#\text{bp}](S_1 - S_2) = 0$, we have the simplification $p^* = \frac{1}{2} + \frac{1}{2C}$, which is at least 0.6 when $C \leq 5$. □

Given a fixed RNA sequence, energies can be used to define a Boltzmann distribution of structures, where a particular structure S with energy $\mathcal{E}(S)$ has the probability in the distribution,

$$\mathbb{P}(S) := \frac{e^{-\mathcal{E}(S)/RT}}{Z}, \tag{2}$$

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.

One would expect that as the number of places where the structures S_1 and S_2 differ increases, the SHAPE data will have more of an influence on interpolations, because a SHAPE penalty is applied more frequently. This would mean that as the number of differences increases, both structures appear simultaneously in a distribution less frequently. The following result confirms this, as the crossover window length is expressed as a function inversely proportional to C and $[\#\text{bp}](S_1 \Delta S_2)$:

Theorem 2. *In the Nussinov-Jacobson SHAPE-directed energy model, the binary $S_2 \to S_1$ interpolation has a crossover window with length at most*

$$\frac{RT \ln 4}{C[\#bp](S_1 \Delta S_2)}.$$

Proof. In order to have at least 20% of each structure appearing in the SHAPE-directed distribution, a necessary (but insufficient) pair of conditions is:

$$\mathbb{P}(S_1|pM_1 + (1-p)M_2) \geq \frac{2}{8}\mathbb{P}(S_2|pM_1 + (1-p)M_2),$$

$$\mathbb{P}(S_2|pM_1 + (1-p)M_2) \geq \frac{2}{8}\mathbb{P}(S_1|pM_1 + (1-p)M_2). \tag{3}$$

The advantage of this condition is that by including probabilities on both sides of the inequality, the partition function Z can be cancelled from the inequality entirely. Using the expressions for $\mathcal{E}_{\mathrm{NJ}}(S_1|pP_1 + (1-p)P_2)$ and $\mathcal{E}_{\mathrm{NJ}}(S_2|pP_1 + (1-p)P_2)$ from the proof of Theorem 1, taking the natural logarithm of the above expression, and rearranging yields the equivalent set of inequalities,

$$\begin{aligned} [\#\mathrm{bp}](S_2) - [\#\mathrm{bp}](S_1) - RT \ln 4 \leq C(2p-1)[\#\mathrm{bp}](S_1 \Delta S_2) \\ \leq [\#\mathrm{bp}](S_2) - [\#\mathrm{bp}](S_1) + RT \ln 4. \end{aligned} \tag{4}$$

In order to find an upper bound for the length of the crossover window, we must find how slowly the expression $C(2p-1)[\#\mathrm{bp}](S_1 \Delta S_2)$ 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 $2RT \ln 4$, and letting ℓ be the length of the crossover window, this is equivalent to:

$$2C[\#\mathrm{bp}](S_1 \Delta S_2)\ell \leq 2RT \ln 4,$$

which can be rearranged to finish the proof. □

Using the VcQrr3 structures W_2 and W_5 from the empirical results in Section 2, we can count $[\#\mathrm{bp}](W_5 \Delta W_2) = 6$. This implies that for $C > 0.5$, the binary $W_5 \to W_2$ interpolation crossover window has length at most 0.29. This is in contrast to the ideal width of 0.6, corresponding to structures W_2 and W_5 both being present with weight at least 20% for the range of p-values from $p = 0.2$ to 0.8. The crossover window is even narrower for other structures. For example, for the binary $W_5 \to W_3$ interpolation (where W_3 is described later), we can count $[\#\mathrm{bp}](W_3 \Delta W_5) = 24$, reducing the crossover window length to at most 0.072. We will find similarly short crossover windows for other prediction models.

3.2. SHAPE data analysis.

Reconstructing a multimodal structural distribution requires both identifying the structures in the distribution, and identifying their weighting. So far, we have found evidence that the structural prediction models may not accurately identify the structures nor the weightings corresponding to mixed sequences of binary SHAPE data. In this section, we consider a related problem: if the structures S_1 and S_2 are known in advance, is it possible to determine the weighting p of structure S_1 from a string of experimental SHAPE data? More precisely, in Rice et al. (2014), the authors propose a cutting-edge version of SHAPE data, described in Section 6.1.1 below. This method of incorporating SHAPE data improves predictions for those sequences that have structures which are hard to predict due to pseudoknots or other non-canonical features. We will consider empirical distributions based on this model: one for paired nucleotides, and one for unpaired nucleotides. Let M_1 and M_2 be SHAPE sequences for S_1 and S_2 modelled by these distributions. Assuming the distributions of SHAPE data are known, we investigate whether it is possible to determine p when given the values of the sequence D, where $D = pM_1 + (1-p)M_2$.

To start, consider the scenario where $[\#\mathrm{bp}](S_1 - S_2) = 0$. In this case, information about the proportion of S_1 and S_2 is contained mostly in positions where the structures differ, counted by $[\#\mathrm{bp}](S_2 - S_1)$. Let $\bar{D}$ be the average value of the SHAPE values in D corresponding to these positions. Let X and Y be random variables distributed as the SHAPE values for unpaired and paired nucleotides, respectively. Then, a method of moments estimator for p is:

$$\hat{p} := \frac{\bar{D} - \mathbb{E}Y}{\mathbb{E}X - \mathbb{E}Y}.$$

We investigate whether $\hat{p}$ reliably reconstructs p, and conclude that when S_1 and S_2 differ by a short-length helix, it does not. In particular, in Section 6.1.1, we find a formula for the variance of $\hat{p}$ in terms of $[\#\text{bp}](S_2 - S_1), p$, and the distributions for X and Y. It turns out that when S_1 and S_2 differ by a helix of length 4, $\hat{p}$ can have a standard deviation as high as 0.5 for $p \in [0.25, 0.75]$. A standard deviation this high means that the value of p cannot reliably be reconstructed. However, we will find potential improvements for SHAPE data that will lower the variance of $\hat{p}$, guaranteeing better predictions of p on average. Then, we analyze the best way to estimate p in the case where structures differ by more than one helix in Section 6.2.

4. Background

Here, we give more details on SHAPE data and SHAPE-directed energy and probability models. The specific models we will analyze are detailed in the section where the analysis occurs.

4.1. **Nearest Neighbor Thermodynamical Model.** The *Nearest Neighbor Thermodynamical Model* assigns an energy score to each potential structure into which a sequence can fold. The model has hundreds of parameters, corresponding to the energies of every possible structural feature and all types of bonds that can appear in an RNA secondary structure, as described in Mathews and Turner (2006), Turner and Mathews (2010), and Doshi et al. (2004). The total energy of a structure is the sum of the energies of all of its components. Because we will be analyzing multimodal distributions, we will analyze the suboptimal structures predicted by this model. Given a fixed RNA sequence, energies can be used to define a Boltzmann distribution of structures, as described in Equation (2). 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 $\mathbb{P}(S)$. But, when a distribution of structures is identified, then there are efficient algorithms which can sample structures from the distribution, as described in Ding and Lawrence (2003) and Mathews (2006), and used in the empirical study in Section 2.

4.2. **SHAPE data.** Auxiliary data can be used to shift the probabilities in the distributions by assigning additional energy terms to each structure, called *pseudoenergies.* The $C|x_i - M_i|$ terms in Equation (1) are an example of pseudoenergies. Experimentally, folded RNA sequences are exposed to chemical reagents in probing experiments, resulting in *SHAPE* data, named after the laboratory method, *selective 2'-hydroxyl acylation analyzed by primer extension.* There are several different reagents that are commonly used: N-methylisatoic anhydride (NMIA), 1-methyl-7-nitroisatoic anhydride (1M7), and 1-methyl-6-nitroisatoic anhydride (1M6). As discussed in Rice et al. (2014), different chemicals can give different structural signals, which we will revisit in Section 6.1.1.

Regardless of the chemical used, once this experiment is completed, each nucleotide in an RNA sequence is assigned a (typically) non-negative score reflecting its reactivity to the reagent. In general, the higher the reactivity, the less likely a nucleotide is to be paired. As mentioned above, there are several ways of incorporating this information into the NNTM, and the survey Eddy (2014) discusses benefits and drawbacks of each. The accuracy of single-conformation predictions is generally high when SHAPE data is included in predictions. However, for those sequences that have structures which are still hard to predict due to pseudoknots or other non-canonical features, the use of differential SHAPE data can help, as introduced in Rice et al. (2014).

Although there is evidence of the existence of multimodal structural distributions, there is not a large collection of data derived from known multimodal distributions yet. As a result, we will model

SHAPE data coming from a multimodal distribution with a linear interpolation of SHAPE data from two structures, as described in Section 2. Throughout the paper, we will use three different types of SHAPE distributions to model mixed data. First, in Section 2, we generated SHAPE data from empirically-derived distributions developed in Sükösd et al. (2013), to illustrate the potential problems with using SHAPE data in existing prediction models. These empirical distributions are described further below. Second, when analyzing the individual prediction models, we will use binary SHAPE sequences. These noiseless sequences represent an idealized version of SHAPE data, and simplify the analysis of each model. Finally, in Section 6 when we investigate how well data can be used to recover the weightings of structures in a distribution, we will use experimental SHAPE data from Rice et al. (2014).

Before moving on, we discuss the empirical distributions for SHAPE data described in Sükösd et al. (2013). Here, the authors fit distributions for paired and unpaired nucleotides to SHAPE data from two *Escherichia coli* ribosomal sequences: a 16S rRNA sequence with 1542 nucleotides, and a 23S rRNA sequence with 2904 nucleotides. The authors formed both a binary model where they fitted the data separately for paired and unpaired nucleotides, and a ternary model where they broke paired nucleotides into *center-paired* and *edge-paired* nucleotides. A paired nucleotide at position i is center-paired if the nucleotides at positions $i-1, i$, and $i+1$ are all paired, and are all part of the same helix. As an added technicality, nucleotides in the center of a helix are still considered center-paired when one side of a helix is interrupted by a single unpaired nucleotide. All other paired nucleotides are called edge-paired.

The ternary model performed better in simulations, and thus is what we used to generate SHAPE data in Section 2 above. However, the binary model will be used for incorporating SHAPE data into SCFGs. Thus, we will let $g(x)$ be the probability distribution function for the paired distribution, and $h(x)$ for the unpaired distribution. $g(x)$ is a generalized extreme value distribution with parameters $\xi = 0.895341, \sigma = 0.0712473$, and $\mu = 0.054239$. $h(x)$ is an exponential distribution with parameter $\lambda = \frac{1}{0.681211}$.

4.3. **Stochastic context-free grammars.** An alternative to NNTM energy models, stochastic context-free grammars (SCFGs) use recursive rules to convert a non-terminal, S, into a string of symbols called terminals. SCFGs were originally introduced in the context of speech recognition and language analysis, as in Baker (1979) and Booth and Thompson (1973). A couple decades later, SCFGs were applied to RNA secondary structures. They have been implemented and analyzed in Knudsen and Hein (1999), Knudsen and Hein (2003), Dowell and Eddy (2004), and Durbin et al. (1998).

For RNA secondary structures, the non-terminal S is converted into the terminals, which are dots '.' and brackets '(' and ')'. Such a sequence of parentheses and dots is the *dot-bracket notation* for an RNA secondary structure, where pairs of brackets correspond to paired nucleotides and dots correspond to unpaired nucleotides. For example, the word ((..).) corresponds to a structure where the first and last nucleotide are paired together, the second and fifth nucleotide are paired together, and the rest are unpaired. Dot-bracket notation was popularized by the ViennaRNA prediction software from Lorenz et al. (2011).

Once the SCFG converts the non-terminal S into a word of terminals, each terminal has a corresponding set of emission probabilities. In the context of RNA structures, emission probabilities correspond to the chance of the non-terminals becoming each type of nucleotide, A, C, G, or U. Typically, given an RNA sequence, the Inside and Outside algorithms from Baker (1979) can be used to compute the structure with the maximum probability of occurring, which is then the structure the SCFG predicts. The SCFG also produces a distribution of structures given an RNA sequence. In order to compute such a distribution, one must compute a partition function similar to in the NNTM: probabilities must be normalized by the sum of the probabilities of all possible structures given the particular RNA sequence.

Name	Dot-bracket Structure	Helices
W_0	((((((((.........))))))))......((((....))))......................................((((((((((......)))))))))).....	W_0
W_1	((((((((.........)))))))).(((((.(((....)))))))).............................((((((((((......)))))))))).....	$W_0 \cup H_1$
W_2	((((((((.........))))))))......((((....))))...((((((......))))))............((((((((((......)))))))))).....	$W_0 \cup H_2$
W_3	((((((((.........)))))))).(((((.(((....))))))))...(((((((...........))))))).((((((((((......)))))))))).....	$W_0 \cup$ $H_1 \cup H_3$
W_4	((((((((.........))))))))...(((.(((....)))))).((((((......))))))............((((((((((......)))))))))).....	$W_0 \cup$ $H_1 \cup H_2$
W_5	((((((((.........))))))))...(((((((....))))...((((((......))))))....))).....((((((((((......)))))))))).....	$W_0 \cup$ $H_2 \cup H_4$

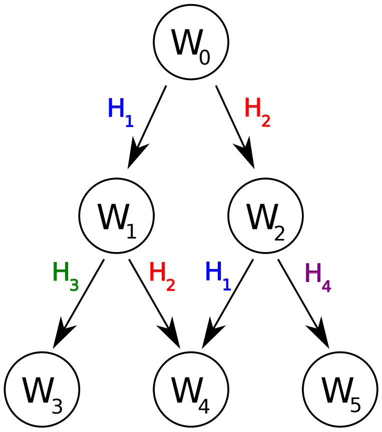

FIGURE 3. The VcQrr3 sequence can fold into several potential structures, all of which contain the same three helices from structure W_0 plus combinations of the four remaining colored helices. On the right, arrows represent adding a helix to a structure, illustrating the relationship between the W_i.

In our case, the SCFG must be modified so that it can also incorporate SHAPE data into its predictions. To include the SHAPE data, we use the empirical binary paired and unpaired distributions from Sükösd et al. (2013). With this information, for each position i, we can multiply the original emission probability by a new factor corresponding to the probability of the SHAPE value at position i being emitted given that the position is paired or unpaired. More details are given in Section 5.3.

4.4. **VcQrr3 structures.** In Rogers and Heitsch (2014), a new method of analyzing distributions of RNA structures was presented, where structures were classified by the helices that they contained, instead of by their exact pairings. By forming equivalence classes of structures in this manner, a landscape of many competing structures could be reduced to a handful of structural classes that combine structures containing similar functional information. The authors applied this analysis to the VcQrr3 sequence discussed in Section 2. VcQrr3 is an RNA sequence with 107 nucleotides that was studied in Lenz et al. (2004) and Tu and Bassler (2007). Such short RNA sequences have gained attention in recent research because their biological roles are more diverse and complicated than originally anticipated, as described in Couzin (2002) and Doudna (2000).

For the RNA sequence VcQrr3, the authors identified seven helices that formed the building blocks for many potential VcQrr3 structures. Representatives from each structural class are summarized in dot-bracket notation in Figure 3. Without SHAPE, a Boltzmann sample of 1000 structures was produced by the NNTM thermodynamic optimization. In the sample, structures W_1 through W_5 all occurred with frequency at least 30. These structures share the helices found in structure W_0, along with combinations of a set of four other helices. The representatives shown in Figure 3 each occurred with positive frequency in a SHAPE-directed Boltzmann sample. Although many of the results below are in terms of these particular representatives, the results rely on the number of pairings and underlying energy of each structure, which does not change greatly between different representatives in the same structural class.

5. MODEL ANALYSIS: BINARY SHAPE DATA

In this section, we analyze the crossover points and crossover windows for binary interpolations under the Deigan energy method, the Zarringhalam energy method, and a popular SCFG, the KH99 grammar from Knudsen and Hein (1999). Throughout this section, we generalize the binary data so that the value of the unpaired data is a constant, A, instead of 1. The results below will resemble the simpler results for the Nussinov-Jacobson energy model.

5.1. **The Deigan model.** The Deigan method, introduced in Deigan et al. (2009), assigns pseudoenergies of the following form:

$$\Delta G_{\text{SHAPE}}(i) := m \ln(1 + a_i) + b. \tag{5}$$

Here, m is a positive constant, b a negative constant, and a_i is the SHAPE reactivity value corresponding to nucleotide i. The penalty $\Delta G_{\text{SHAPE}}(i)$ is used only if the ith nucleotide is paired. This penalty can be incorporated into the dynamic programming algorithm that computes the NNTM energy for a structure S. In Deigan et al. (2009), the authors showed that shifting the probabilities defined in Equation (2) by incorporating these pseudoenergies can dramatically improve structural predictions.

Here, we investigate how the Deigan pseudoenergies from Equation (5) behave under binary SHAPE interpolations. Due to the way the Deigan penalty is incorporated into dynamic programming algorithms, $\Delta G_{\text{SHAPE}}(i)$ is applied once for edge-paired nucleotides, twice for center-paired nucleotides, and never for unpaired nucleotides. Thus, we will need to count slightly different quantities than in the Nussinov-Jacobson model. Let $\mathcal{E}(S)$ be the NNTM energy for $\mathcal{S}$ before SHAPE data is incorporated. Let $[\#\text{m}](S)$ be the number of nucleotides in edge-pairs in S, plus twice the number of nucleotides in center-pairs. We will refer to this as counting nucleotides *with multiplicity.* Additionally, let $[\#\text{m}](S_1 - S_2)$ be the number of paired nucleotides in S_1, counted with multiplicity, that are not in S_2. In this computation, edge- or center-pairedness depends only on S_1, and not on S_2.

This method of counting paired nucleotides can lead to counterintuitive results when many edge-paired nucleotides from one structure align with center-paired nucleotides from another. Consider the structures below:

$$T_1:\ \ \texttt{........((((((....))))))........}$$
$$T_2:\ \ \texttt{.((..((..))))......((((..))..))..}$$

Here, $[\#\text{m}](T_1) = 20$ due to the high number of center-paired nucleotides, and $[\#\text{m}](T_2) = 16$. However, even though T_1 has more paired nucleotides with multiplicity, $[\#\text{m}](T_1 - T_2) = 4$ while $[\#\text{m}](T_2 - T_1) = 8$. Although this combinatorial example illustrates the scenario, VcQrr3 structures W_5 and W_4 have a similar relationship. For structures with this sort of relationship (and some other conditions), we will show that the crossover point must be skewed towards the structure with fewer pairings overall. Thus, we have the following version of Theorem 1 for the Deigan model:

Theorem 3. *Let S_1 and S_2 be two structures for the same RNA sequence. Assume that $\mathcal{E}(S_1) \geq \mathcal{E}(S_2)$, $[\#m](S_1) \leq [\#m](S_2)$, and $[\#m](S_1 - S_2) = D[\#m](S_2 - S_1) > 0$ for some constant $D > 1$. Then, there is no choice of $A > 0, m > 0$, and $b < 0$ such that the crossover point for the binary $S_2 \to S_1$ interpolation is in the range, $[0, D/(D + 1)]$.*

Before proving Theorem 3, we investigate its application to VcQrr3. For structures W_5 and W_4, we can count that $[\#\text{m}](W_5 - W_4) = 5$ while $[\#\text{m}](W_4 - W_5) = 3$, that $[\#\text{m}](W_5) = [\#\text{m}](W_4) = 104$, and verify that $\mathcal{E}(W_5) > \mathcal{E}(W_4)$. Then, in the statement of the theorem, $D = 5/3$, so that $D/(D+1) = 0.625$. This yields the following corollary, showing the existence of uncentered crossover points in RNA structures for a real RNA sequence:

Corollary 4. *There are no values for m, b, and A with $m, A > 0$ such that the crossover point p in the $W_4 \to W_5$ interpolation has $p \leq 0.625$.*

Theorem 3. Let $\mathcal{E}_D(S|M)$ be the total Deigan model energy for the structure S with SHAPE sequence M. Although the dynamic programming algorithm recursively adds the penalties $\Delta G_{\text{SHAPE}}(i)$ to each paired nucleotide while simultaneously computing the free energy of the structure using the

NNTM, the SHAPE pseudoenergy can be computed separately and added to the NNTM energy to calculate $\mathcal{E}_D(S|M)$.

We consider the possible contributions of $\Delta G_{\text{SHAPE}}(i)$ for mixtures of binary SHAPE data. Every paired nucleotide in a structure contributes b to the structure's Deigan SHAPE-directed energy, counted with multiplicity, regardless of the SHAPE sequence. Additionally, in $\mathcal{E}_D(S_1|pP_1 + (1-p)P_2)$, a term of the form $m\ln[1+A(1-p)]$ appears for each paired nucleotide in S_1 that is unpaired in S_2, counted with multiplicity. Similarly, in the energy for S_2, terms of the form $m\ln[1+Ap]$ correspond to paired nucleotides in S_2 that are unpaired in S_1.

Thus, the crossover point for the $S_2 \to S_1$ interpolation satisfies the equation $\mathcal{E}_D(S_1|pP_1 + (1-p)P_2) = \mathcal{E}_D(S_2|pP_1 + (1-p)P_2)$, which can be rearranged to the following expression:

$$\begin{aligned}(6)\quad &\frac{\mathcal{E}(S_1) - \mathcal{E}(S_2) + \big([\#\text{m}](S_1) - [\#\text{m}](S_2)\big)b}{m} \\ &\qquad = [\#\text{m}](S_2 - S_1)\ln(1+Ap) - [\#\text{m}](S_1 - S_2)\ln(1+A(1-p)).\end{aligned}$$

Our goal will be to show that for all $A, m > 0, D > 1, b < 0$, and $p \in [0, D/(D+1)]$, the right side of the expression is negative, while the left side is positive, which will yield a contradiction. The left side is simple to analyze: the numerator is positive because $\mathcal{E}(S_1) > \mathcal{E}(S_2)$ and $[\#\text{m}](S_1) \geq [\#\text{m}](S_2)$ by assumption. And, the denominator is positive because $m > 0$.

The right side of Equation (6) is more challenging to analyze. We see that when $A = 0$, the right side of the equation is zero. Then, we will show that with the conditions on A, D, and p, the derivative is negative with respect to A, which will prove that the expression is negative. The derivative of the right side of Equation (6) is:

$$\begin{aligned}(7)\quad &\frac{\mathrm{d}}{\mathrm{d}A}\Bigg([\#\text{m}](S_2 - S_1)\ln(1+Ap) - [\#\text{m}](S_1 - S_2)\ln(1+A(1-p))\Bigg) = \\ &\qquad \frac{1}{(1+Ap)(1+A(1-p))} \cdot \Bigg(A\big([\#\text{m}](S_1 - S_2) - [\#\text{m}](S_2 - S_1)\big)(p^2 - p) \\ &\qquad\qquad + \big([\#\text{m}](S_1 \Delta S_2)\big)p - [\#\text{m}](S_1 - S_2)\Bigg).\end{aligned}$$

The denominator of the fractional term is positive for all $A > 0$ and all $p \in (0,1)$, which leads us to investigate the sign of the remaining expression enclosed in parentheses. We substitute the assumed condition $[\#\text{m}](S_1 - S_2) = D[\#\text{m}](S_2 - S_1)$ into this expression, and it becomes:

$$(8)\qquad [\#\text{m}](S_2 - S_1) \cdot \big[(AD - A)p^2 + (-AD + A + D + 1)p - D\big].$$

We view the expression in brackets as a polynomial in p. When $p = 0$, this expression is negative, because $D > 1$ by assumption. Additionally, the coefficient of p^2 is positive while the constant term, $-D$, is negative, which means that the derivative in Equation (7) has a zero for a negative p value and a positive p value. We will show that the the positive p root is greater than $D/(D+1)$, which will imply that the derivative is always negative for $p \in [0, D/(D+1)]$. In order to do so, we substitute $p = D/(D+1) + t$ for an indeterminate, t, into Equation (8), so that it becomes:

$$\begin{aligned}(9)\quad &\frac{[\#\text{m}](S_2 - S_1)}{(D+1)^2} \cdot \Bigg[\big(A(D-1)(D+1)^2\big)t^2 + \\ &\qquad \big((D+1)(AD^2 - 2AD + D^2 + A + 2D + 1)\big)t - AD^2 + AD\Bigg].\end{aligned}$$

Here, the zeroes are determined by the expression in brackets, which we view as a polynomial in t. Again, the coefficient of t^2 is positive, because $A > 0$ and $D > 1$. Also, the constant term is negative, since $D > 1$. Thus, there is a root of this expression when $t > 0$. Because this corresponds to the positive root of $p = D/(D+1) + t$, this proves that the positive root of p is always greater than $D/(D+1)$. Therefore, the derivative in Equation (7) is negative for all $p \in [0, D/(D+1)]$. Note that this condition on the negativity of the derivative is sharp. To see this, we can write the zeroes Equation (7) in terms of p explicitly. Taking the limit as A approaches zero gives $D/(D+1)$.

Now, since the right hand side of Equation (6) is zero for $A = 0$ and has a negative derivative for $A > 0$ and $p \in [0, D/(D+1)]$, this implies that the expression is negative for $A > 0$ and $p \in [0, D/(D+1)]$. This completes the proof. □

The previous result showed a problem with centering crossover points for a specific class of pairs of RNA structures. The next result instead shows that crossover points can be centered whenever one structure contains all of the helices of the other, plus more. In fact, in this scenario, crossover points can be pushed to any interval $[c_1, c_2]$. This highlights that for the Deigan model, the reason that crossover points cannot consistently be centered involves structures with competing helices.

Theorem 5. *Let $\{S_{i,1}, S_{i,2}\}_{i=1}^{k}$ correspond to pairs of structures for the same RNA sequence where $[\#m](S_{i,1} - S_{i,2}) = 0$ for all i. Then, for any constants $0 < c1 < c2 < 1$, and any $A > 0$, there exist infinitely many Deigan parameters $m > 0$ and $b < 0$ so that all k crossover points for the binary interpolations are between c_1 and c_2.*

Proof. The proof relies on the fact that when two structures differ by exactly one helix, then the difference of their energies has a Deigan term of the form $m \ln(1 + Ap)$ occur as many times as the term, $-b$. This special relationship will imply that for any two structures $S_{i,1}$ and $S_{i,2}$ that differ by one helix, the set of m and b forcing the crossover point to be between c_1 and c_2 is a cone in the (m, b)-plane, where the upper and lower slopes of the lines depend only on c_1, c_2, and A, but not on the structures themselves. The intersection of finitely many such cones is always another infinite cone, which will complete the proof. Now, for the details.

In the case that $S_{i,2}$ has strictly more helices than $S_{i,2}$, we have that $[\#\text{m}](S_{i,2} - S_{i,1}) = [\#\text{m}](S_{i,2}) - [\#\text{m}](S_{i,1})$ and $[\#\text{m}](S_{i,1} - S_{i,2}) = 0$. Thus, setting the interpolated Deigan energies of $S_{i,1}$ and $S_{i,2}$ equal yields:

$$\mathcal{E}(S_{i,1}) + [\#\text{m}](S_{i,1})b = \\ \mathcal{E}(S_{i,2}) + [\#\text{m}](S_{i,2})b + \Big([\#\text{m}](S_{i,2}) - [\#\text{m}](S_{i,1})\Big) m \ln(1 + Ap).$$

Rearranging yields:

$$-b = m \ln(1 + Ap) + \frac{\mathcal{E}(S_{i,2}) - \mathcal{E}(S_{i,1})}{[\#\text{m}](S_{i,1}) - [\#\text{m}](S_{i,2})}.$$

Let $L_i = \dfrac{\mathcal{E}(S_{i,2}) - \mathcal{E}(S_{i,1})}{[\#\text{m}](S_{i,1}) - [\#\text{m}](S_{i,2})}$. The bound $c_1 \le p \le c_2$ is equivalent to the following:

$$m \ln(1 + c_1 A) + L_i \le -b \le m \ln(1 + c_2 A) + L_i.$$

As a result, for any particular $A > 0$, the set of solutions satisfying the condition that $c_1 < p < c_2$ is a cone in the (m, b)-plane with a point at the intercept of the two lines, $(m, b) = (0, L_i)$. Note that while the intercept L_i of the two bounding lines depends on the structures themselves, the slope depends only on c_1, c_2, and A. As a result, the upper lines of all k cones are parallel, and the lower lines of all k cones are parallel. So, the intersection of all of the cones is the space between the lowest cone's upper bound, and the highest cone's lower bound, which must be another infinite cone of solutions, completing the proof. □

Next, we examine the crossover windows for the Deigan interpolations. As in the Nussinov-Jacobson model, the simplest scenario is when we have structures where $[\#\text{m}](S_2 - S_1) = 0$. Below, the Deigan analogue of Theorem 2 will imply that for two of the VcQrr3 structures, the crossover window is of width at most 14%, which is much shorter than the expected width of 60%.

Theorem 6. *Let S_1 and S_2 be two potential structures for the same RNA sequence, where $[\#m](S_2 - S_1) = 0$. Then, for all values of $A, m,$ and b with $A > 1$ and $m > 1$, the crossover window for the $S_2 \to S_1$ binary interpolation has length at most $2 - 2 \cdot 4^{-2RT/[\#m](S_1 - S_2)}$.*

Proof. We use the condition in Equation (3) from the proof of Theorem 2. Plugging in Deigan pseudoenergies yields the following equivalent condition, which is the analogue of Equation (4):

$$\begin{aligned}(10)\quad [\mathcal{E}(S_2) - \mathcal{E}(S_1)] + \Big([\#\mathrm{m}](S_2) - [\#\mathrm{m}](S_1)\Big)b - RT\ln 4 \\ \le m[\#\mathrm{m}](S_1 - S_2)\ln(1 + A(1-p)) \le \\ [\mathcal{E}(S_2) - \mathcal{E}(S_1)] + \Big([\#\mathrm{m}](S_2) - [\#\mathrm{m}](S_1)\Big)b + RT\ln 4.\end{aligned}$$

Now, in order to find an upper bound for the length of the crossover window, we must find the how slowly the expression $m[\#\mathrm{m}](S_1 - S_2)\ln(1 + A\tilde{p})$ can jump between the lower bound and upper bound in (10). In other words, we investigate how slowly $m[\#\mathrm{m}](S_1 - S_2)\ln(1 + A\tilde{p})$ can increase by $2RT\ln 4$, as a function of $\tilde{p}$.

Note that the derivative of $m[\#\mathrm{m}](S_1 - S_2)\ln(1 + A\tilde{p})$ with respect to $\tilde{p}$ is decreasing in $\tilde{p}$, which implies that the longest crossover window occurs when one of the endpoints in the window is $\tilde{p} = 1$. Thus, we assume that the crossover window is of the form $\tilde{p} \in [p_L, 1]$, and we search for the lower bound of the window, p_L, where the following is true:

$$m[\#\mathrm{m}](S_1 - S_2)\ln(1 + A) - m[\#\mathrm{m}](S_1 - S_2)\ln(1 + Ap_L) = 2RT\ln 4.$$

Solving for p_L yields:

$$\begin{aligned}p_L &= 4^{-\frac{2RT}{m[\#\mathrm{m}](S_1 - S_2)}} + \frac{1}{A}\left[4^{-\frac{2RT}{m[\#\mathrm{m}](S_1 - S_2)}} - 1\right] \\ &\ge 2 \cdot 4^{-\frac{2RT}{m[\#\mathrm{m}](S_1 - S_2)}} - 1,\end{aligned}$$

where the last line is true because $A \ge 1$ by assumption. This implies that the width of the crossover window is at most:

$$1 - p_L \le 2 - 2 \cdot 4^{-\frac{2RT}{m[\#\mathrm{m}](S_1 - S_2)}}.$$

This is a decreasing function in m, and thus is at most $2 - 2 \cdot 4^{-2RT/[\#\mathrm{m}](S_1 - S_2)}$. □

We can apply this result to the VcQrr3 structures. We have that $[\#\mathrm{m}](W_1 - W_3) = 0$ and $[\#\mathrm{m}](W_3 - W_1) = 24$, which yields:

Corollary 7. *For all values of $A, m,$ and b with $A, m > 1$ and the temperature $T = 310K$, the crossover window for the $\{W_3, W_1\}$ interpolation has length at most* 0.14.

Thus, for the Deigan method, have found an upper bound on the crossover window length for a class of RNA structures, and identified an actual RNA sequence with structures within this class. Our analysis has focused on the Deigan method because it is the dominant approach to identifying secondary structures. However, we will now turn towards finding similar results for the other structural prediction models, illustrating that these challenges are not specific to the Deigan model. Because the other models are more complicated, the results will be more constrained. Nonetheless, we will still be able to identify structures with uncentered crossover points and short crossover windows.

5.2. **The Zarringhalam model.** Introduced in Zarringhalam et al. (2012), the Zarringhalam model for incorporating SHAPE data uses a pseudoenergy term, much like the Deigan model. However, unlike the Deigan model, a penalty is added exactly once for each nucleotide, regardless of whether the nucleotide is center-paired, edge-paired, or unpaired. Thus, we will be interested in the [#bp] function from the analysis of the Nussinov-Jacobson model. As opposed to the Deigan model, the Zarringhalam model uses all of the SHAPE data to assign energies to a structure.

Explicitly, a new pseudoenergy penalty of the form,

$$\Delta G_{\text{Z-SHAPE}}(i) = \beta |x_i - f(M_i)|,$$

is added to each term, where x_i is 1 if the ith nucleotide in the structure is unpaired, and x_i is 0 if the ith nucleotide is paired. β is a parameter with the default value, 0.89. As before, M_i is the SHAPE reactivity corresponding to the ith nucleotide. The function f rescales SHAPE data in a piece-wise linear fashion:

$$f(y) = \begin{cases} 0, & y \leq 0, \\ \dfrac{7}{5}y, & 0 < y \leq 0.25, \\ 4y - \dfrac{13}{20}, & 0.25 < y \leq 0.4, \\ \frac{3}{4}y + \frac{13}{40}, & 0.4 < y \leq 0.7, \\ \dfrac{y}{10} + \dfrac{39}{50}, & 0.7 < y \leq 2.2, \\ 1, & y \geq 2.2. \end{cases}$$

Thresholds for f were determined by examining distributions of SHAPE data and determining which ranges of values corresponded roughly to nucleotides being highly unreactive, slightly unreactive, slightly reactive, and highly reactive, and were partly based on data from Deigan et al. (2009). In particular, f linearly maps the intervals $[0, 0.25]$, $[0.25, 0.3]$, $[0.3, 0.7]$, and $[0.7, 2.2]$ to the intervals, $[0, 0.35]$, $[0.35, 0.55]$, $[0.55, 0.85]$, and $[0.85, 1]$.

Despite the fact that SHAPE penalties are assessed for both paired and unpaired nucleotides, it turns out that the Zarringhalam model also does not guarantee crossover points between 0.4 and 0.6, even for structures with $[\#\text{bp}](S_1 - S_2) = 0$, with the default value $\beta = 0.89$. However, the crossover point result for the Zarringhalam model below is less general than Theorem 6 for the Deigan model because the Zarringhalam model is more complicated. This is due to the fact that penalties are now applied to paired and unpaired nucleotides, and due to the piecewise function, f.

Theorem 8. *When $\beta = 0.89$ in the Zarringhalam model, there is no value of A such that both the $W_5 \to W_2$ and the $W_0 \to W_2$ binary interpolations have crossovers between 0.4 and 0.6.*

Proof. Again, the total contribution of the $\Delta G_{\text{Z-SHAPE}}$ terms can be tabulated separately from the rest of the NNTM energy calculations. Because the Zarringhalam pseudoenergy is assessed on both paired and unpaired nucleotides, the Zarringhalam pseudoenergies for S_j in a $S_k \to S_j$ interpolation can take on four different forms, depending on whether the ith nucleotide is paired or unpaired in each of S_j and S_k. Let c_1 and c_2 correspond to the crossover points in the $W_5 \to W_2$ and $W_0 \to W_2$ interpolations, respectively. By tabulating the paired and unpaired nucleotides in each structure and how they overlap, we find the equations,

$$\begin{aligned} &-33.6 + 6\beta(1 - f((1 - c_1)A)) + 45\beta(1 - f(A)) \\ &\qquad\qquad\qquad\qquad = -31.6 + 6\beta f((1 - c_1)A) + 45\beta(1 - f(A)), \end{aligned}$$

$$-33.6 + 12\beta f(c_2 A) + 51\beta(1 - f(A)) = -27.4 + 12\beta f(c_2 A) + 51(1 - f(A)).$$

To simplify matters, let $\tilde{c}_1 = 1 - c_1$. Then, the equations above simplify to the following, when considering that $\beta = 0.89$:

$$f(A\tilde{c}_1) = \frac{1}{2} - \frac{1}{6\beta} = \frac{167}{534} \approx 0.31, \tag{11}$$

$$f(Ac_2) = \frac{1}{2} + \frac{62}{240\beta} = \frac{211}{267} \approx 0.79. \tag{12}$$

From equation (11), we know that because the value of $f(A\tilde{c}_1)$ is between 0 and 0.35, $f(A\tilde{c}_1) = 1.4A\tilde{c}_1$, and hence

$$A = \frac{835}{3738\tilde{c}_1}. \tag{13}$$

Likewise, from equation (12), we have that $f(Ac_2) = \frac{3}{4}Ac_2 + \frac{13}{40}$, yielding:

$$A = \frac{4969}{8010c_2}. \tag{14}$$

The fact that the crossover points are restricted to $[0.4, 0.6]$ can be combined with Equations (13) and (14) to show that $0.372 \leq A \leq 0.559$ and $1.03 \leq A \leq 1.56$, respectively. These contradictory intervals complete the proof. □

Next, for the Zarringhalam model, we find the analogue of Theorem 2. Here, we add the restriction that $A \leq 2.2$, corresponding to the maximum SHAPE value anticipated by f, the piecewise function that rescales SHAPE data.

Theorem 9. *Consider any structures S_1 and S_2 for the same RNA sequence, with $[\#bp](S_2 - S_1) = 0$. Then, for $A \leq 2.2$ and $\beta > 0$, the Zarringhalam crossover window has length at most $10RT\ln 4/(\beta A[\#bp](S_1 - S_2))$.*

Proof. The proof has the same structure as Theorem 2. Let c_1 be the number of nucleotides which are paired in both S_1 and S_2, and let c_2 be the number of nucleotides which are unpaired in S_1 and S_2. Then, the Zarringhalam energy for S_1 given data $pP_1 + (1-p)P_2$ is:

$$\mathcal{E}_Z(S_1|pP_1 + (1-p)P_2) = \mathcal{E}(S_1) + \beta\Big[c_1 + [\#\text{bp}](S_1 - S_2)f(\tilde{p}A) + c_2(1 - f(A))\Big],$$

where $\tilde{p} = 1 - p$, and a similar equation holds for S_2. Again using the condition in Equation (3) and simplifying, we obtain the equivalent condition,

$$\mathcal{E}(S_1) - \mathcal{E}(S_2) - RT\ln 4 \leq \beta[\#\text{bp}](S_1 - S_2)\big(1 - 2f(\tilde{p}A)\big) \leq \mathcal{E}(S_1) - \mathcal{E}(S_2) + RT\ln 4.$$

Thus, we investigate how slowly $\beta[\#\text{bp}](S_1 - S_2)(1 - 2f(\tilde{p}A))$ can increase, as a function of p. We have:

$$\frac{d}{dp}\left[1 - 2f((1-p)A)\right] = 2Af'((1-p)A) \geq 0.2A,$$

by looking at the piecewise function, f. Therefore, if ℓ is the length of the crossover window,

$$0.2[\#\text{bp}](S_1 - S_2)\beta\ell A \leq 2RT\ln 4,$$

and rearranging completes the proof. □

Noticing that $[\#\text{bp}](W_3 - W_0) = 24$ and $[\#\text{bp}](W_0 - W_3) = 0$ gives the following corollary for $A = 2.2$, the theoretical maximum SHAPE value imposed by the rescaling function, f:

Corollary 10. *The crossover window for the Zarringhalam $W_3 \to W_0$ interpolation with $\beta = 0.89$ and $A = 2.2$ has width at most* 0.2.

Thus, we have found that crossover points can be uncentered and crossover windows can be short for the Zarringhalam energy model as well. Next, we will consider stochastic context free grammars, which are an alternative to energy models. Because the SCFG will depend on incorporating empirical distributions for SHAPE data, the analysis becomes complicated and relies on numerical computations. However, we will find that even with SCFGs, we cannot guarantee that crossover points are centered.

5.3. **Stochastic context-free grammars.** In this section, we consider the SCFG introduced in Knudsen and Hein (1999), called the KH99 grammar. This grammar has been implemented in the Pfold program, Knudsen and Hein (2003). The grammar is described by the following rules (where w.p. stands for "with probability"):

$$\begin{array}{ccccc} S & \rightarrow & LS & \text{w.p.} & p_1 \\ & & L & \text{w.p.} & q_1 \\ \hline F & \rightarrow & (F) & \text{w.p.} & p_2 \\ & & LS & \text{w.p.} & q_2 \\ \hline L & \rightarrow & . & \text{w.p.} & p_3 \\ & & (F) & \text{w.p.} & q_3 \end{array}$$

Beginning with the non-terminal S, the rules are repeatedly applied until only dots and brackets (the terminals) remain. Then, each terminal has a corresponding set of emission probabilities, corresponding to the chance of the non-terminals becoming each type of nucleotide: A, C, G, or U. Thus, we will have rules as follows:

$$. \rightarrow \begin{array}{ccc} A & \text{w.p.} & p_A \\ C & \text{w.p.} & p_C \\ G & \text{w.p.} & p_G \\ U & \text{w.p.} & p_U \end{array} \Bigg| \quad () \quad \rightarrow \begin{array}{ccc} AU \text{ or } UA & \text{w.p.} & p_{AU} \\ GC \text{ or } CG & \text{w.p.} & p_{GC} \\ GU \text{ or } UG & \text{w.p.} & p_{GU} \end{array}$$

Above, () represents two paired nucleotides, which are treated simultaneously when emission probabilities are considered. AU and UA pairs are emitted with equal probability, and likewise for the other two pairs. In Knudsen and Hein (1999), optimal values for the emission probabilities were computed by training the grammar on a test set of known structures.

As mentioned in Section 4.3, to include the SHAPE data, we use the empirical binary paired and unpaired distributions from Sükösd et al. (2013) to modify the original emission probabilities. In our implementation of the SCFG, the distributions are discretized. As before, let $g(x)$ be the probability distribution function for the paired distribution, and let $h(x)$ be the probability distribution for the unpaired distribution. Then, if the SHAPE sequence is $\{M_i\}_{i=1}^n$, the emission probability for nucleotide j has an additional factor of $g(M_j)$ if position j is paired (a '(' or a ')'), and an additional factor of $h(M_j)$ if position j is unpaired (a '.').

Now, we are ready to analyze the crossovers for VcQrr3. We see that for the SCFG, it is impossible for the crossover points to be centered while simultaneously allowing the SHAPE signal to be strong enough to guide the predictions towards the correct structures, even for distance-one structures. The SCFG version of Theorem 1 is as follows:

Theorem 11. *Consider the distributions of secondary structures generated by the KH99 grammar with the standard parameters and empirical distributions for paired and unpaired SHAPE data. There is no value of A for which the crossover point p for the $W_4 \rightarrow W_2$ binary interpolation is between* 0.4 *and* 0.6 *and the probability of structures W_2 and W_4 are above the probability for structure W_5 at the crossover point.*

Proof. Because the empirically-derived paired distribution is a generalized extreme value distribution with a complicated probability density function, we will have to rely on numerical computations in order to prove this result. First, we employ the Cocke-Younger-Kasami algorithm to trace how the KH99 grammar can produce each of the structures, W_2, W_4, and W_5. Upon such a traceback, we have that the following transition probabilities T_2, T_4, and T_5 for W_2, W_4, and W_5, respectively:

$$T_2 := (1/2)^{28} p_1^{46} (1-p_1)^5 p_2^{24} (1-p_2)^4 p_3^{51} (1-p_3)^4 p_A^{12}\, p_U^{25}\, p_G^5\, p_C^9\, p_{AU}^{13}\, p_{GC}^{13}\, p_{GU}^2$$
$$T_4 := (1/2)^{30} p_1^{41} (1-p_1)^6 p_2^{25} (1-p_2)^5 p_3^{47} (1-p_3)^5 p_A^{10}\, p_U^{22}\, p_G^5\, p_C^{10}\, p_{AU}^{15}\, p_{GC}^{12}\, p_{GU}^3$$
$$T_5 := (1/2)^{31} p_1^{39} (1-p_1)^6 p_2^{26} (1-p_2)^5 p_3^{45} (1-p_3)^5 p_A^{11}\, p_U^{23}\, p_G^3\, p_C^8\, p_{AU}^{14}\, p_{GC}^{14}\, p_{GU}^3$$

Then, we also need to keep track of the product of the emission probabilities for each structure. Much like in the Zarringhalam model, the emission probabilities for a structure W_j with a SHAPE sequence that is a mixture of the sequences for W_j and W_k come in four different varieties, depending on whether the nucleotides in a given position i are paired or unpaired in each of the structures. Now, we can compute the total probabilities for each structure in the SCFG distribution:

$$\mathbb{P}(W_2|pP_2 + (1-p)P_4) = T_2 g(0)^{55} g(A(1-p)) h(Ap)^5 h(A)^{46},$$
$$\mathbb{P}(W_4|pP_2 + (1-p)P_4) = T_4 g(0)^{55} h(A(1-p)) g(Ap)^5 h(A)^{46},$$
$$\mathbb{P}(W_5|pP_2 + (1-p)P_4) = T_5 g(0)^{55} g(A(1-p)) g(Ap)^3 h(Ap)^2 g(A)^3 h(A)^{43}.$$

First, we aim to show that if $A \leq 0.2$, then for all p, $\mathbb{P}(W_5) > \mathbb{P}(W_4)$. On the range, $[0, 0.2]$, the ratio $g(x)/h(x)$ is strictly in the interval $[1, 5]$. As a result, we have for all $p \in [0,1]$ and all $A \in [0, 0.2]$:

$$\frac{g(A(1-p)) g(A)^3 h(Ap)^2}{h(A(1-p)) h(A)^3 g(Ap)^2} > \left(\frac{1}{5}\right)^2 \geq 0.031 > \frac{T_4}{T_5}$$

Rearranging this inequality and comparing to the probabilities for W_2, W_4, and W_5 illustrates that the probability assigned to W_5 is always greater than the probability assigned to W_4 for $A \leq 0.2$.

Next, equating the probabilities for structures W_2 and W_4 when $A = 0.2$ yields a crossover point of $p \approx 0.289$. As A increases, structure W_2 is favored more because it has strictly more unpaired nucleotides. This means that the crossover point becomes smaller as A increases, which can be verified numerically for the empirical distributions. Thus, the crossover point is at most 0.289 for $A \geq 0.2$. □

We now have evidence that both SCFGs and energy models cannot consistently center crossover points. This indicates that these models will have trouble reconstructing what structures appear in a multimodal distribution, along with the correct weightings of the structures. We next turn to a related problem, and investigate whether the correct weightings can be recovered in a multimodal distribution, if the structures are known in advance.

6. Parameter analysis: estimating p

In this section, we again consider an interpolation of SHAPE data coming from two structures, S_1 and S_2, with proportion p coming from structure S_1. However, we now investigate whether or not it is possible to determine the proportion, p, if the structures S_1 and S_2 are known in advance. In contrast to the previous sections, we consider experimental SHAPE data in this section. The goal is to determine whether or not the experimental SHAPE data distinguishes between paired and unpaired nucleotides enough to recover the weights of structures in a distribution.

6.1. **Interpolations with $[\#\mathbf{bp}](S_1 - S_2) = 0$.** Like when analyzing the models before, let us restrict our attention to the simple scenario where $[\#\text{bp}](S_1 - S_2) = 0$. The SHAPE values corresponding to nucleotides where the two structures agree give limited information about the proportion, p. So, we do not consider these SHAPE values, and instead only look at the SHAPE values corresponding to the positions where S_1 and S_2 differ structurally. Then, in the mixed SHAPE signal, we focus our attention on a set of n data points for the SHAPE values, $\{D_i\}_{i=1}^n$, where each point $D_i = pX_i + (1-p)Y_i$, with X_i being the SHAPE value corresponding to the unpaired nucleotide i from S_1, and Y_i being the SHAPE value corresponding to nucleotide i from S_2. With the set of points $\{D_i\}_{i=1}^n$, we would like to recover the value of p.

Due to its flexibility in analyzing different types of distributions, we analyze the estimator $\hat{p}$ generated by the method of moments:

$$\hat{p} := \frac{\bar{D} - \mathbb{E}Y}{\mathbb{E}X - \mathbb{E}Y}.$$

Here, $\bar{D}$ represents the average value of the data points D_i. X is a random variable with the same distribution as an unpaired SHAPE value, and Y is a random variable corresponding to paired SHAPE values. Since $\mathbb{E}\bar{D} = p\mathbb{E}X + (1-p)\mathbb{E}Y$, we have that $\mathbb{E}\hat{p} = p$, and thus $\hat{p}$ is an unbiased estimator of p.

In order to obtain reliable predictions of p using $\hat{p}$, we will need $\hat{p}$ to have a small variance. As a first check, we examine how the estimator $\hat{p}$ performs using the paired and unpaired distributions, $g(x)$ and $h(x)$, originally derived in Sükösd et al. (2013) and used in Section 5.3 from above. The paired distribution, which is a generalized extreme value distribution, has an infinite variance for $\xi = 0.895341 > 0.5$, and thus, the variance of $\hat{p}$ is unbounded. Additionally, randomly generating data from these distributions and testing the accuracy of $\hat{p}$ against this data reveals that $\hat{p}$ performs poorly, and is unable to reconstruct the value of p. Thus, we will consider what happens with other models of SHAPE data where the variance is finite. In this case, we can compute:

$$\operatorname{Var}\hat{p} \quad = \quad \frac{p^2 \operatorname{Var} X + (1-p)^2 \operatorname{Var} Y}{n(\mathbb{E}X - \mathbb{E}Y)^2}. \tag{15}$$

Ideally, the variance will be small enough to give consistent estimates for p. However, the variance will depend on underlying distributions for paired and unpaired nucleotides.

We will analyze the success of an estimator given distributions for X (unpaired SHAPE) and Y (paired SHAPE) by first choosing a minimal scenario where the estimator should perform well, and then by choosing a cutoff for the variance. We choose helices of length 4 as a minimal difference for structures S_1 and S_2, so that we will have $n = 8$ data points. Then, we will call the distributions for X and Y *separable* if the variance of $\hat{p}$ is at most 0.0225 for all p in $[0.25, 0.75]$ and for all $n \geq 8$. This corresponds to a standard deviation of 0.15. The motivation for this choice is that for large n, $\hat{p}$ has approximately a normal distribution, and this requirement then aligns with the case that approximately 95% of the distribution for $\hat{p}$ is within 0.3 of the actual value, p, whenever $0.25 \leq p \leq 0.75$. In the subsections 6.1.1 and 6.1.2, we examine two potential types of SHAPE distributions for paired and unpaired nucleotides, and investigate whether the distributions are separable.

6.1.1. *Differential shape distributions.* Introduced in Rice et al. (2014) in 2014, the inclusion of *differential SHAPE data* into existing secondary structure prediction models has been shown to improve secondary structure predictions by refining the data so that it reveals noncanonical pairings and tertiary structure interactions. To use predictions with differential SHAPE, three different sequences of SHAPE data are required, using the reagents NMIA, 1m6, and 1m7. The 1m7 SHAPE data is used as in the Deigan method, where it is included in the term $m \ln(1 + M_i) + b$. However, the NMIA and 1m6 data are processed, and then the positive amplitude signal from the difference of NMIA and 1m6 is calculated, and included in a new pseudoenergy term. More precisely, if $\{K_i\}_{i=1}^n$ is the NMIA data, and $\{L_i\}_{i=1}^n$ is the 1m6 data, then we have a new penalty of the form,

$$\Delta G_{\text{Diff}}(i) = d[K_i - L_i]_+,$$

for a parameter $d > 0$, where $[x]_+ = x$ if x is positive, and zero otherwise. The default value for d is 2.11.

Although the SHAPE pseudoenergies are written as two terms, it is possible to rewrite them as one term. Let $J_i = (K_i - L_i)_+$ be the positive amplitude in the differential SHAPE data penalty. Then, we have:

$$\Delta G_{\text{SHAPE}}(i) + \Delta G_{\text{Diff}}(i) = m \ln\Big(1 + \Big[e^{J_i d/m}(1 + M_i) - 1\Big]\Big) + b. \tag{16}$$

Therefore, we can view the differential SHAPE pseudoenergy as a rescaling of the SHAPE data, which is then used in the Deigan pseudoenergy from before. In Rice et al. (2014), the authors test the differential SHAPE method against 14 different RNA sequences. Here, we use the NMIA, 1m6, and 1m7 SHAPE data from these 14 sequences along with the transformation suggested by (16) with the default Deigan parameter $m = 2.8$ and default differential SHAPE parameter $d = 2.11$ to create new empirically-derived distributions for paired and unpaired nucleotides. We use the MATLAB statistics toolbox to fit the data to distributions, choosing gamma distributions as in Rice et al. (2014). For the unpaired distribution, we obtain a shape parameter $a_U = 0.400637$ and a scale parameter $b_U = 4.24848$. For the paired distribution, we have a shape parameter $a_P = 0.31486$ and a scale parameter $b_P = 0.869797$. This gives us that the unpaired distribution has mean 1.7021 and variance 7.2313 and that the paired distribution has mean 0.27386 and variance 0.23821.

Consider using these distributions with the estimator $\hat{p}$ described above. To check separability, we plug these statistics and $n = 8$ into Equation (15) to obtain the following estimator variance for the differential SHAPE distributions:

$$\operatorname{Var}\hat{p} = 0.4577198448p^2 - 0.02919426956p + 0.01459713478.$$

It is easy to check that this function is minimized at $p = 0.031891$, and the maximum value for $p \in [0.25, 0.75]$ is 0.25017, which is much too great for the distributions to be separable. In fact, even if the variance of the paired and unpaired distributions are cut by a factor of 10 each, the maximum value of $\hat{p}$ is scaled down by a factor of 10 to 0.025017, which is still too high to be separable.

6.1.2. *Hypothetical shape distributions.* In Süкösd et al. (2013), the authors proposed a hypothetical distribution for the unpaired SHAPE values that corresponded to a six-fold increase in reagent reactivity. In this case, the authors modelled the unpaired distribution as a Gaussian with mean 3.51 and standard deviation 1.78. Let X_{Hyp} be a random variable with this hypothetical distribution, and let X_{Diff} be a random variable with the differential SHAPE unpaired distribution. We consider a linear interpolation between the two distributions as follows: let $X_t = tX_{\text{Hyp}} + (1-t)X_{\text{Diff}}$ where X_{Hyp} and X_{Diff} are independent. We leave the unpaired distribution Y the same as the empirical distribution from before. We have the following result, which gives an indication of much the unpaired differential SHAPE distribution needs to be modified before separability is achieved.

Lemma 12. *The distributions corresponding to the random variables X_t and Y are separable for $t \geq 0.62828$.*

Proof. Because X_{Hyp} and X_{Diff} are independent, we have:

$$\operatorname{Var} X_t = t^2 \operatorname{Var} X_{\text{Hyp}} + (1-t)^2 \operatorname{Var} X_{\text{Diff}}.$$

Plugging into equation (15) (and emphasizing that $\hat{p}$ is now a function of t) gives the following:

$$\operatorname{Var}\hat{p}_t = \frac{p^2[t^2\operatorname{Var} X_{\text{Hyp}} + (1-t)^2\operatorname{Var} X_{\text{Diff}}] + (1-p)^2\operatorname{Var} Y}{n(t\mathbb{E}X_{\text{Hyp}} + (1-t)\mathbb{E}X_{\text{Diff}} - \mathbb{E}Y)^2} \tag{17}$$

For any fixed value of t and n, $\operatorname{Var}\hat{p}$ is a quadratic in p, meaning that its maximum value on the interval $p \in [0.25, 0.75]$ occurs at either $p = 0.25$ or $p = 0.75$. Plugging in $p = 0.25, n = 8$, and the variance values yields the following:

$$\operatorname{Var}\hat{p}_t\big|_{p=0.25,n=8} = \frac{0.08124765625t^2 - 0.1129890625t + 0.06335703125}{3.268502410t^2 + 5.503609180t + 2.316788410}.$$

This function is minimized at $t = 0.88809$, and is easily verified to be below 0.0225 for $t \in [0.0475, 1]$. Likewise, plugging in $p = 0.75$ and $n = 8$ yields:

$$\operatorname{Var} \hat{p}_t \big|_{p=0.75, n=8} = \frac{0.7312289062t^2 - 1.016901562t + 0.5092132812}{3.268502410t^2 + 5.503609180t + 2.316788410}.$$

This function is minimized at $t = 0.83382$, and is also easily verified to be below 0.0225 for $t \in [0.62828, 1]$. Combining these conditions on t completes the proof. □

6.2. **Beyond distance-one interpolations.** Now, consider the case where the structures S_1 and S_2 are known, but they differ by more than one helix, and we would like to reconstruct the proportion p of the mixture that comes from structure S_1. As before, let X be a random variable with the distribution of an unpaired SHAPE value, and let Y be a random variable with the distribution of a paired SHAPE value. Then, given a SHAPE sequence corresponding to a mixture of structures S_1 and S_2, we will have a sample of $n_1 > 0$ data points distributed as $D_1 = pX + (1-p)Y$ and $n_2 > 0$ data points of the form, $D_2 = (1-p)X + pY$. For each of these data sets, we can find an estimator for p as before:

$$\hat{p}_1 = \frac{\bar{D}_1 - \mathbb{E}Y}{\mathbb{E}X - \mathbb{E}Y}, \quad \text{and} \quad \hat{p}_2 = \frac{\bar{D}_2 - \mathbb{E}X}{\mathbb{E}Y - \mathbb{E}X}.$$

This suggests a family of unbiased estimators for p,

$$\hat{p} = s\hat{p}_1 + (1-s)\hat{p}_2, \tag{18}$$

where $s \in [0,1]$. We investigate which one is best to use. We can compute:

$$\operatorname{Var} \hat{p} = \frac{s^2}{n_1}\left[\frac{p^2 \operatorname{Var} X + (1-p)^2 \operatorname{Var} Y}{(\mathbb{E}X - \mathbb{E}Y)^2}\right] + \frac{(1-s)^2}{n_2}\left[\frac{p^2 \operatorname{Var} Y + (1-p)^2 \operatorname{Var} X}{(\mathbb{E}X - \mathbb{E}Y)^2}\right].$$

Let $\alpha = \frac{n_2}{n_1}$. Then, the minimum occurs at the s-value,

$$s = \frac{p^2(\operatorname{Var} X + \operatorname{Var} Y) - 2p\operatorname{Var} X + \operatorname{Var} X}{p^2(\alpha+1)(\operatorname{Var} X + \operatorname{Var} Y) - 2p(\operatorname{Var} X + \alpha \operatorname{Var} Y) + (\operatorname{Var} X + \alpha \operatorname{Var} Y)}.$$

For all of the distributions we have investigated so far, the maximum variance of $\hat{p}$ for $p \in [0,1]$ occurs at $p = 1$. With this in mind, we substitute in $p = 1$ to find the minimum maximum variance for these distributions occurs at:

$$s = \frac{\operatorname{Var} Y}{\operatorname{Var} Y + \alpha \operatorname{Var} X}.$$

For example, when using a mixture of the differential SHAPE and hypothetical distributions as in the previous section, this implies that one should use the estimator in Equation (18) with

$$s = \frac{1}{1 + (106.5543033t^2 - 148.1823771t + 74.09118853)\alpha}.$$

7. Conclusion

The SHAPE-directed energy models we have analyzed were designed to predict a single dominant structure in a distribution of RNA structures. We have seen that this resulted in the energy models collapsing the distributions to a single structure, even in the presence of simulated or binary SHAPE data derived from a mixture of structures. Because all noise is removed in binary SHAPE data, one would expect the prediction models to work best here. The uncentered crossover points and short crossover windows indicate that the energies in the models are influenced too heavily by the data. There are other options that may help improve multimodal distribution identification. SHAPE data may be preprocessed before being input into an existing model, or a new model must be produced that does not change the energies of the distributions as much.

Additionally, experimental SHAPE data currently does not distinguish enough between paired and unpaired distributions to reliably predict the proportion of structures contributing to a distribution, even when the structures are known in advance. However, by separating the unpaired distribution slightly more from the paired distribution, it may be possible to improve the prediction of proportions.

8. Acknowledgements

T. Greenwood was supported in part by NSF DMS-1344199. C. E. Heitsch was supported in part by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and by NIH R01 GM126554.

References

Baker, J. (1979). Trainable grammars for speech recognition. *Journal of the Acoustic Society of America*, 65(S132):54–550.

Booth, T. L. and Thompson, R. A. (1973). Applying probability measures to abstract languages. *IEEE Transactions on Computers*, C-22(5):442–450.

Clote, P. (2005). An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov–Jacobson energy model. *Journal of Computational Biology*, 12(1):83–101.

Clote, P., Kranakis, E., Krizanc, D., and Stacho, L. (2007). Asymptotic expected number of base pairs in optimal secondary structure for random RNA using the Nussinov–Jacobson energy model. *Discrete Applied Mathematics*, 155(6):759 – 787. Computational Molecular Biology Series, Issue V.

Couzin, J. (2002). Small RNAs make big splash. *Science*, 298(5602):2296–2297.

Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009). Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, 106(1):97–102.

Ding, Y., Chan, C. Y., and Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Research*, 32(Web Server issue):W135–W141.

Ding, Y. and Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301.

Doshi, K. J., Cannone, J. J., Cobaugh, C. W., and Gutell, R. R. (2004). Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105–105.

Doudna, J. A. (2000). Structural genomics of RNA. *Nature Structural Biology*, 7:954 EP –.

Dowell, R. D. and Eddy, S. R. (2004). Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5(1):71.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Eddy, S. R. (2014). Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annual Review of Biophysics*, 43(1):433–456. PMID: 24895857.

Eddy, S. R. and Durbin, R. (1994). Rna sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088.

Knudsen, B. and Hein, J. (1999). RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454.

Knudsen, B. and Hein, J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428.

Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S., and Bassler, B. L. (2004). The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. *Cell*, 118(1):69–82.

Leonard, C. W., Hajdin, C. E., Karabiber, F., Mathews, D. H., Favorov, O. V., Dokholyan, N. V., and Weeks, K. M. (2013). Principles for understanding the accuracy of SHAPE-directed RNA structure modeling. *Biochemistry*, 52(4):588–595. PMID: 23316814.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology : AMB*, 6:26–26.

Mathews, D. H. (2006). Revolutions in RNA secondary structure prediction. *Journal of Molecular Biology*, 359(3):526 – 532.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911 – 940.

Mathews, D. H. and Turner, D. H. (2006). Prediction of RNA secondary structure by free energy minimization. *Current Opinion in Structural Biology*, 16(3):270 – 278.

Nussinov, R. and Jacobson, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11):6309–6313.

Quarrier, S., Martin, J. S., Davis-Neulander, L., Beauregard, A., and Laederach, A. (2010). Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA*, 16(6):1108–1117.

Rice, G. M., Leonard, C. W., and Weeks, K. M. (2014). RNA secondary structure modeling at consistent high accuracy using differential shape. *RNA*, 20(6):846–854.

Rogers, E. and Heitsch, C. E. (2014). Profiling small RNA reveals multimodal substructural signals in a Boltzmann ensemble. *Nucleic Acids Research*, 42(22):e171–e171.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C., and Haussler, D. (1994). Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22(23):5112–5120.

Sükösd, Z., Swenson, M. S., Kjems, J., and Heitsch, C. E. (2013). Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Research*, 41(5):2807–2816.

Tu, K. C. and Bassler, B. L. (2007). Multiple small RNAs act additively to integrate sensory information and control quorum sensing in Vibrio harveyi. *Genes & Development*, 21(2):221–233.

Turner, D. H. and Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38(Database issue):D280–D282.

Zarringhalam, K., Meyer, M. M., Dotu, I., Chuang, J. H., and Clote, P. (2012). Integrating chemical footprinting data into RNA secondary structure prediction. *PLOS ONE*, 7(10):1–13.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148.