

Communication

Hee Rhang Yoon, Aaztli Coria, Alain Laederach, and Christine Heitsch*

Towards an understanding of RNA structural modalities: a riboswitch case study

<https://doi.org/10.1515/cmb-2019-0004>

Received July 22, 2019; accepted October 10, 2019

Abstract: A riboswitch is a type of RNA molecule that regulates important biological functions by changing structure, typically under ligand-binding. We assess the extent that these ligand-bound structural alternatives are present in the Boltzmann sample, a standard RNA secondary structure prediction method, for three riboswitch test cases. We use the cluster analysis tool RNAstructProfiling to characterize the different modalities present among the suboptimal structures sampled. We compare these modalities to the putative base pairing models obtained from independent experiments using NMR or fluorescence spectroscopy. We find, somewhat unexpectedly, that profiling the Boltzmann sample captures evidence of ligand-bound conformations for two of three riboswitches studied. Moreover, this agreement between predicted modalities and experimental models is consistent with the classification of riboswitches into thermodynamic versus kinetic regulatory mechanisms. Our results support cluster analysis of Boltzmann samples by RNAstructProfiling as a possible basis for de novo identification of thermodynamic riboswitches, while highlighting the challenges for kinetic ones.

Keywords: RNA secondary structures; Boltzmann distribution; suboptimal sample; multimodality; profiling

1 Introduction

Unlike Deoxy riboNucleic Acid (DNA), RiboNucleic Acid (RNA) exists in the cell as a single-stranded polymer molecule [14, 44]. As such, the bases of RNA are able to pair intramolecularly forming complex secondary structures [8, 18, 24, 39, 45] which are organized by tertiary interactions into the three-dimensional structure. Since 3D structural determination remains challenging for RNA molecules, computational predictions of RNA secondary structures from sequence are still an important resource for experimentalists.

Most algorithms used to predict RNA secondary structure are based on the nearest neighbor thermodynamic model (NNTM) [25, 41]. Originally, the goal was to predict the minimum free energy (MFE) structure, which remains a popular approach to this day [22, 30]. However, predictions of suboptimal structures [48, 49] and of base pairing probabilities under the Boltzmann partition function [12, 23, 26] have long been used to complement MFE predictions. These two approaches are unified by the method of sampling suboptimal secondary structures from the Boltzmann ensemble [5]. Moreover, these Boltzmann sample predictions often reveal that the suboptimal secondary structures are organized into two or more distinct modalities [4, 20, 31, 42].

In recent years, it has become increasingly clear that such distinct structural modalities should not be treated as an artifact of thermodynamic prediction methods. For instance, the existence of different base pairing configurations for the same sequence can be experimentally confirmed using a variety of RNA structural determination techniques including Nuclear Magnetic Resonance (NMR), single molecule fluorescence resonance energy transfer (sm-FRET) and chemical structure probing [1, 10, 15, 19, 29, 37]. It remains unclear,

Hee Rhang Yoon: School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 30332

Aaztli Coria: Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC, 27599

Alain Laederach: Department of Biology, University of North Carolina, Chapel Hill, NC, 27599

***Corresponding Author: Christine Heitsch:** School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 30332

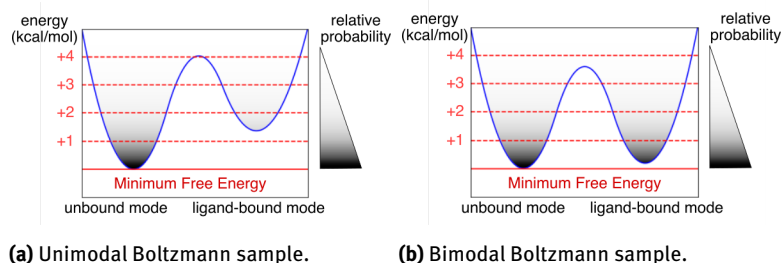


Figure 1: Schematic representations of two Boltzmann samples for a 2-state riboswitch with unbound and ligand-bound structural modes. The relative probability gradient indicates the sampling ratio with the MFE prediction. Recall that small changes in free energy result in significant differences in probability under the Boltzmann distribution; see Equation 2 on page 51 and related discussion. On the left, the ligand-bound mode is too thermodynamically unfavorable under the NNTM to appear in the Boltzmann sample. On the right, however, it would be sampled with sufficient frequency to be identified by RNAstructProfiling as a structural mode. Although it might seem like all riboswitches should resemble Figure 1a, our results demonstrate that thermodynamic riboswitches are better described by Figure 1b, although kinetic ones do indeed follow the first model.

however, to what extent predicted structural modalities are a true biological signal of alternative base pairing configurations.

Riboswitches are the canonical example of RNA sequences which are known to assume distinct structural conformations. However, the change in structure is typically mediated by the binding of a small molecule called a ligand. One might well expect, then, that only the unbound structure should be sufficiently thermodynamically favorable under the NNTM to appear in a Boltzmann sample prediction. If so, then although the sequence is known to assume more than one base pairing configuration, the Boltzmann sample (which does not include the ligand-binding biophysics) is effectively unimodal. This expectation is captured schematically in Figure 1a. If indeed this is true, the prospect of using Boltzmann sampling to identify new riboswitches and other multimodal RNA molecules does not look promising.

We investigate the validity of this assumption for three known riboswitches which have proposed base pairing models grounded in experimental data. Using the RNA suboptimal structure cluster analysis tool RNAstructProfiling [31], we find that there exist riboswitches whose ligand-bound thermodynamics are accessible to Boltzmann sampling, as represented in Figure 1b.

More precisely, we confirm that the extent to which helices from the proposed models are present in the predicted modalities aligns with the classification of riboswitches into “thermodynamic” and “kinetic” regulatory mechanisms [2, 11, 28]. This suggests that new thermodynamic riboswitches might be identifiable via Boltzmann sampling while also highlighting some challenges to be overcome as the sequence length/structural complexity of the putative riboswitch increases.

For the simple thermodynamic riboswitch considered in Section 3.1, both proposed models are clearly identified as structural modalities by profiling. In other words, this ligand-bound conformation is sufficiently favorable under the NNTM that the Boltzmann sample resembles Figure 1b rather than Figure 1a. Hence, the modalities reported by profiling are a true structural signal.

For the more complex thermodynamic riboswitch considered in Section 3.2, the situation is more complicated. To begin, all helices from the proposed models *do* appear in the Boltzmann sample. Hence, the assumption captured in Figure 1a again does not hold. In this case, though, the agreement between the modalities predicted by profiling and the proposed base pairing models is not as good as the previous one.

In particular, there is one crucial helix from the ligand-bound structure which, although present in the sample, has low frequency/estimated probability. While this could be interpreted as evidence for Figure 1a, that would be incorrect since one of the unbound structure models also includes this “missing” helix. In other words, factors such as folding kinetics [7] not included in the NNTM are likely a consideration here, but not ones directly related to the ligand-binding. As discussed in Section 3.2, this means that — subject to the accuracy of the NNTM approximation to folding biophysics — profiling identifies the structural signal in the Boltzmann sample with both high precision and high recall. Hence, while not perfect, this approach of

profiling Boltzmann samples to identify putative structural modes may yield useful insights into potential new thermodynamic riboswitches.

The situation for the kinetic riboswitch considered in Section 3.3 is markedly different. The helices from the unbound structure do indeed dominate the Boltzmann sample as expected from Figure 1a. In fact, no helices exclusive to the ligand-bound model are even sampled. Interestingly, the ligand-binding domain is clearly thermodynamically favorable if the kinetics of folding are simulated by sequential Boltzmann sampling for sequence prefixes. This indicates that it just might be possible to identify kinetic riboswitches by Boltzmann sampling as well, but a great many challenges remain.

Hence, we find that the original assumption holds for the kinetic riboswitch considered, but is violated by the thermodynamic ones. In the best/simplest case, the predicted modalities for the latter are in good agreement with the experimental models. This supports profiling of Boltzmann sample predictions as a useful method for identifying the conformational changes of simple thermodynamic riboswitches. For a more complex case, there is still a strong correlation between prediction and experiment, but the agreement is not quite as good — most likely because of the quality of the NNTM approximation independent of the ligand-binding question. This highlights that thermodynamic riboswitch identification based on the structural modes predicted by RNAstructProfiling is possible, but additional work is needed for longer sequences and/or more complex switches.

2 Materials and Methods

2.1 Riboswitches

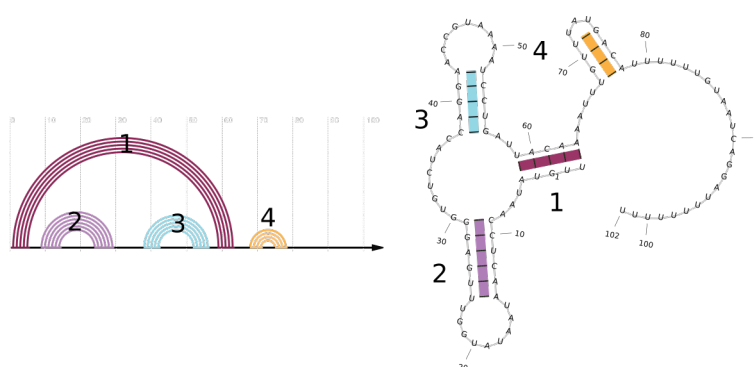


Figure 2: Two visualizations of the same RNA secondary structure: arc diagram (left) and planar model (right). Runs of consecutive base pairs, called helices, are both colored and numbered to indicate their correspondence. The helix $\{(i, j), \dots, (i + k - 1, j - k + 1)\}$ which begins with the base pairing between nucleotides in positions i and j and has total length k will be denoted by the triple (i, j, k) . Hence, the four helices pictured are $(1, 63, 5)$, $(9, 29, 6)$, $(38, 56, 5)$, and $(68, 78, 4)$ respectively.

A riboswitch is a segment of a messenger RNA (mRNA) molecule that regulates gene expression through a conformational change induced by the binding of a small molecule ligand. The structure of a riboswitch consists of two parts: an aptamer domain and an expression platform. Ligand binding to the aptamer domain causes the expression platform to change its conformation, resulting in the regulation of gene expression [6, 47]. Regulation can occur by controlling transcription (synthesis of RNA from DNA) or translation (synthesis of proteins from RNA). Transcriptional riboswitches control the formation of a terminator stem, whose presence causes transcription termination [32, 40]. On the other hand, translational riboswitches control the exposure of key regulatory elements like the Shine-Dalgarno (SD) sequence and the AUG start codon.

Riboswitches can be categorized into thermodynamic and kinetic switches [2]. Thermodynamic switches exist as a mix of conformations whose energies are similar to the minimum free energy [28]. Ligand bind-

ing stabilizes a particular conformation. Moreover, thermodynamic switches can switch back and forth between conformations depending on the concentration of the ligand. Kinetic switches, on the other hand, are riboswitches whose conformations have a large energy difference [11]. RNA folding can occur faster than transcription, and ligand binding to the RNA during co-transcriptional folding allows the RNA to fold into a high energy state.

RNA conformations and their changes can manifest in different ways. We focus here on differences in the base-pairing of the secondary structures. We will compare the predicted structural modes identified by profiling to putative models obtained via NMR spectroscopy or fluorescence spectroscopy. Each model will be presented in secondary structure visualizations, using both arc diagrams and planar models, as in Figure 2. We compute the free energy change under the nearest neighbor thermodynamic model (NNTM) using GT-fold [38], and the relative probability, which is the ratio of sampling probability of the model and sampling probability of the MFE structure, according to Equation 2 with $T = 310K$.

2.2 Boltzmann sample

A popular approach for RNA secondary structure prediction [22, 25, 30, 41] is to find a unique minimum free energy (MFE) structure under the nearest neighbor thermodynamic model (NNTM). However, since the NNTM is only an approximation to the folding biophysics, prediction accuracy has long been increased by considering suboptimal structures [48, 49] and/or base pairing probabilities under the Boltzmann partition function [12, 23, 26]. The ability to sample from the Boltzmann ensemble [5] allows one to examine structural alternatives to the MFE structure in proportion to their estimated probability under the NNTM. Hence, it can be used to search for signals of multimodality in RNA secondary structures [9, 20, 31, 33, 43].

A Boltzmann sample is a set of structures, typically of size 1000, sampled from the Boltzmann (i.e., Gibbs) ensemble [5]. In this distribution, a structure S with energy $\epsilon(S)$ exists with probability

$$P(S) = \frac{\exp[-\epsilon(S)/RT]}{Z}, \quad (1)$$

where R is the Boltzmann constant, T is the absolute temperature, and Z is the partition function $Z = \sum_{S \in \Omega} \exp[-\epsilon(S)/RT]$ summed over all possible states (in this case, secondary structures for the given sequence) $S \in \Omega$.

Since the sampling probability is exponential in energy, a small difference in energy leads to a much larger difference in sampling probability. Given two structures S_1 and S_2 with energies ϵ_1 and ϵ_2 , the ratio of their sampling probabilities is

$$\frac{P(S_1)}{P(S_2)} = \exp\left[\frac{\epsilon_2 - \epsilon_1}{RT}\right]. \quad (2)$$

When $T = 310K$, a decrease in energy by 1 kcal/mol leads to approximately 5 fold increase in sampling probability. Hence, one of the many distinct structures which is 4 kcal/mol above MFE prediction is unlikely to be sampled at all since the relative probabilities would $1.4 \times 10^{-3} : 1$.

2.3 Profiling

Intuitively, we consider a mode in a Boltzmann sample to be a set of similar structures whose collective frequency is high enough to be considered “signal” rather than “noise.” More precisely, we will consider each (selected) profile identified by the cluster analysis method RNAstructProfiling [31] to be a structural mode. In contrast to other cluster analysis methods [3, 17, 36], profiling is explicit about filtering the thermodynamic noise from the stochastic sampling. Additionally, profiling facilitates comparisons between different clusters in the Boltzmann sample by highlighting structural similarities and differences in the summary profile graph, like the one pictured in Figure 3.

Given a Boltzmann sample, typically with 1000 structures, profiling first identifies the helix classes present. A helix class is an equivalence class of helices which are all subsets of the same maximal helix.

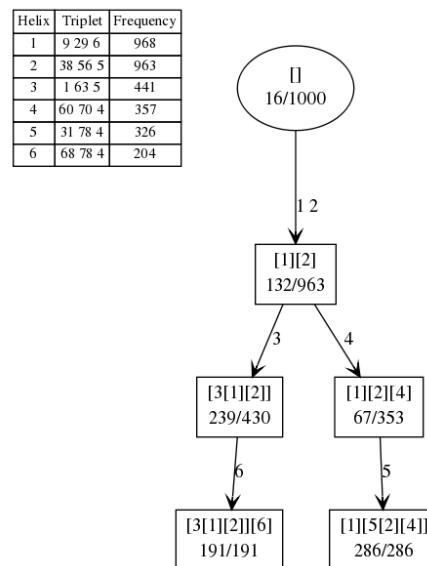


Figure 3: A profile graph for the sequence from Figure 2 up to position 78. Profiling identified 5 distinct structural modes present in the Boltzmann sample, corresponding to the five rectangles/selected profiles listed. The table lists the feature locations by (i, j, k) triplet and sampled frequencies. For instance, pairings from the purple helix (which is maximal) are the first feature and occurred in 968 of the structures sampled. The four colored helices from Figure 2 correspond to the profile $[3[1][2]][6]$, where the brackets indicate the nesting relationship. This selected profile appears in the bottom left rectangle. It differs from the selected profile above by the sixth feature. The profile $[3[1][2]]$ has a specific frequency of 239, and a general frequency of 430 since the latter includes the structures from the $[3[1][2]][6]$ profile.

A helix is maximal if the run of base pairings cannot be extended any further. Next, profiling selects high frequency helix classes as features. The frequency cutoff is determined by maximizing the average Shannon information entropy. After feature selection, the next step identifies the profiles present in the Boltzmann sample. A profile is an equivalence class of sampled structures which have the same set of features. Some profiles may have been sampled with very low frequency, so the Shannon entropy method is again used to select only the most informative. The relationships among these selected profiles are then presented in the summary profile graph.

The summary profile graph is generated by computing a transitive reduction on the set of selected profiles, which are drawn as rectangles. If additional vertices are needed to connect the graph, as in Figure 4 on page 53, these “intersection” profiles are drawn in dashed ovals. The “root” node, which is the oval at the top of the graph, indicates the profile common to all sampled structures. In this example, there were no features common to all sampled structures, so this is the empty profile. The “numerator” indicates the number of structures with exactly this profile, called the specific frequency. The “denominator” gives the number of structures which contain at least this set of features, called the general frequency. In this example, 132 of the 1000 structures sampled contained only the features 1 and 2, along with other lower frequency pairings, whereas 963 contained those plus other additional features. Directed edges are labeled by the features which are added as the profiles grow larger, progressing down through the graph from the root node. Although the graph pictured in Figure 3 is a tree, this is by no means a requirement.

The web version of profiling is available at <http://rnaprofiling.gatech.edu/>. The code is freely available at <https://github.com/gtDMMB/RNAstructProfiling>. By default, profiling samples 1000 suboptimal structures using *Turner 99* energy parameters [24] and dangle option d2, which adds dangling energies for nucleotides on the ends of multi-loops and external loops.

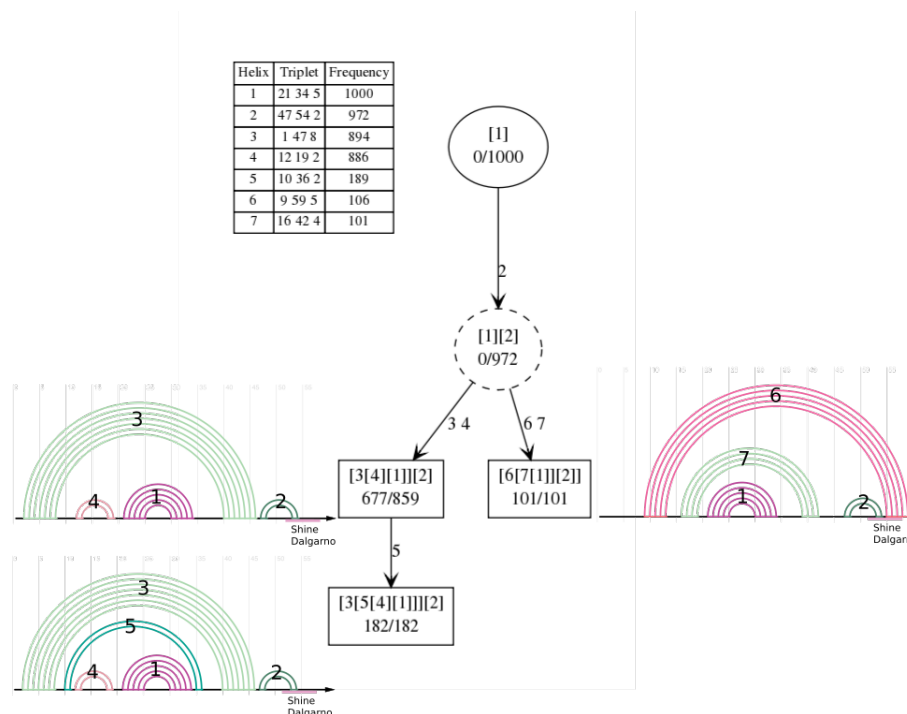


Figure 4: Profiling graph of SAM-III riboswitch showing the three structural modes identified. The arc diagrams visualize the featured base pairings for each selected profile. Recall that the dashed oval is an intersection profile, included to connect the graph.

3 Analysis and Results

We use profiling to identify predicted structural modalities for three riboswitches, two whose regulation is thermodynamically controlled and one kinetic. We note that a second kinetic switch was analyzed, but results were essentially the same. Some additional details are given at the end of Section 3.3.

3.1 SAM-III riboswitch: thermodynamic switch

The SAM-III riboswitch regulates *metK* gene expression in bacteria by controlling translation initiation. In the absence of the ligand, the riboswitch folds so that the Shine Dalgarno (SD) sequence is exposed, allowing translation to take place. Upon binding of *S*-adenosylmethionine (SAM), the riboswitch refolds into an alternate conformation in which the SD sequence is sequestered [46].

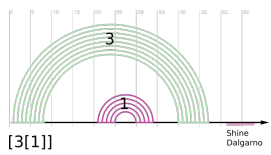
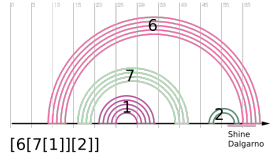
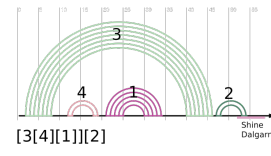
Profiling found 28 helix classes present in the Boltzmann sample generated, and selected 7 as features. With these features, the sampled structures partitioned into 8 profiles, and 3 had significant enough presence to be selected as structural modes. Note that profile $[3[4][1]][2]$ includes the MFE structure.

The summary profile graph displayed in Figure 4 illuminates the relationship among these 3 modes. In particular, one can see that the primary distinction is between structures which contain features 3 and 4 versus 6 and 7. Among the former, there are two types, depending on the presence of feature 5 which is a short (maximal) helix of length 2.

Comparing this analysis to putative models of the SAM-III riboswitch shows that profiling captures the important conformation change. Table 1 illustrates NMR spectroscopy based models of the unbound and ligand-bound structures presented in [46], along with the MFE structure.

Selected profile $[6[7[1]][2]]$ which was found in 10.1% of the Boltzmann sample coincides exactly with the proposed ligand-bound model whereas the unbound structure has the profile $[3[1]]$. Although this exact

Table 1: The two NMR-based models (unbound and ligand-bound) proposed in [46] and minimum free energy (MFE) prediction of SAM-III riboswitch. The helices are labeled to match Figure 4, and the profile is also given. Energies were computed using GTfold [38]. The difference of +0.60 kcal/mol between the ligand-bound and MFE structures results in a relative sampling probability of 0.38:1 according to Equation 2 with $T = 310K$. The +1.7 kcal/mol difference from the MFE gives a 0.06:1 relative sampling probability for the unbound structure.

Structure	Structure representations	Energy
unbound		-24.90 kcal/mol
ligand-bound		-26.00 kcal/mol
MFE		-26.60 kcal/mol

combination of features was not identified, 85.9% of the Boltzmann sample included features 1 and 3, along with 2 and 4. Since these differences are all short (maximal) helices with length 2, we understand profiles [3[4][1]][2] and [3[5[4][1]][2] as variants on the unbound model whose energies were lowered via the additional base pairings.

To appreciate why profile [3[1]] was not among the selected structural modes, we have only to consider the predicted free energy changes under the NNTM listed in Figure 4. In the Boltzmann distribution, each 1 kcal/mol difference is a factor of ~ 5 in probability. Given the difference of 1.7 kcal/mol from the MFE, the probability of sampling exactly features 1 and 3 is quite low. This is an aspect of how thermodynamic optimization methods can over-predict base pairs, although one which is easily addressed since the additional pairings are all short helices.

In other words, the recall is excellent, but the precision could be improved. Here, the recall of 1 is being computed as the ratio of “true positive” helices shared by predicted features and proposed models to the total number of helices in the proposed model. The precision of 0.71 is being computed as the ratio of the same true positive helices to the total number of predicted features.

Nonetheless, for the SAM-III riboswitch, profiling identifies that structures closely related to the proposed unbound model dominant the Boltzmann sample, while the ligand-bound model is a significant alternative structural mode. This analysis provides a proof-of-principle result that thermodynamic riboswitches can be detected from the different structural modalities identified by profiling from a Boltzmann sample prediction.

3.2 add-riboswitch: thermodynamic switch

The adenine sensing riboswitch is found on chromosome II of *Vibrio vulnificus*. In the absence of the ligand, the SD sequence and the AUG start codon are sequestered, thereby repressing translation. The ligand binding

changes the riboswitch conformation so that SD sequence and the AUG start codon are exposed, allowing translation initiation [29].

The previous SAM-III analysis is a best case, with a clear correspondence between predicted structural modes and putative base pairing models. However, the add-riboswitch is a more complicated example. To begin, its length is 112 nucleotides, compared to the 59 length SAM-III sequence. This additional 53 nucleotides substantially increases the number of low energy suboptimal secondary structures possible and hence the structural diversity in the predicted Boltzmann sample.

For this riboswitch, profiling finds 193 helix classes (versus 28 for SAM-III) present in the Boltzmann sample. However, nearly all are low or very low frequency, since only 9 (versus 7 before) are selected as features. With these features, the sampled structures are partitioned in 41 profiles (versus 8). Again, though, nearly all are low frequency, since only 5 (versus 3) have significant enough presence to be selected as structural modes.

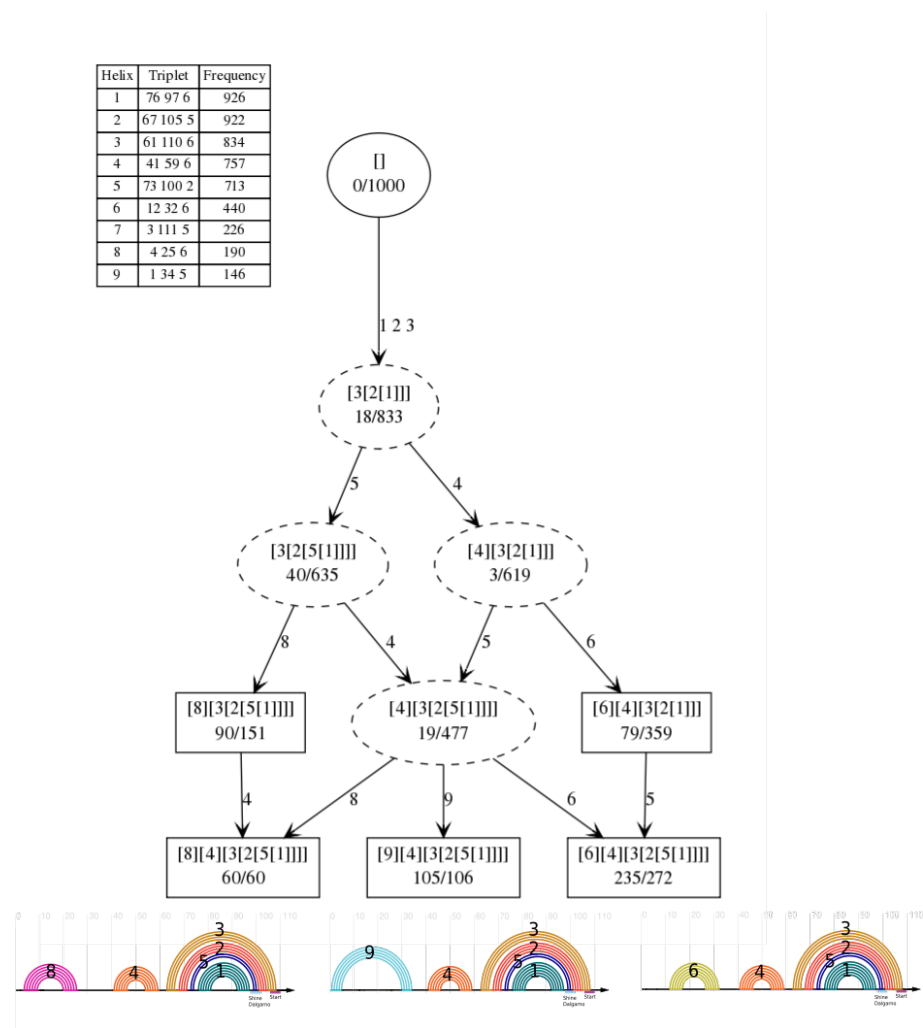


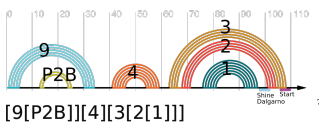
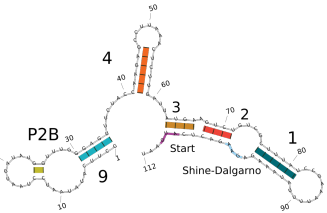
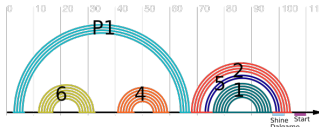
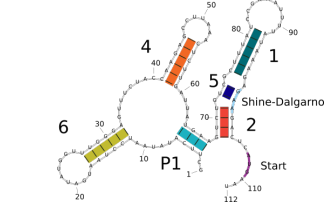
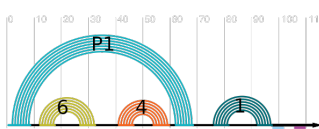
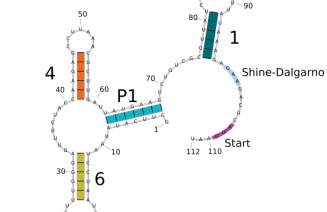
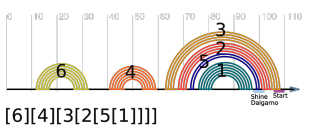
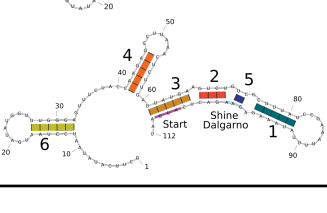
Figure 5: Profiling graph of add-riboswitch showing the five structural modes identified. In this case, four intersection profiles were needed to connect the graph. The arc diagrams visualize the three selected profiles with the most featured base pairings.

As before, the summary profile graph displayed in Figure 5 illuminates the relationship among these 5 modes. All selected profiles include features 1, 2, and 3, which nest as $[3[2[1]]]$. All but one also contain the two base pairs from feature 5, which are inserted into this run as $[3[2[5[1]]]$. If this extended helix is considered a structural unit, then the structural variation identified by profiling is due to features 4, 6, 8, and 9.

All but one of the modes contain feature 4. Since it is present in 75.7% of the sample, but the feature threshold cutoff is below 14.6%, we know that alternatives are lower frequency. Features 6, 8, and 9 are all competing for base pairings in the sequence region from nucleotides 1 to 34. From this we understand that thermodynamic optimization predicts a stable helix in this region, but the NNTM cannot resolve exactly which pairings are involved. Hence, although profiling predicts 5 structural modes, they are all closely related since 88.3% of the sample contained features 1, 2, and 3, and 61.9% contained 4 as well.

We note that no profiles containing feature 7 were selected as structural modes. Under closer inspection of the Boltzmann sample, this seems reasonable since the two most frequent were [7[4]] which had 37 structures and [7[6][4][3[2[5[1]]]] with 36. However, we also observed that feature 7 can be considered an alternative to feature 3 since their maximal helices overlap; of the 166 structures without #3, 118 contain #7. (There were 108 structures which contained pairings from both #3 and #7.) This indicates that it may be useful to consider low frequency alternatives to high probability pairings.

Table 2: The three NMR spectroscopy based models (apoB, apoA, and holo) and the MFE predicted structure for the add-riboswitch. According to the NNTM and Equation 1, the relative probabilities of sampling the proposed models versus the MFE prediction are 0.0021 , 6.7×10^{-4} , and 1.5×10^{-4} , respectively.

Structure	Structure representation	Energy
 [9[P2B]][4][3[2[1]]]		-20.30 kcal/mol
 [P1[6][4]][2[5[1]]]		-19.60 kcal/mol
 [P1[6][4]][1]		-18.70 kcal/mol
 [6][4][3[2[5[1]]]]		-24.1 kcal/mol

As with our previous SAM-III analysis, we compare the features identified by profiling with the putative helices in the NMR spectroscopy based models. Of the 9 features listed in Figure 5, all but 2 (#7 and #8) are present in the model. Conversely, of the 9 helices in the putative models from Table 2, all but 2 (P1 and P2B) are predicted features. Hence, the recall and precision are both 0.78.

Since 3 of 4 helices from the proposed ligand-bound model are featured, we conclude that profiling finds evidence for this conformation. Under further inspection, we will see that the remaining ligand-bound helix

is present in the Boltzmann sample, and hence that the original assumption captured by Figure 1a on page 49 is too strong. Since apoA mediates the transition from apoB to holo, it would seem that Figure 1b is a better approximation to the folding landscape for this riboswitch — subject to the NNTM approximation to folding biophysics.

As illustrated in Figure 6, it is not possible to increase this recall without substantially decreasing the precision. Of the two “false negative” helices, the critical missing one is P1. This helix is a crucial difference between the two proposed unbound structures, apoB and apoA, since it forms the aptamer domain, which persists in the ligand-bound holo conformation. However, P1 was only sampled in 36 structures (out of 1000), and there are 20 other helix classes with higher frequency which were also not featured by profiling.

This low frequency is a result of P1 being in direct competition with features 3 and 7. The maximal helix for P1 is (1, 69, 8) which also overlaps with feature 2. However, we see from the proposed apoA structure that portions of the P1 and feature 2 maximal helices can coexist as (3, 67, 4) and (68, 104, 4). Given the relative probabilities, these ‘apoA’-like structures were only 2.7% of the sample. In contrast, no structures containing both P1 and either feature 3 or 7 were sampled, although features 3 and 7 coexisted in 10.8% of the structures.

Similarly, only 5 ‘holo’-like structures were sampled which contained P1 and features 1 but not 2. (Every structure which contained P1 also contained #4 and #6.) These structures all contained additional base pairings, and each was assigned its own profile. Hence, unlike the SAM-III riboswitch, there is no structural signal for the proposed ligand-bound holo conformation in the Boltzmann sample.

In contrast, though, ‘apoB’-like structures which contain features 3, 4, and 9 account for 14.5% of the sample. (All but 1 of 834 structures which contain #3 also contain #1 and #2.) Of these, 6.7% contain P2B, 10.6% contain #5, and 1.9% contain P2B but not #5. Both P2B and #5 are short helices of length 2.

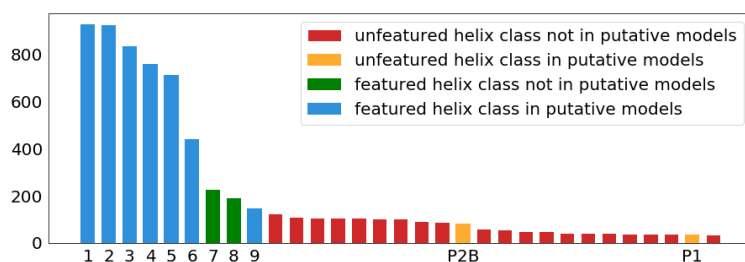


Figure 6: Distribution of helix classes with frequency 30 or higher in the Boltzmann sample prediction for the add-riboswitch analyzed. P1 is 30th with frequency 36, and P2B is 19th with 82. Taking the helices of the putative model as the target and the profiling features as the prediction, 7 (blue) are “true positives,” 2 (green) are “false positives,” and 2 (yellow) are “false negatives.” This yields a recall, a precision, and an F1-score (harmonic mean of precision and recall) of 0.78. Increasing the recall to 0.89 by including helix classes up to P2B (resp. 1 for P1) as features decreases the precision to 0.42 (resp. 0.3) and reduces the F1-score from 0.78 to 0.57 (resp. 0.42). Hence, profiling achieves a good balance between these two competing objectives.

Finally, we note that all of the helices in the MFE structure occur in one or more of the putative models, as illustrated in Figure 7. This supports the general understanding that high probability base pairings are most likely to be a true structural signal while illustrating that the open challenge is to identify which of the lower probability ones are also.

Given the precision and recall of the predicted features for the add-riboswitch, profiling is clearly extracting biologically relevant information from the Boltzmann ensemble. In this case, though, the correspondence between the structural modes selected and the proposed models based on NMR spectroscopy is less good than the SAM-III riboswitch. A particular challenge is to identify pairings like P1 that characterize structural modes but which are only sampled at very low frequency. Since the critical missing helix also occurs in an unbound model, other factors besides ligand-binding (such as folding kinetics) are more likely to be the cause of the lower accuracy. To increase recall without decreasing precision, approaches for improving the NNTM approx-

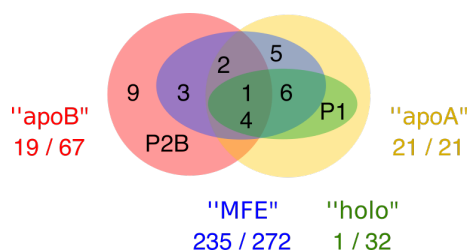


Figure 7: Venn diagram highlighting the similarities and differences among helix classes for the add-riboswitch base pairing models with the MFE structure's profile. Frequencies are given as specific / general. Recall that the specific frequency is the number of sampled structures with exactly that profile while the general is the number which contain at least those helix classes (and possibly others).

imation to folding biophysics — such as including auxiliary information from SHAPE and other footprinting data — may well be needed.

3.3 pbuE riboswitch: kinetic switch

We now turn from riboswitches whose regulation is classified as “thermodynamic” to a “kinetic” one. The pbuE riboswitch found in *Bacillus subtilis* is also adenine sensing. Unlike the previous two switches, pbuE regulates transcription, and not translation. By default, a long terminator stem forms, resulting in transcription attenuation. However, ligand binding during co-transcriptional folding stabilizes an alternative structure, disrupting formation of the terminator stem and allowing transcription to continue [21].

The pbuE sequence has length 102, and its structural diversity does fall between the SAM-III and add-riboswitch. The Boltzmann sample has 73 helix classes reduced to 6 features, and 5 structural modes selected from 7 profiles. In this case, however, the profile graph given in Figure 8 reveals that 4 of these modes are very closely related.

Unsurprisingly, all structures sampled included features 1 and 2, which is the transcription-attenuating terminator stem with 22 base pairs. The profiles then split on feature 3. Features 4 and 5 are almost identical. Feature 6 explicitly fills in 3 additional base pairs, however structures belonging to the 2 related profiles without #6 do sometimes include pairings in the region 1,...,17 — just with much lower frequency. Hence, this is probably not a significant structural difference.

On further review, then, the Boltzmann sample has at most two significant structural modes. Additional analysis reveals that all but one structure in the profile [1[2]] also contains pairings from helix class #7 which has frequency 58 and triplet (9, 29, 6). This is interesting because it is the helix P0 from the putative models.

Table 3 illustrates the fluorescence spectroscopy based models of the pbuE riboswitch proposed in [21]. The unbound model consists of the terminator stem along with the helix P0 just discussed. When the aptamer domain, which consists of the first 63 nucleotides, has been transcribed, the riboswitch folds into a three-way junction formed by helices P0, P1, and P2 which can be stabilized by ligand-binding. Since the helices P1 and P2, whose maximal triples are (1, 63, 5) and (38, 56, 5), directly conflict with feature 1, ligand binding disrupts the terminator stem formation and transcription can proceed.

In comparing the predicted structural modes to the putative models, we did find the unbound model present in 5.8% of the sample, although manual inspection of low frequency alternatives was necessary to identify the P0 pairings as the (9, 29, 6) helix in the [1[2]] profile. Clearly, the terminator stem dominates the sample, and this is consistent with the kinetics of co-transcriptional regulation as well as Figure 1a on page 49.

The P1 and P2 aptamer helices are not sampled at all, which agrees with the negligible probability of the ligand-bound structure relative to the MFE one. Interestingly, though, these helices are clearly present when we simulate the transcription process. To do this, we generated a Boltzmann sample and profiled it for initial prefixes of the pbuE sequence from length 65 up to length 101. For each of the 37 new samples, plus the original full sequence, we tracked whether the helices P0, P1, and P2 were selected as features and also

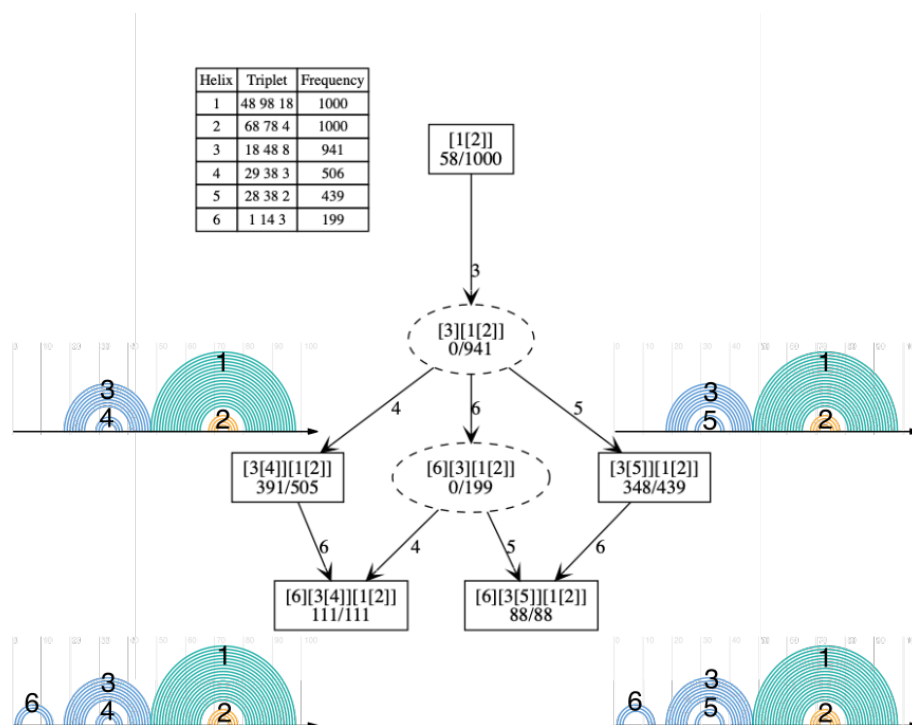


Figure 8: Profiling graph of the Boltzmann sample of pbuE riboswitch with five modes identified from the Boltzmann sample.

whether the terminator stem helices (now denoted TS1 and TS2) were. Figure 9 demonstrates that the aptamer domain helices are featured up to length 85, at which point the terminator stem begins to dominate.

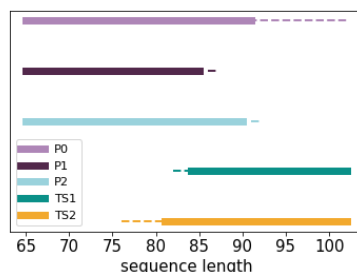


Figure 9: Graph indicating pbuE prefix length at which helices from the aptamer domain (P0, P1, and P2) and the terminator stem (now denoted TS1 and TS2) are either featured (solid), sampled (broken), or nonexistent (blank). The broken line shows if the given helix is present in at least one of the sampled structures. A clear transition occurs as the transcription simulation proceeds from position 80 to 90. We conclude that this is the critical window for ligand-binding to stabilize the aptamer domain and prevent transcription attenuation.

In fact, when we examine the profiling graphs for prefixes of length 65 to 77, the profile $[P1[P0][P2]]$ is the dominant structural mode. For lengths 78 to 85, profiles with this three-way junction were selected as structural modes, along with other ones in which it was disrupted. For lengths 86 to 91, all structural modes contained the terminator stem helices. Beyond 92, the results were very similar to Figure 8.

As we had expected, our profiling analysis of the pbuE riboswitch shows that Boltzmann sampling does not take into account folding kinetics, especially when mediated by ligand-binding. A novel outcome, however, was the simplicity with which the co-transcriptional folding process could be simulated and analyzed, yielding useful insights.

Table 3: Fluorescence spectroscopy based models (unbound and ligand-bound) and MFE prediction for the pbuE riboswitch. Given the differences in free energy with the MFE, the relative probability of sampling the unbound structure is 0.012:1 while the ligand-bound is 6.8×10^{-15} :1.

Structure	Structure representation	Energy
unbound		-32.30 kcal/mol
ligand-bound		-14.90 kcal/mol
MFE		-35.00 kcal/mol

As remarked, we analyzed a second kinetic switch, the thiM riboswitch from E.coli (TPP riboswitch), and the results were sufficiently similar to pbuE not to report separately. In particular, we simulated the co-transcriptional folding process by profiling the initial prefixes of the TPP sequence from length 80. Initially, profiling featured helices from the putative ligand-bound model. However, as the length of the sequence increased, such helices ceased to be features as helices from the unbound model started to dominate.

4 Discussion

We draw several conclusions from our RNAstructProfiling cluster analysis of Boltzmann sample predictions for three different riboswitches whose conformational change under ligand-binding is manifested at the base pairing level, resulting in different secondary structures from the unbound model.

First, in agreement with Figure 1a on page 49, we should not expect thermodynamic optimization methods to detect “kinetic” riboswitches like pbuE — at least not from a single Boltzmann sample prediction. This is because the conformational change is fundamentally dependent on the folding dynamics mediated by ligand-binding as the sequence is transcribed. Nonetheless, we were intrigued that the pbuE aptamer domain helices were featured by the profiler for initial sequence prefixes. This suggests that it may be possible to simulate folding kinetics for riboswitch detection with this sequential sampling and profiling approach.

It would certainly be interesting to compare against existing methods for detecting locally stable secondary structure motifs [13, 16, 34].

Conversely, though, our results indicate that Figure 1b is a better representation for the two thermodynamic riboswitches studied since all helices from the proposed ligand-bound models were sufficiently favorable under the NNTM to appear in the Boltzman sample predictions. However, the potential for profiling to identify the (approximate) structural modes for thermodynamic riboswitches is dependent on the diversity of the Boltzmann sample, which in turn correlates with sequence length and switch complexity. As remarked, the SAM-III riboswitch is a best case. Nonetheless, we expect that other short, 2-state thermodynamic riboswitches might well be equally successful.

In contrast, our second thermodynamic riboswitch was almost twice as long, as well having a proposed 3-state system. Out of the 9 features predicted by profiling, 7 agreed with the putative model (which has 9 proposed helices). As remarked, this is both good precision and good recall at the helix level. To improve the accuracy of the predicted structural modes, however, it would be necessary to identify the critical missing P1 base pairings, which is one of at least 20 other low energy helices. To solve this needle/haystack problem, it is likely that additional computational approaches like recent advances in nonredundant Boltzmann sampling [27] and in sampling with SHAPE data [35] may be needed to improve the quality of the NNTM approximation to folding biophysics.

Acknowledgement: The authors would like to thank Professor Shan Zhao and the Department of Mathematics at the University of Alabama for hosting the excellent 2019 NSF-CBMS Conference: Mathematical Molecular Bioscience and Biophysics, Professor Guowei Wei for his inspirational lectures at that meeting, and the Computational and Mathematical Biophysics journal for organizing this special issue. Thanks also are due to the anonymous reviewers whose thoughtful feedback significantly improved the paper.

This work was supported by funds from the National Institutes of Health (R01GM126554 to CH, R01GM101237 to AL, and F31GM130040 to AC) and the National Science Foundation (DMS1344199 to CH).

References

- [1] Bothe, J. R., Nikolova, E. N., Eichhorn, C. D., Chugh, J., Hansen, A. L., and Al-Hashimi, H. M. (2011). Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. *Nature methods*, 8(11):919–931.
- [2] Coppins, R. L., Hall, K. B., and Groisman, E. A. (2007). The intricate world of riboswitches. *Current opinion in microbiology*, 10(2):176–81.
- [3] 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(suppl_2):w135–w141.
- [4] Ding, Y., Chan, C. Y., and Lawrence, C. E. (2005). RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11(8):1157–1166.
- [5] Ding, Y. and Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301.
- [6] Edwards, T. E., Klein, D. J., and Ferré-D'Amaré, A. R. (2007). Riboswitches: small-molecule recognition by gene regulatory RNAs. *Current Opinion in Structural Biology*, 17(3):273–279.
- [7] Flamm, C. and Hofacker, I. L. (2008). Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatsh Chem*, 139(4):447–457.
- [8] Fresco, J. R., Alberts, B. M., and Doty, P. (1960). Some molecular details of the secondary structure of ribonucleic acid. *Nature*, 188(4745):98–101.
- [9] Freyhult, E., Moulton, V., and Clote, P. (2007). Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, 23(16):2054–2062.
- [10] Fürtig, B., Richter, C., Wöhrner, J., and Schwalbe, H. (2003). NMR Spectroscopy of RNA. *ChemBioChem*, 4(10):936–962.
- [11] Gilbert, S. D., Stoddard, C. D., Wise, S. J., and Batey, R. T. (2006). Thermodynamic and Kinetic Characterization of Ligand Binding to the Purine Riboswitch Aptamer Domain. *Journal of Molecular Biology*, 359(3):754–768.
- [12] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188.
- [13] Hofacker, I. L., Priwitzer, B., and Stadler, P. F. (2004). Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2):186–190.

- [14] Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and Zamir, A. (1965). Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465.
- [15] Homan, P. J., Favorov, O. V., Lavender, C. A., Kursun, O., Ge, X., Busan, S., Dokholyan, N. V., and Weeks, K. M. (2014). Single-molecule correlated chemical probing of RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 111(38):13858–13863.
- [16] Horesh, Y., Wexler, Y., Lebenthal, I., Ziv-Ukelson, M., and Unger, R. (2009). RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry. *BMC bioinformatics*, 10(1):76.
- [17] Huang, J., Backofen, R., and Voß, B. (2012). Abstract folding space analysis based on helices. *RNA*, 18(12):2135–2147.
- [18] Jaeger, J. A., Turner, D. H., and Zuker, M. (1989). Improved predictions of secondary structures for RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 86(20):7706–7710.
- [19] Krokhotin, A., Mustoe, A. M., Weeks, K. M., and Dokholyan, N. V. (2017). Direct identification of base-paired RNA nucleotides by correlated chemical probing. *RNA*, 23(1):6–13.
- [20] Kutchko, K. M., Sanders, W., Ziehr, B., Phillips, G., Solem, A., Halvorsen, M., Weeks, K. M., M, N., Moorman, N., and Laederach, A. (2015). Multiple conformations are a conserved and regulatory feature of the RB1 5' UTR. *RNA*, 21(7):1274–1285.
- [21] Lemay, I.-F., Penedo, J. C., Tremblay, R., Lilley, D. M., and Lafontaine, D. (2006). Folding of the Adenine Riboswitch. *Chemistry & Biology*, 13(8):857–868.
- [22] Lorenz, R., Bernhart, S. H., Höner zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26.
- [23] Mathews, D. H. (2004). Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190.
- [24] 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.
- [25] Mathews, D. H. and Turner, D. H. (2006). Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol*, 16(3):270–278.
- [26] McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119.
- [27] Michálik, J., Touzet, H., and Ponty, Y. (2017). Efficient approximations of RNA kinetics landscape using non-redundant sampling. *Bioinformatics*, 33(14):i283–i292.
- [28] Quarta, G., Sin, K., and Schlick, T. (2012). Dynamic energy landscapes of riboswitches help interpret conformational rearrangements and function. *PLOS Computational Biology*, 8(2):1–14.
- [29] Reining, A., Nozinovic, S., Schlepckow, K., Buhr, F., Fürtig, B., and Schwalbe, H. (2013). Three-state mechanism couples ligand and temperature sensing in riboswitches. *Nature*, 499(7458):355–359.
- [30] Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):129.
- [31] Rogers, E. and Heitsch, C. E. (2014). Profiling small RNA reveals multimodal substructural signals in a Boltzmann ensemble. *Nucleic Acids Research*, 42(22).
- [32] Sashital, D. G. and Butcher, S. E. (2006). Flipping Off the Riboswitch: RNA Structures That Control Gene Expression. *ACS Chemical Biology*, 1(6):341–345.
- [33] Schroeder, S. J. (2018). Challenges and approaches to predicting RNA with multiple functional structures. *RNA*, 24(12):1615–1624.
- [34] Soldatov, R. A., Vinogradova, S. V., and Mironov, A. A. (2014). RNASurface: fast and accurate detection of locally optimal potentially structured RNA segments. *Bioinformatics*, 30(4):457–463.
- [35] Spasic, A., Assmann, S. M., Bevilacqua, P. C., and Mathews, D. H. (2018). Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Research*, 46(1):314–323.
- [36] Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. (2006). RNASHapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503.
- [37] Stephenson, J. D., Kenyon, J. C., Symmons, M. F., and Lever, A. M. (2016). Characterizing 3D RNA structure by single molecule FRET. *Methods*, 103:57–67.
- [38] Swenson, M. S., Anderson, J., Ash, A., Gaurav, P., Sükösd, Z., Bader, D. A., Harvey, S. C., and Heitsch, C. E. (2012). GTfold: enabling parallel RNA secondary structure prediction on multi-core desktops. *BMC research notes*, 5:341.
- [39] Tinoco, I., Uhlenbeck, O. C., and Levine, M. D. (1971). Estimation of Secondary Structure in Ribonucleic Acids. *Nature*, 230(5293):362–367.
- [40] Tucker, B. J. and Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Current Opinion in Structural Biology*, 15(3):342–348.
- [41] Turner, D. H. and Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, 38(suppl 1):D280–D282.
- [42] Voß, B., Giegerich, R., and Rehmsmeier, M. (2006). Complete probabilistic analysis of RNA shapes. *BMC Biology*, 4(1):5.
- [43] Voss, B., Meyer, C., and Giegerich, R. (2004). Evaluating the predictability of conformational switching in RNA. *Bioinformatics*, 20(10):1573–1582.

- [44] Waterman, M. (1978). Secondary structure of single-stranded nucleic acids. *Advances in Mathematics, Supplementary Studies*, 1:167—212.
- [45] Westhof, E. and Jaeger, L. (1992). RNA pseudoknots. *Current Opinion in Structural Biology*, 2(3):327–333.
- [46] Wilson, R. C., Smith, A. M., Fuchs, R. T., Kleckner, I. R., Henkin, T. M., and Foster, M. P. (2011). Tuning riboswitch regulation through conformational selection. *Journal of Molecular Biology*, 405(4):926–38.
- [47] Winkler, W. C. and Breaker, R. R. (2003). Genetic control by metabolite-binding riboswitches. *ChemBioChem*, 4(10):1024–1032.
- [48] Wuchty, S., Fontana, W., Hofacker, I. L., and Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165.
- [49] Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52.