RESEARCH ARTICLE

Structural prediction of RNA switches using conditional base-pair probabilities

Amirhossein Manzourolajdad*, John L. Spouge

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

* manzouro@ncbi.nlm.nih.gov

Abstract

An RNA switch triggers biological functions by toggling between two conformations. RNA switches include bacterial riboswitches, where ligand binding can stabilize a bound structure. For RNAs with only one stable structure, structural prediction usually just requires a straightforward free energy minimization, but for an RNA switch, the prediction of a less stable alternative structure is often computationally costly and even problematic. The current sampling-clustering method predicts stable and alternative structures by partitioning structures sampled from the energy landscape into two clusters, but it is very time-consuming. Instead, we predict the alternative structure of an RNA switch from conditional probability calculations within the energy landscape. First, our method excludes base pairs related to the most stable structure in the energy landscape. Then, it detects stable stems ("seeds") in the remaining landscape. Finally, it folds an alternative structure prediction around a seed. While having comparable riboswitch classification performance, the conditional-probability computations had fewer adjustable parameters, offered greater predictive flexibility, and were more than one thousand times faster than the sampling step alone in sampling-clustering predictions, the competing standard. Overall, the described approach helps traverse thermodynamically improbable energy landscapes to find biologically significant substructures and structures rapidly and effectively.

OPEN ACCESS

Citation: Manzourolajdad A, Spouge JL (2019) Structural prediction of RNA switches using conditional base-pair probabilities. PLoS ONE 14 (6): e0217625. https://doi.org/10.1371/journal.pone.0217625

Editor: Yaakov Koby Levy, Weizmann Institute of Science, ISRAEL

Received: December 10, 2018

Accepted: May 15, 2019

Published: June 12, 2019

Data Availability Statement: All relevant data are within the manuscript and Supporting Information files.

Funding: This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

In many organisms, structural rearrangements of RNA switches trigger biological functions. In bacteria, RNA switches regulate gene expression of mRNA downstream from them [1, 2]. In eukaryotes, they regulate alternative splicing [3]. In viruses, they can be critical in various stages of the viral life cycle, regulating rates of replication, transcription, RNA dimerization, etc. [4–9]. Plasticity, the ability to assume more than one structure, can also enhance the adaptability of RNA [10] by permitting it to accommodate distinct conformational phenotypes with only small perturbations to its genotype.

Current scientific strategies have produced an abundance of prokaryotic data, so most known RNA switches are in bacteria. To find putative riboswitches in prokaryotic sequences, common approaches rely on the assumption that the biologic mechanism of the riboswitch is at least partially conserved across certain bacteria. Hence, they locate conserved structural elements upstream of homologous coding regions [11, 12]. The Rfam database [13] provides a publicly available set of regulatory RNAs, including riboswitches. Although known riboswitch families are already structurally and functionally diverse [14–16] they probably represent only a small fraction of all bacterial riboswitches [17]. In any case, they provide invaluable examples of RNA molecules with stable alternative conformations.

After selective binding to a specific ligand, metabolite, or uncharged transfer RNA, the structure of a bacterial riboswitch toggles from unbound to the bound state. The conformational change then influences the expression of a proximal downstream gene. Most riboswitches contain two substructures: (1) an aptamer, which binds the ligand or metabolite, and which is usually structurally conserved; and (2) an expression platform, which undergoes allosteric rearrangement to regulate the gene. The thiamine pyrophosphate (TPP) riboswitch is a typical case illustrating regulation through a structural rearrangement. Fig 1 illustrates schematics of the bound and unbound conformations of the TPP riboswitch in *Bacillus subtilis* [18]. Fig 1A illustrates the bound state, where the aptamer in the RNA substructure surrounds TPP. In the bound state, the aptamer (sequence coordinates 0nt-120nt) stabilizes, causing the downstream expression platform (coordinates 120nt-190nt) to form a terminator stem loop (Fig 1A). In the unbound state without the TPP ligand, however, the anti-terminator substructure in the expression platform is stable, disrupting the terminator substructure (Fig 1B). The structure of the riboswitch *cis*-regulates downstream genes by influencing their expression.

In typical riboswitches, the unbound structure has a lower computed folding energy than the bound structure, because the ligand is required to stabilize the bound state. Furthermore, folding kinetics sometimes fine-tune binding. Typically (with some exceptions, e.g., [19, 20]), the bound state becomes energetically more favorable as the mRNA transcribes and elongates, so in the presence of an adequate ligand concentration the structure increasingly switches to the bound state. For instance, for the TPP riboswitch at a sequence elongation of 170nt, the minimum free energy (MFE) structure [21–23] corresponds to the unbound state; but at an elongation of 190nt, the MFE structure corresponds to the bound state (Fig 1C). Therefore, the relative stability of the bound and unbound states of some riboswitches can vary under different sequence elongations or segment selections. For simplicity and consistency, we always refer to the less stable structure in the computed energy landscape, whether bound or unbound, as *the alternative* structure.

Prediction of alternative structures generally requires extensive analysis of the secondary structure energy landscape. There are many strategies for alternative structure prediction (reviewed in [24, 25]). Structural entropy of the RNA folding landscape [26, 27], graph-based representations [28], Markov-state models [29], abstract shapes [30–34], energy-band-based sampling [35] base-pair-distance-based alternative structure prediction [36, 37], and the more recent probability-corrected sampling [38] are just a few examples of using various modeling techniques to explore the RNA folding landscape.

The Boltzmann ensemble thermodynamic model is the basis of most predictive strategies. The model conceptualizes the RNA structure of the sequence under scrutiny as a Boltzmann ensemble, estimating ensemble energies from experimentally derived nearest-neighbor free energy parameters [39–43]. In the Boltzmann ensemble, the probability $\Pr(S)$ of each structure S is proportional to $\exp[-E(S)/RT]$, where $E(S)$ is the free energy of S; R, the universal gas constant; and T, the folding temperature. Prediction of the alternative structure requires an efficient exploration of suboptimal (i.e., lower probability) structural configurations.

The Sampling-Clustering (SC) method is an important strategy for finding alternative structures [18, 44, 45]. First, suboptimal structures are sampled from the energy landscape, possibly under various temperatures to increase structural diversity. Then, the samples are

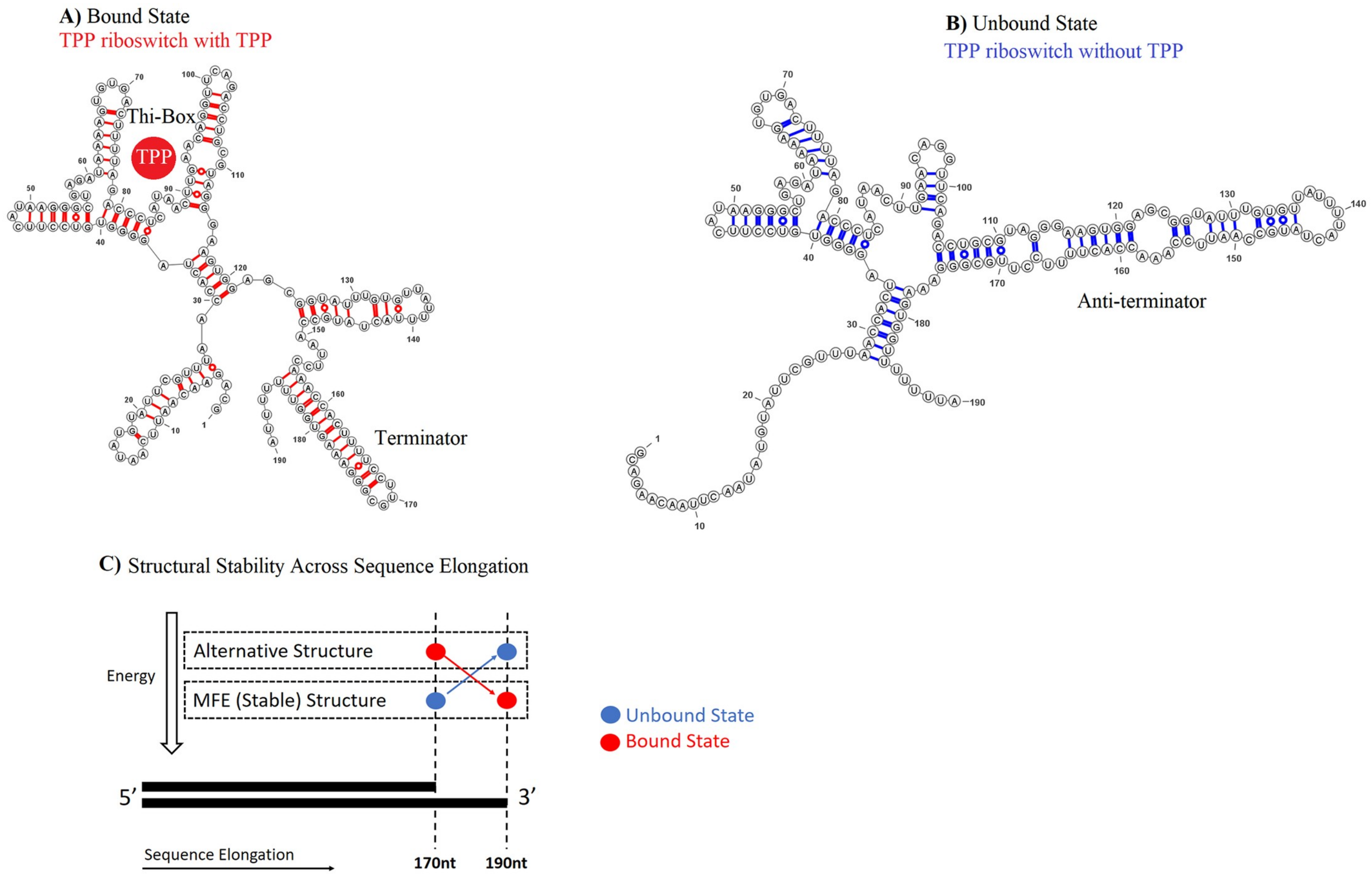

Fig 1. Two structural states of the thiamine pyrophosphate (TPP) riboswitch in *Bacillus subtilis*. [18]. A) Riboswitch when TPP is present. TPP ligand approximate location shown as solid circle. Thi-Box and terminator are formed. B) Riboswitch when TPP is absent. Anti-terminator is formed. The aptamer region is roughly between 1nt-120nt and the expression platform is roughly between 120nt-190nt (A and B). C) Folding energy of the TPP riboswitch. The unbound state is more stable, when the elongation is 170nt; the bound state, when the elongation increases to 190nt.

https://doi.org/10.1371/journal.pone.0217625.g001

partitioned into two clusters. The cluster containing the MFE structure corresponds to the stable structure while the other cluster corresponds to the alternative structure of the RNA under investigation. Although SC predicts some alternative structures accurately, it fails badly for others. Indeed, Boltzmann probabilities decrease exponentially with the free energy, with the potential to increase the SC computational runtimes exponentially [46]. For most riboswitches, the theoretical computational complexity poses no practical problem, however, because (1) the candidate riboswitch has only moderate length, or (2) exhaustive sampling is not required. For other riboswitches, however, the energy difference between the stable and the alternative structures (as much as 20 kcal/mol [37]) can severely slow the sampling. Furthermore, as the length *n* of the RNA sequence increases, the number of structures in the Boltzmann ensemble explodes exponentially, so traversing the ensemble explicitly becomes infeasible [47].

Instead of the time-consuming method of explicitly traversing structures of the ensemble, as SC does, in this work we infer the alternative structure from base-pair probabilities. Our Conditional-Probability (CP) method is as follows: First, the MFE structure is determined. Then, base-pair probabilities, conditioned on the absence of bonds between the MFE base pairs and their neighbors are calculated. McCaskill's algorithm [48–50] efficiently calculates the exact base-pair probabilities. We then select a stem whose base pairs have high conditional probabilities as a "seed". Finally, we determine the alternative structure by computing the structure with the lowest free energy of all structures containing the seed. UNAFold and Mfold [21, 50–53] implement the exact calculations, and in particular we use UNAFold (version 4.0.0) to calculate base-pair probabilities conditioned on the inclusion or exclusion of a given set of base pairs. Using an independent benchmark [54] consisting of riboswitches [18, 20, 24, 44, 55–71] (S1 Table) and the purine riboswitch family taken from Rfam [13], the Results section compares CP and SC predictions of alternative structures for both the prediction accuracy and computational times. Although CP prediction has very few adjustable parameters, the Results section also briefly describes the parameter space anecdotally in a single riboswitch to display the parameters' impact on CP alternative structure prediction. The Discussion then examines some implications of our study.

Materials and methods

Sampling-Clustering (SC) procedure for predicting the alternative structure. Given an RNA sequence, the energy landscape of the RNA is sampled at different temperatures starting from 300 structures at 37°C and 150 structures at each temperature value at six decile intervals towards the melting temperature of the RNA strand, totaling to 1200 samples per RNA. Sample numbers were selected according to same SC procedure used in [54] for comparison purposes. The samples are then partitioned into two clusters using k-means clustering, using base-pair Hamming distance d_H (the distance $d_H\{S',S''\}$ counts the base pairs that are in either of the structures S' and S'' but not in both). Denote the most energetically stable secondary structure configuration of a riboswitch by S_1; the alternative secondary structure, by S_2. As in [18], the computed MFE structure (denoted by S_1^*) then was used to predict the most stable structure S_1. The lowest-energy structure of the cluster not containing the MFE structure (denoted by $\hat{S}_2^*$, the over-hat often denoting a sampled quantity in statistics) was used to predict the alternative structure S_2.

Barsacchi Dataset consisting of 20 Riboswitches and 20 Non-coding (Non-riboswitch) RNAs. An extant benchmark, unchanged, provides sequences and structures for our evaluation of riboswitch classification [54]. Sequence lengths are identical to [54] for comparison purposes. We refer to the set of 20 riboswitches as 20-riboswitch set. The structures corresponding to twelve of those twenty riboswitches were provided by the authors and are used here for prediction accuracy calculations. We refer to this subset of sequences as the Barsacchi Structural dataset set (See S1 Table for names, lengths and number of actual structures available for each of the 20 riboswitches).

The Purine Riboswitch dataset consisted of the purine riboswitch family. First, the structurally conserved aptamers of purine riboswitch family RF00167 were downloaded from Rfam (133 sequences roughly 100-nt long), along with the consensus secondary structure which represented the bound state of the aptamer. Then, each aptamer sequence was extended 100nt downstream to capture the expression platform of the riboswitch as well. We mapped the consensus structure of the aptamer to the beginning of individual sequences.

Structure visualization. Structures in all figures were done with the aid of the VARNA software [72].

Conditional-Probability prediction

We first computed the MFE secondary structure S_1^* to predict S_1. We then constructed an alternative structure S_2^* to predict S_2, as follows: Let $[i{\cdot}j]$ denote a base pair between sequence positions i and j. Given S_1^*, a base-pair-to-structure distance, and a dissimilarity threshold τ, we first build an excluded set $\mathcal{E}^{(\tau)}$ consisting of base pairs close to base pairs $[i \cdot j] \in S_1^*$. Let us first describe base-pair-to-structure distance δ_{bs}:

Base-pair distance. Our conditional probability predictions require a base-pair distance $\delta(i{\cdot}j, i'{\cdot}j')$ between $[i{\cdot}j]$ and $[i'{\cdot}j']$. Base-pair distance δ_{bb} as defined in [22], is

$$\delta_{\text{bb}}(i \cdot j, i' \cdot j') = \max\{|i - i'|, |j - j'|\}. \tag{1}$$

Given a set of base pairs in a structure S', [73] define the base-pair-to-structure distance

$$\delta_{\text{bs}}(i \cdot j, S') = \min_{i' \cdot j' \in S'} \delta_{\text{bb}}(i \cdot j, i' \cdot j'). \tag{2}$$

Now, given a non-negative integer dissimilarity threshold τ, consider the excluded set $\mathcal{E}^{(\tau)}$. All the base pairs in $\mathcal{E}^{(\tau)}$ are selected such that their δ_{bs}-distance to S^* is less than or equal to τ:

$$\mathcal{E}^{(\tau)} = \{[i \cdot j] : \delta_{\text{bs}}(i \cdot j, S_1^*) \leq \tau\}, \tag{3}$$

McCaskill algorithm. Our methods make heavy use of the McCaskill algorithm [48] for calculating various probabilities related to the existence of a given base pair $[i{\cdot}j]$ in the structure of an RNA sequence at thermal equilibrium. Given the RNA sequence, the Boltzmann probability of its being in the structure S at thermal equilibrium is $\Pr(S) = e^{-E(S)/RT}/Z$, where the partition function $Z = \sum e^{-E(S)/RT}$ is a normalizing factor. The McCaskill algorithm efficiently computes the exact probability $P_{i,j}$ that S contains the base pair $[i{\cdot}j]$ through recursive calculations. The calculations combine various partition functions, given here in McCaskill's notation. The restricted partition function $Q_{i,j}^b$ for bases from i to j inclusive is the sum of the Boltzmann weights $e^{-E(S)/RT}$ over the structures $S_{i,j}$ where the base pair $[i{\cdot}j]$ closes $S_{i,j}$ into a loop structure. The full partition function $Q_{i,j}Q_{i,j}$ for bases from i to j inclusive is the sum of $e^{-E(S)/RT}$ over the corresponding structures $S_{i,j}$ with and without the base pair $[i{\cdot}j]$, so for example the full partition function is $Z = Q_{1,n}$. The probabilities $P_{i,j}$ of individual base pairs $[i{\cdot}j]$ within the ensemble of structures S can be computed as $P_{i,j} = \sum_{S \ni [i \cdot j]} e^{-E(S)/RT}/Z$ where the sum is over structures S containing $[i{\cdot}j]$. McCaskill's recursive calculations follow combinatorial patterns resembling algorithms for determining the MFE structure, computing first $Q_{i,j}^b$, then $Q_{i,j}$, and finally $P_{i,j}$ from previously computed values for smaller substructures. Due to complexity of McCaskill's algorithm, we must refer the reader to [48] for details.

Conditional base-pair probabilities. Given a set of excluded base pairs $\mathcal{E}$ (e.g., $\mathcal{E}^{(\tau)}$ above), the McCaskill algorithm can also compute the probability that a given base pair $[i{\cdot}j]$ exists, under the constraint that none of the base pairs in $\mathcal{E}$ are allowed. We denote the probability as

$$P_{i,j}^{\sim\mathcal{E}} \equiv \Pr\{[i \cdot j] \in S | [i' \cdot j'] \notin S, \forall [i' \cdot j'] \in \mathcal{E}\}. \tag{4}$$

(In set theory, "~" often denotes complementation, so it helps suggest exclusion of $\mathcal{E}$ in the superscript of $P_{i,j}^{\sim\mathcal{E}}$ in Eq (4)). Like calculating the unconstrained base-pair probabilities $P_{i,j}$ above, the McCaskill algorithm can calculate the conditional base-pair probabilities in Eq (4) by using constrained partition functions. Our general plan is to constrain the alternative structure by avoiding all the base pairs in a neighborhood $\mathcal{E}^{(\tau)}$ of the MFE structure, then to identify base pairs that the constrained structure is likely to contain, and then to fold the alternative structure around those base pairs.

The UNAfold (version 4.0.0) Software Package [50] calculates both the unconditional probabilities $P_{i,j}$ and the conditional probabilities $P_{i,j}^{\sim\mathcal{E}}$ efficiently with the McCaskill algorithm [48].

Given an excluded set $\mathcal{E}$ (e.g., $\mathcal{E}^{(\tau)}$ above), next select a seed substructure L^*, a longest stem whose every base pair *individually* has conditional probability higher than 0.5. Let L denote any stem (i.e., any ladder), where we permit L to contain bulges (i.e., single unpaired bases). Let #L count its base pairs.

$$L^* = \arg\max_L\{\# \ L|P_{i,j}^{\sim\mathcal{E}} > 0.5 \text{ for every } [i \cdot j] \in L\}$$

$$S_2^* = \arg\max_S \Pr[S| \ L^* \in S]$$

Finally, S_2^* is predicted as the lowest-energy structure that contains L^*. Structure S_2^* of the given RNA sequence serves to predict S_2.

Because the seed lies outside $\mathcal{E}$, by definition it always lies outside S_1^*, so $S_2^* \neq S_1^*$. Note that S_2^* may still incidentally contain some originally excluded base pairs from $\mathcal{E}$ (or S_1^*, for that matter): the excluded base pairs in $\mathcal{E}$ are only used to compute the conditional probability distribution for selecting the seed. Once a seed is selected, it is the *only* constraint in energy minimization of S_2^* (i.e., S_2^* must contain it). Default calculations for seed-based predictions were performed at temperature 37˚C and favored the version 3.0 energies of UNAfold, the most current energy parameters for RNA folding [40].

Generalization to other choices of seeds. We explored several criteria for seed selection. Empirically, the stem seed L^* was most promising. Other seeds included the single base pair with highest $P_{i,j}^{\sim\mathcal{E}}$ value and the union of all base pairs with $P_{i,j}^{\sim\mathcal{E}} > 0.5$.

Generalization to other temperatures. The most recent version of UNAfold (version 4.0.0) can make energy calculations at temperatures $T = T_0$ other than 37˚C, permitting the following generalized procedure. Predict structure S_1^* at 37˚C and construct the excluded set $\mathcal{E} = \mathcal{E}^{(\tau)}$, as above. Calculate all conditional base-pair probabilities $\{P_{i,j}^{\sim\mathcal{E}}\}$ and stem seeds L^* at various temperatures $T = T_0$, to predict the alternative structure $S_2^{*T_0}$.

Generalization by iteration number. The procedure above could be considered the first step of an iterative process, where subsequent predictions S_i^* (i = 3,4. . .) can be made by excluding the union $\mathcal{E} = \mathcal{E}_i = \mathcal{E}_{i-1} \cup \mathcal{E}_{i-2}$ of base pairs in previous predictions, and then identifying a seed L_i^* from the resulting conditional probability distribution $\{P_{i,j}^{\sim\mathcal{E}}\}$.

Notations. The alternative structure prediction S_2^* is associated with adjustable parameters: threshold $\tau = \tau_0$, temperature $T = T_0$, iteration i, and the type of seed. It is denoted in full by *type-of-seed* $S_i^{*T_0}(\tau = \tau_0)$. For example, the alternative structure predicted with iteration 2 and using τ = 15, at temperature 70˚C, and using a single base pair as the seed is denoted as *single-base-pair-seed* prediction $S_2^{*70^\circ\text{C}}(\tau = 15)$. Our notation usually drops the default values, i.e. $S_2^* \equiv S_2^{*37^\circ\text{C}}(\tau = 5)$ using stem seed L^*. Unless otherwise stated, results used default parameters.

Normalized seed length. The number of base pairs within the seed was divided by the log of the sequence length SL, (#L^*)/$\log_{10}$(SL), to derive Normalized Seed Length.

Results

The Results section focuses on predicting the alternative structure. Here, the most energetically stable structure prediction is the MFE structure S_1^*, so its precise quality depends on the thermodynamic model. Furthermore, both the CP and SC methods use the MFE structure to predict the most stable structure. Therefore, the Results section omits comment on the stable structure prediction.

On the other hand, alternative structure prediction is sensitive to sequence elongation (i.e., the exact length chosen as the sequence under scrutiny). To avoid the possibility of cherry picking sequence lengths, our study took an independent "Barsacchi dataset" [54] (without change) as our dataset of 20 riboswitches and 20 non-riboswitches. The Barsacchi dataset provides a gold standard for RNA sequence and actual structures (if actual structures were available). The next subsection introduces the notations and concepts for the CP prediction using the TPP *tenA* riboswitch in *Bacillus subtilis* [18, 66] (*tenA* TPP) as a specific example.

The CP prediction method applied to the *tenA* TPP riboswitch

The CP prediction method uses the computed MFE structure S_1^* to predict the actual stable structure S_1 (a standard prediction method, e.g., [54]); and prediction S_2^* (as described in the Materials and methods section) to predict the actual alternative structure S_2. The predicted alternative structure S_2^* always has a higher computed energy than the predicted MFE structure S_1^*.

To predict S_2^*, we first determine a set of base pairs that are "close to" base pairs in S_1^*, namely all base pairs having distance less than or equal to dissimilarity threshold τ, $\delta_{\mathrm{bs}}(i \cdot j, S_1^*) \leq \tau$ (see the Materials and methods section for precise mathematical definitions). Increasing τ therefore excludes progressively more base pairs from the alternative structure S_2^*. Our default dissimilarity measure is $\tau = 5$. Let $\mathcal{E}^{(\tau)}$ be the set of excluded base pairs. Conditioned on the absence of base pairs in $\mathcal{E}^{(\tau)}$, the McCaskill algorithm yields the conditional probability of any other base pair. The Materials and Methods section defines a longest bulge-containing stem, L^*, where every base pair has a conditional probability higher than 0.5, e.g., for *tenA* TPP, L^* contained 7 base pairs. Finally, using the stem L^* as a seed, constrained energy minimization yields S_2^*, which is the minimum energy structure containing L^*.

Empirical results motivated our defaults for L^* and τ. Our choice of τ provides some insight into our methods. Briefly, for $\tau = 0$ the excluded set $\mathcal{E}^{(\tau=0)} = \{[i \cdot j] : [i \cdot j] \in S^*\}$ is precisely the set of base pairs in S_1^*. Empirically, $\mathcal{E}^{(\tau=0)}$ often excluded too few base pairs to make $S_2^*(\tau = 0)$ markedly different from S_1^*. We selected $S_2^* \equiv S_2^*(\tau = 5)$ as our default for alternative structure prediction. Fig 2 illustrates the actual alternative structure (here, unbound structure) of *tenA* TPP and its corresponding predictions. Fig 2A illustrates the actual alternative structure S_2 with its aptamer region and anti-terminator stem. Fig 2B illustrates S_2^*, with the stem seed L^* shown in red.

The established approach for predicting the two structural states of a riboswitch uses a Sampling-Clustering (SC) method. Every SC computation in this article followed standard protocol by sampling RNA structures at many different temperatures. Fig 2B illustrates alternative structure prediction under SC (see the Materials and methods section for details of our SC implementation). As we can see, for this elongation of TPP riboswitch, the CP method recognizes critical substructures of the alternative structure better than the SC method.

Fig 3 displays the energy landscape for the *tenA* TPP example. Fig 3A displays a standard depiction of an energy landscape, where the X-axis gives base-pair Hamming distance to S_1^* for each sampled structure S. Because our main interest is structural prediction of S_2, Fig 3B depicts our Alternative-Structure-Referenced (ASR) energy landscape, where the X-axis gives base-pair Hamming distance of each sampled structure S to S_2. Structural dissimilarities of predictions to S_2 are better illustrated in the ASR energy landscape than in the standard energy landscape.

Conditional-Probability gave base-pair prediction performance superior to Sampling-Clustering

To assess predictive accuracy of CP and SC, a subset of the Barsacchi dataset provided 12 riboswitches with known actual alternative structures S_2 (our "Barsacchi Structural dataset").

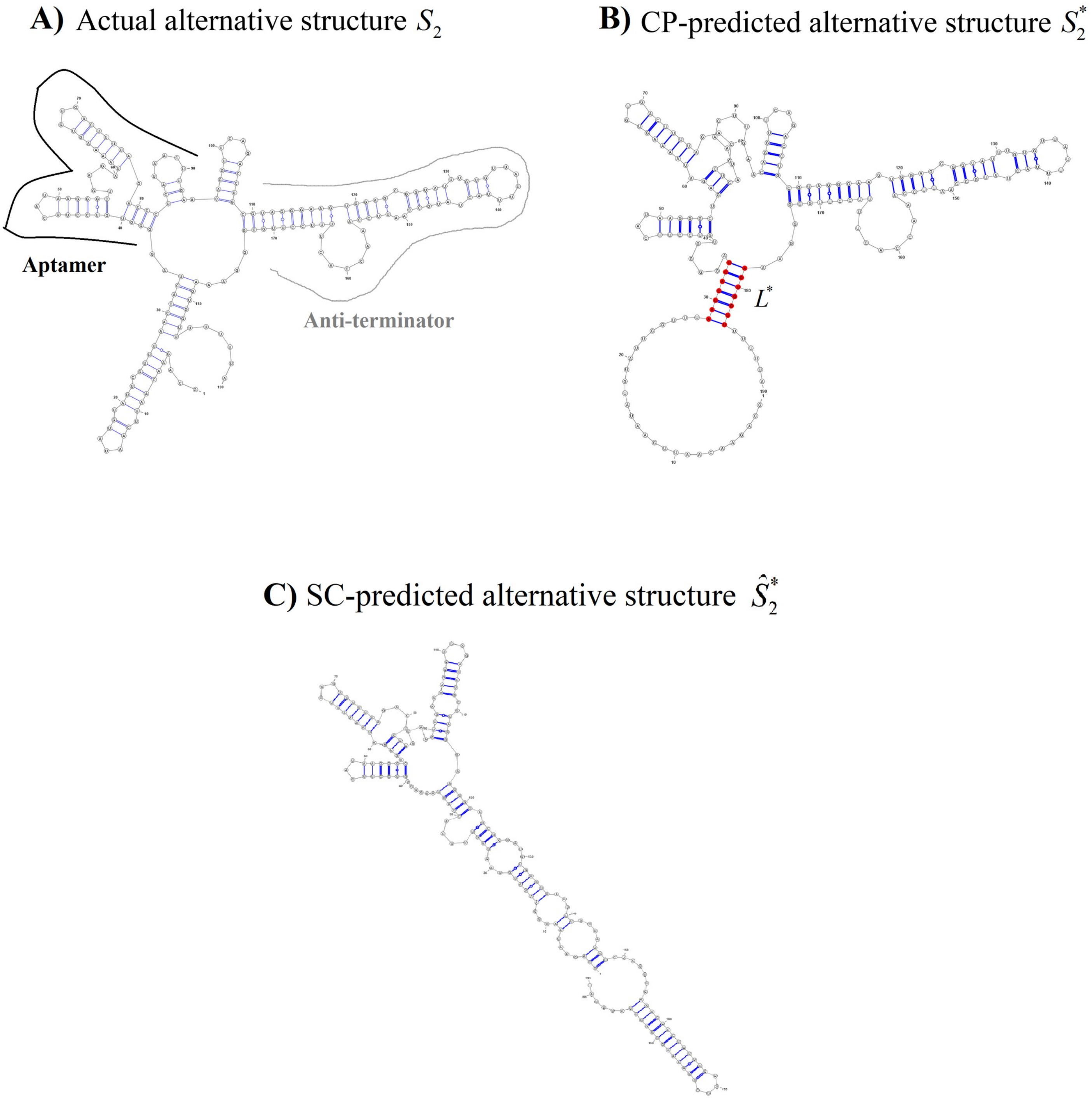
A) Actual alternative structure S_2
Aptamer
Anti-terminator
B) CP-predicted alternative structure S_2^*
L^*
C) SC-predicted alternative structure $\hat{S}_2^*$

Fig 2. Prediction of the alternative structure of *tenA* TPP riboswitch. Sequence length in the Barsacchi Dataset was 190nt. A) The actual alternative structure S_2 (here, the unbound state), with the approximate locations of the aptamer substructure and the anti-terminator stem as indicated. B) The CP-predicted alternative structure S_2^*. The stem seed $L^* = \{28...34 \cdot 178...184\}$ is shown in red. C) The SC-predicted alternative structure $\hat{S}_2^*$ (See Materials and methods for details).

https://doi.org/10.1371/journal.pone.0217625.g002

Within the Barsacchi Structural dataset, we compared S_2, our gold standard, to three predicted alternative structures: $\hat{S}_2^*$ (from SC), $S_2^*(\tau = 0)$ (from CP, with $\tau = 0$), and S_2^* (from CP, with default $\tau = 5$). For the three predicted alternative structures $\hat{S}_2^*$, $S_2^*(\tau = 0)$, and S_2^*, we calculated their sensitivity and specificity values (i.e., total-SEN and total-PPV in [74]), along with the F

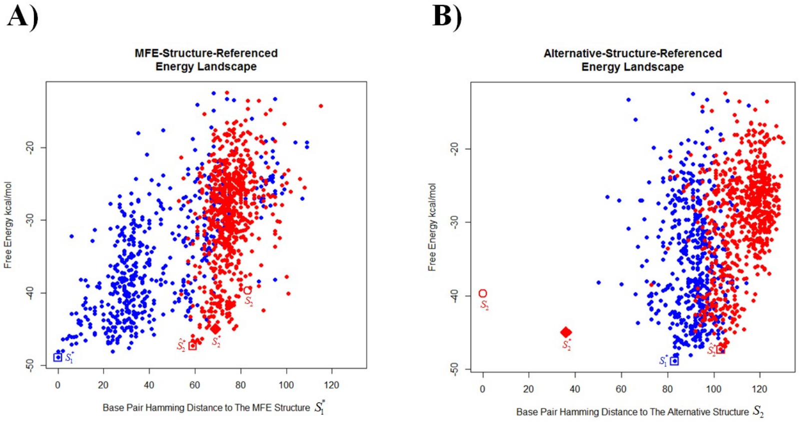

Fig 3. Energy landscape of *tenA* TPP. The open blue square represents the computed MFE structure S_1^*; the open red circle, the actual alternative structure S_2; the solid red diamond, the CP-predicted alternative structure S_2^*; and the open red square, the SC-predicted alternative structure $\hat{S}_2^*$. Blue and red dots represent samples of the MFE-containing and alternative clusters, respectively (see the Materials and methods section for details on our SC implementation). A) A standard depiction of an energy landscape, where the MFE structure S_1^* provides the reference structure at distance 0. B) The Alternative-Structure-Referenced (ASR) energy landscape, where the actual alternative structure S_2 provides the reference structure at distance 0. Some important base-pair Hamming distances can be read directly from Fig 2A and 2B: $d_H\{S_2,S_1^*\} = 83$ (from both 2A and 2B); $d_H\{\hat{S}_2^*,S_1^*\} = 59$ and $d_H\{S_2^*,S_1^*\} = 69$ (from 2A). $d_H\{\hat{S}_2^*,S_2\} = 103$ and $d_H\{S_2^*,S_2\} = 36$ (from 2B).

https://doi.org/10.1371/journal.pone.0217625.g003

measure, another measure of prediction accuracy [75]. Table 1 gives the overall performance of the three predictions (S2 Table gives the details of performance for each riboswitch). The total-SEN, total-PPV, and F measure ordered the performance of the three predictions consistently: $S_2^*(\tau = 0)$ was slightly inferior to $\hat{S}_2^*$, but S_2^* was noticeably superior to them both.

Conditional-Probability was about 1000 times faster than Sampling-Clustering

In computational speed on the Barsacchi Structural dataset, CP was more than three orders of magnitude faster than SC (see Table 1). Computations were parallelized, so the longest sequence in the dataset, here the *LysC* Lysine riboswitch (243nt), determined the speeds. Typical CP runtimes increased with increasing dissimilarity thresholds, but even a dissimilarity threshold of half the length of the sequence, $\tau = 122$, only increased the CP runtime to 6 seconds. Our SC runtimes included the sampling step, but not the clustering step. The whole CP computation of S_2^* (i.e., under default value $\tau = 5$) took around a second, whereas the sampling step of SC alone took more than 18 mins.

Table 1. Alternative-structure prediction performance in the Barsacchi Structural dataset.

Predicting the alternative structure via	total-SEN (%)	total-PPV (%)	F measure	Time
SC	43.0	37.4	0.100	00:18:23^{x}
CP under zero threshold	36.2	31.1	0.084	00:00:01
CP under the default threshold	51.3	46.1	0.121	00:00:01

The text explains the performance measures. The (parallelized) computational time reflects the longest sequence in the dataset, the *LysC* Lysine riboswitch (243nt). The SC time (marked x) does not include the clustering computation. All computations were performed on a 12 core Intel(R) Xeon(R) 2.93GHz CPU. The performance of SC $\hat{S}_2^*$, CP under zero threshold $S_2^*(\tau = 0)$, and CP under the default threshold S_2^* are shown below.

https://doi.org/10.1371/journal.pone.0217625.t001

Riboswitch classification

To compare how well CP and SC can classify RNA sequences into riboswitch or non-riboswitch, we used the full Barsacchi Dataset (20 riboswitches vs. 20 non-riboswitches). SC classified each sequence according to the average silhouette value of the two clusters of samples; CP, according to normalized seed length (see the Materials and methods section). Compared to non-riboswitches, riboswitches tend to have both a higher average silhouette value [54] and normalized seed length (4.00bp for riboswitches; 3.50bp for non-riboswitches). Fig 4 displays the corresponding ROC curves for SC (in blue) and CP (in red).

We tuned the SC computations more extensively in riboswitch classification than in other parts of our study, exploring multiple temperatures or 37˚C alone to generate the samples, since performance of SC can highly depend on sampling parameters. To make our comparison to the best SC performance, we included Vienna RNA results, which had been previously used on Barsacchi Dataset. SC classification performance was better at 37˚C alone (area under curve 0.5825, solid blue in Fig 4) than at multiple temperatures (area under curve 0.455) under the UNAFold software. Area under curve was higher using Vienna RNA at 37˚C alone (0.676, dashed blue in Fig 4). In our hands and on the Barsacchi Dataset, the performance of the CP-based classifier was comparable to SC under Vienna RNA sampling when differentiating riboswitches from non-riboswitches (0.6385 for CP compared to 0.676 for SC, difference 0.0375±0.1227 with two-tailed p-value 0.76 [76]), and higher compared to SC under UNAFold (0.5825).

The seeds L^* pointed to critical substructures in the purine riboswitch family

The Materials and Methods section details the construction of our Purine Riboswitch dataset from Rfam. The Purine Riboswitch dataset contained 133 sequences, each of length about 200nt, constructed with the intent that the sequences were long enough to fold into both unbound and purine-bound structures. Rfam provides a consensus structure for the ligand-bound aptamer, which we mapped onto each of the 133 sequences. Stem P1 of the ligand-bound aptamer (Fig 5C) is fully formed only if the aptamer has bound a purine nucleotide, and not otherwise [77]. As above, we predicted S_1^*, S_2^*, and $\hat{S}_2^*$ for each sequence in the dataset and examined whether the predicted structure contained Stem P1, a substructure of interest in the bound state of a purine riboswitch. (Here, the comparison of the CP and SC methods must consider the computed MFE structure S_1^*, because it can correspond to either of the bound and unbound structures.) To describe Fig 5A, of 133 sequences, S_1^* (predicted by both CP and SC) contained Stem P1 for 50 sequences. Of the 50 sequences where S_1^* contained Stem P1, S_2^* (predicted by CP) contained Stem P1 for 7 sequences; and $\hat{S}_2^*$ (predicted by SC), for 21 sequences. Of the remaining 83 = 133–50 sequences where S_1^* did not contain Stem P1, S_2^* (predicted by CP) contained Stem P1 for 24 sequences, and $\hat{S}_2^*$ (predicted by SC) contained Stem P1 for 17 sequences.

For the 24 sequences where CP predicted Stem P1 in S_2^* but not in S_1^*, we investigated the location of the stem seed L^*. Interestingly, of the 24 sequences, L^* was almost identical to Stem P1 in 12 sequences, and in 10 of the remaining sequences, L^* was distant from the aptamer and near the terminator stem. In those cases, the CP stem seed L^* contained many consecutive base pairs overlapping the actual terminator stem (see Fig 5B). In the remaining 2 cases, L^* had no clear biological interpretation. In summary, therefore, in 22/24 purine riboswitches where Stem P1 of the aptamer was in S_2^* but not S_1^*, the stem seed L^* pointed to a regulatory substructure in the purine riboswitch.

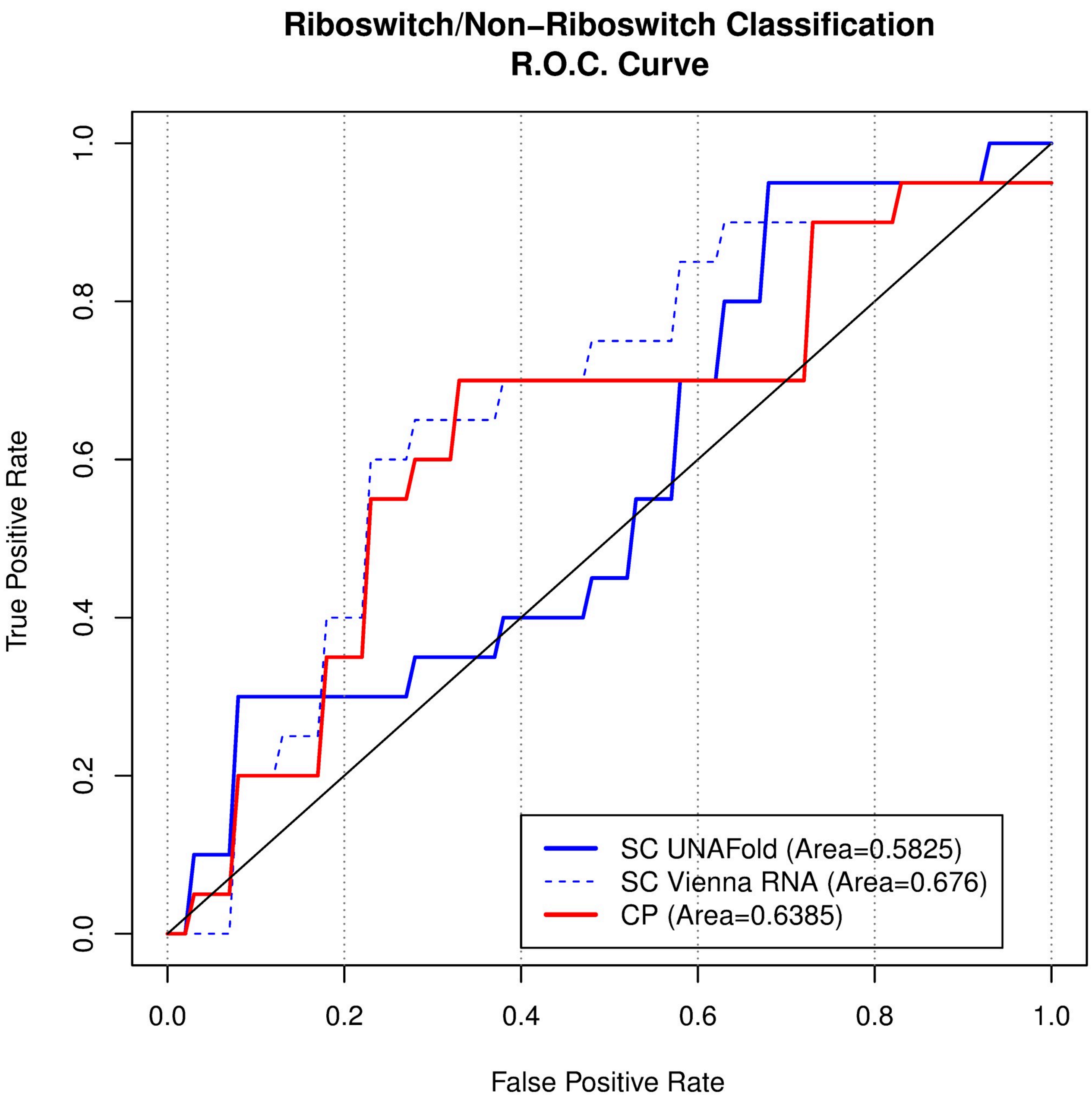

Fig 4. ROC. curves for SC- and CP-based riboswitch classifiers. The blue and red lines correspond to SC- and CP-based classifiers. The legend gives the ROC value, the area under the curves for the two classification methods. Calculations were done at 37˚C for all three classifiers.

https://doi.org/10.1371/journal.pone.0217625.g004

A case study: How adjustable parameters in CP affect prediction accuracy

The generalized CP prediction S_2^* of alternative structure has only a few adjustable parameters: dissimilarity threshold τ (default: 5), temperature T (default: 37˚C), *type-of-seed* (default: stem seed L^*), and iteration number i (default: 2). The Materials and Methods section describes the parameters in detail. Our default parameters (particularly for τ and L^*) were tuned empirically by their predictions on the Barsacchi Structural dataset, but the sparsity of the data prevented methodical optimization (and any formal separation of training and test datasets). The

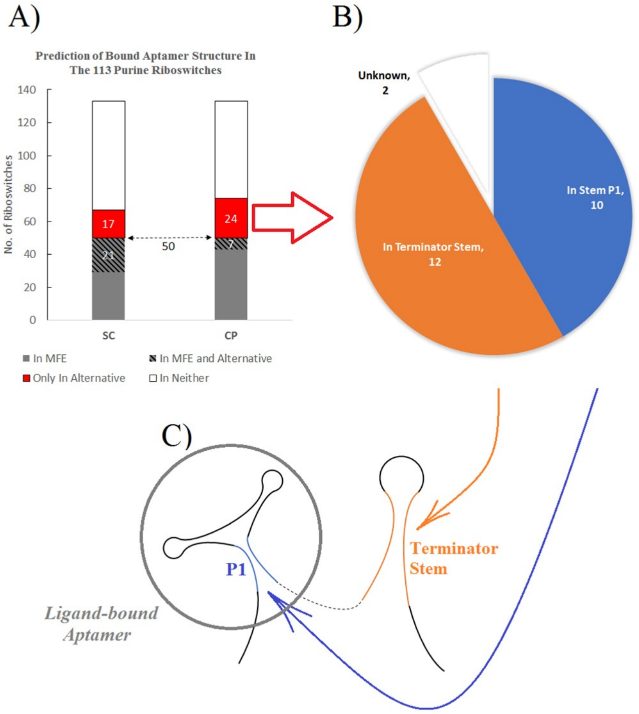

Fig 5. Alternative structure prediction in the purine riboswitch dataset. We examined the predicted structures S_1^*, S_2^*, and $\hat{S}_2^*$ for Stem P1, a substructure of the bound aptamer. A) The number of predicted structures containing Stem P1. The left bar corresponds to SC (which predicts S_1^* and $\hat{S}_2^*$); the right bar, to CP (which predicts S_1^* and S_2^*). B) The location of the stem seed L^* for the 24 sequences where CP predicted the bound aptamer in S_2^* but not S_1^*. C) The schematic structure of a typical ligand-bound purine riboswitch.

https://doi.org/10.1371/journal.pone.0217625.g005

following uses the *xpt* Guanine riboswitch [2, 57], 162nt long in the Barsacchi Structural dataset, as an anecdotal example of how S_2^* varies with the parameters.

In Fig 6, ASR energy landscapes display the actual unbound structure S_2 of *xpt* Guanine on their left, as a standard for comparing other structures. For example, Fig 6A displays structural samples from the Boltzmann distribution, along with the predicted alternative structures S_2^* (CP) and $\hat{S}_2^*$ (SC). Fig 6B–6D vary τ, T, the *type-of-seed*, or the number of iterations, showing the effects of changing parameters on CP predictions of the alternative structure. Fig 6B shows that the 61 dissimilarity thresholds $\tau = 0...60$ produced only five distinct predictions S_2^*. Fig 6C shows that the 12 temperatures $T = 37,40,43,46,49,52,55,58,61,64,67,70$ (from 37˚C to an approximation of the melting temperature 69.8˚C in 3˚C increments) produced only four distinct predictions S_2^*. We also could have folded S_2^* around a seed structure other than a stem. Fig 6D shows different predictions resulting from: (1) various seeds, e.g., base pairs with conditional probabilities higher than 0.5 and 0.8, and stems of various fixed lengths; and (2) various numbers of iterations of the exclusion-conditional probability calculation (e.g., predictions S_3^*, S_4^*, and S_5^* from 3, 4, and 5 iterations using a stem seed: see the Materials and methods subsection, Generalization by iteration number). Overall, base-pair Hamming distance $d_H\{S_2^*, S_2\}$ from S_2^* to S_2 took very few different values as the individual parameters varied (i.e., relatively few distinct predicted structures S_2^* appeared as the parameters varied).

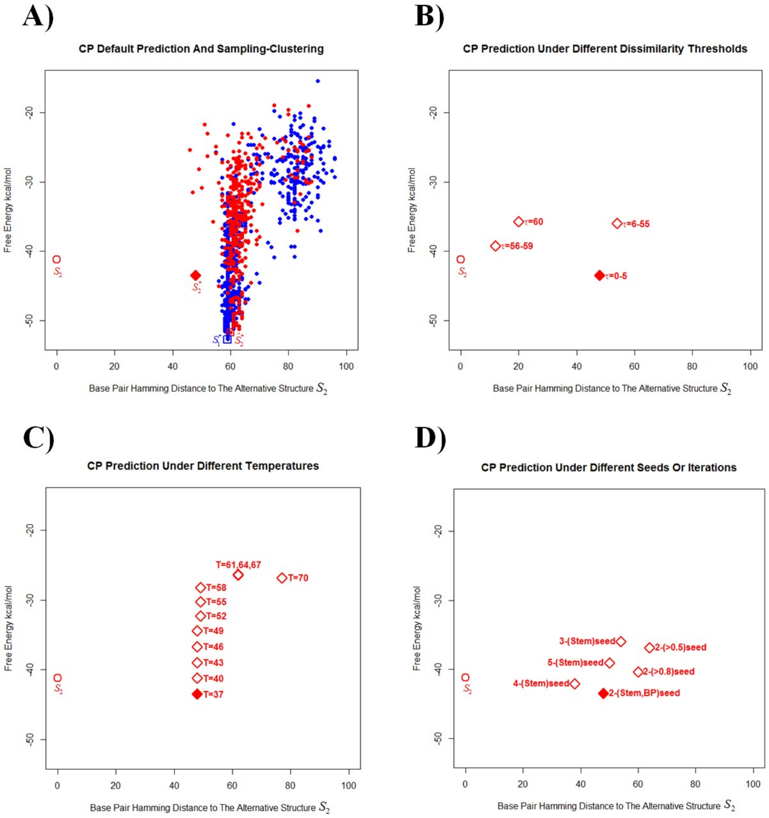

Fig 6. Alternative-Structure-Referenced (ASR) energy landscape of *xpt* Guanine riboswitch. The sequence length was 162nt, directly from the Barsacchi dataset. In each subfigure, the open red circle on the left represents S_2. In Fig 5B–5D, the solid diamond represents CP prediction under default parameters; and the empty diamonds, CP prediction under non-default parameters. A) The solid red diamond represents the CP-predicted alternative structure S_2^*; and the open red square, the SC-predicted alternative structure $\hat{S}_2^*$. Blue and red dots represent samples of the MFE-containing and alternative clusters, respectively (see Materials and methods for details on the Sampling-Clustering procedure). B) CP-predicted alternative structures S_2^* under different dissimilarity thresholds $\tau = 0. . .60$. If predictions were identical, they are shown with a label indicating a range for τ. C) CP-predicted alternative structures S_2^* under different temperatures T from 37˚C to approximate melting temperature 70˚C. The free energy of any fixed structure varies with temperature, and the underlying RNA structure remained fixed for temperatures $37 \leq T \leq 49$, $50 \leq T \leq 58$, and $61 \leq T \leq 67$. D) CP-predicted alternative structures S_2^* under different seeds and iterations. The initial number in the label of each point gives the iteration $i = 2. . .5$. The seed was either a stem seed L^*, a base-pair seed (the base pair with highest conditional probability, denoted by "BP"), or all base pairs with conditional probability higher than a threshold (either 0.5 or 0.8). For iteration number $i = 2$, the CP-predicted S_2^* from stem and base-pair seeds were identical: hence, its label "2-(stem,BP)-seed".

https://doi.org/10.1371/journal.pone.0217625.g006

Discussion

Computational identification of RNA switches and substructures solely from sequence could elucidate biological control mechanisms in many species, and successful identification of switches could accelerate the discovery of new sensors and control mechanisms, particularly in bacteria and other prokaryotes. For an RNA switch (or more specifically, for a typical riboswitch), predicting an alternative biologically functional structure can be challenging. Such predictions hold the key to understanding many regulatory mechanisms, however. Although riboswitches are a very diverse subset of RNA switches, there is not enough experimentally verified data for optimizing a set of parameters of a universal RNA Switch Predictor.

Conditional-Probability and Sampling-Clustering produced comparable predictions in the Barsacchi dataset [54]

We used the Barsacchi dataset [54] as a benchmark mainly to have independently selected sequence lengths for our study. In its first step, the CP prediction of alternative structure excludes from the energy landscape those base pairs belonging to a metastable state, here, the MFE structure. Then, it explores the remaining energy landscape for stable substructures. The prediction has relatively few adjustable parameters, so we gave them a preliminary default tuning with a structural subset of the Barsacchi dataset.

We predicted the CP alternative structure using only the Boltzmann ensemble at 37˚C. In our hands, compared to a implementation of the current SC procedure [18, 44, 54], which included sampling from many temperatures, CP predicted alternative structures better than SC (Table 1). Thus, although the sparsity of relevant data required us to use the Barsacchi dataset [54] for both training and testing, Table 1 suggests (at the very least) that Conditional-Probability (CP) predictions can be competitive with current Sampling-Clustering (SC) methods for predicting the alternative structure. In fact, Table 1 gives very conservative performance values for CP, because only the bound structure was available for three of the twelve riboswitches (see S2 Table). For these three riboswitches, the MFE prediction S_1^* corresponded to the actual bound state S_2, so the CP-predicted alternative state S_2^* was different, lowering CP prediction performance. Nonetheless, prediction performance of CP under the default dissimilarity threshold was still higher than that of SC.

Conditional-Probability vs Sampling-Clustering produced comparable predictions in the purine riboswitch dataset

To avoid cherry-picking in our Purine Riboswitch dataset, a purine riboswitch family of 133 Rfam sequences, we fixed all sequence lengths at 200nt. Fig 5A again suggests that CP predictions are also competitive with SC predictions in the Purine Riboswitch dataset. In fact, CP confined predictions of the Stem P1 in the bound aptamer substructure more to the alternative structure than SC (cf., 7/50 under CP to 21/50 under SC). This relatively exclusive prediction is probably desirable in the alternative structure predictor, because the ligand-bound aptamer is exclusive to only one of the two functional structures of a purine riboswitch. Overall, however, the CP default parameter predictions made at a single temperature of 37˚C again had comparable performance to an SC prediction at multiple temperatures.

Conditional-Probability was more than 1000 times faster than Sampling-Clustering

The computed free energy of the bound structure of a riboswitch does not include ligand binding energies, and consequently it may be arbitrarily high. Regardless, CP prediction always has a polynomial time-complexity. Denote sequence length by n and the dissimilarity threshold by τ. CP prediction has five steps (each step is followed by its time-complexity): MFE prediction ($O(n^4)$); excluded set determination ($\tau^2 O(n)$); conditional probability calculation ($O(n^3)$); seed selection ($O(n^2)$); and prediction of the alternative structure by constrained energy minimization ($O(n^4)$). Thus, the total computational time $t_{cp}(n,\tau) = \tau^2 O(n)+O(n^4)$. In contrast, SC prediction can be sensitive to the computed free energy of the bound structure, because folding probabilities exponentially decrease with energy. The time-complexity of exhaustive sampling is exponential in both the sequence length n and the maximum energy explored. The complexity becomes more manageable under statistical sampling, but then it depends on the number of samples (s) as well as n. A typical SC procedure such as ours consists of three steps: MFE prediction ($O(n^4)$); sampling ($sO(n^2)$); and clustering ($s^2O(n)$). Thus, the total computational time $t_{sc}(n,s) = O(n^4)+sO(n^2)+s^2O(n)$. Thus, the time complexities of the two procedures have a term $O(n^4)$. The SC prediction requires more sampling as the sequence lengthens, burdening computer memory and increasing the computation times accordingly. CP predictions, on the other hand, require more memory as τ increases. We generally expect CP predictions to be much faster than SC even for RNA sequences of moderate length (e.g., n equaling a few hundred).

Typically, riboswitch lengths are moderate, so the sample sizes s necessary for effective prediction remain feasible. To compare the speed of CP and SC predictions, Table 1 shows that despite its far superior speed, even when t_{sc} excludes the time for clustering the samples, CP can still outperform SC in predictive accuracy (e.g., for *LysC* Lysine, $t_{cp}(n = 243,\tau = 5)$ equals about 1 second; $t_{sc}(n = 243,s = 1200)$, more than 18 minutes). The time for CP alternative structure prediction does increase as τ increases from its default value $\tau = 5$, but it still remains far faster than SC prediction (e.g., $t_{cp}(n = 243,\tau = 122)$ equals only 6 seconds). To summarize, in our hands CP predictions were about 1000 times faster than SC predictions, with at least comparable prediction accuracy.

The effect of the dissimilarity threshold τ on CP predictions

CP generates a seed around which we fold an alternative structure. Loosely, when generating the CP seed, the dissimilarity threshold τ effectively excludes the formation of base pairs "close to" the MFE structure. It therefore enforces topological differences between alternative and MFE structures (see S1 File, which relates τ to the Relaxed Base Pair score $\rho_t(S_1,S_2)$ between structures [73]). The threshold $\tau = 0$ excludes all base pairs in the MFE structure, and as τ increases, the exclusion procedure progressively excludes more and more base pairs. Empirically, our default $\tau = 5$ had a better performance than simply excluding MFE base pairs with $\tau = 0$ (see Table 1). The CP prediction accuracy was relatively robust near the default $\tau = 5$, i.e., τ close to $\tau = 5$ predicted similar alternative structures (e.g., $0\leq\tau\leq5$ for *xpt* Guanine in Fig 6B). In fact, for the few sequences in our Barsacchi Structural dataset, predictions often did not change much for small values of τ. On the other hand, some high values of τ ($56\leq\tau\leq59$) improved the prediction accuracy for *xpt* Guanine dramatically (see Fig 6B). Unfortunately, attempts to tune τ through intricate dependencies on (e.g.) sequences or conditional probabilities did not change prediction accuracies much. Large τ values sometimes improved predictions dramatically for individual riboswitches, but unfortunately no general and systematic theme of improvement became apparent.

The effect of the seed on CP predictions

The seed is another adjustable CP parameter. After excluding base pairs close to the MFE structure from the energy landscape (see above), a seed is merely a substructure that occurs frequently in the remaining structures. Empirically, the longest bulge-containing consensus stem L^* made a good default seed for alternative structure prediction. We tested other seeds (e.g., seeds consisting of a single base pair), but the corresponding alternative structure predictions were generally less accurate (see the anecdotal case study in Fig 6D). Seeds with too many base pairs often produce poor predictions, because MFE and alternative structures often share many of the aptamer substructures.

Riboswitches tend to have longer stem seeds than non-riboswitches

A riboswitch classifier can use normalized seed length, because riboswitches tend to yield longer stem seeds than RNAs with a single structure. With our data and under the CP method, the normalized seed length of a riboswitch is 4.00bp; of a non-riboswitch, 3.50bp (see the Results section). Thus, a riboswitch of length 100nt has a stem seed of expected length 8bp ($4.00\times\log_{10} = 8$); a non-riboswitch of the same length; 7bp. The use of seed length in classification takes advantage of the notion that base pairs may have evolved to stabilize an alternative structure in riboswitches [45] while this evolutionary pressure does not exist in non-switching RNAs. In many purine riboswitches, the stem seed pointed to critical control substructures exclusive to the bound structure (Fig 5B). Although we did not explore other riboswitch families, it is interesting to speculate that if a folding pathway does not contain the MFE structure, a CP seed may often point to important regulatory substructures. Although different regulatory mechanisms of different RNA switches may influence the *type-of-seed*, for the purine riboswitches at least, biologically functional substructures often included our stem seed.

The effect of other parameters on CP predictions

The CP temperature parameter T may be particularly useful for predictions in riboswitches with thermosensory functions. Because RNA thermosensors lay beyond our purview, however, the present article relied on the default temperature 37˚C for CP prediction of alternative structures (note that predictions from sampling always sampled at several temperatures, however). The use of other temperatures in CP prediction did not change prediction accuracy dramatically. We also considered iterating our three-step CP prediction process: (1) exclude base pairs (2) determine a stem seed; and (3) find the lowest free-energy structure containing the stem seed (see the Materials and methods section for details). The threshold parameter τ seemed to offer a more gentle and graduated exclusion of base pairs than iteration, however. Iteration may prove useful for riboswitches with more than two metastable structures, but here, it only improved CP alternative structure prediction very occasionally (see Fig 6D, where the iterated predicted structure S_4^* is closer to the actual alternative structure S_2 than the default prediction S_2^*).

Conclusion

In this article, we use *exact* conditional base-pair probabilities to predict the alternative structure of RNA switches. Our approach selects base pairs associated with high conditional probabilities, after it excludes substructures in the primary metastable structure (here, the MFE structure). Conditioning on exclusion improves the chance that an exact (McCaskill) probability calculation finds base pairs in an alternative structure. Within the limitations imposed on our ROC tests by available data, Conditional-Probability (CP) computations had classification accuracy comparable to Sampling-Clustering (SC) computations. In contrast, however, the results for computational speed were not at all tentative: CP was more than 1000 times faster than SC, its speed making it a much more promising as a predictor of alternative structures in computationally demanding settings like genomic RNA.

CP predictions have very few adjustable parameters. Specifically, our dissimilarity threshold parameter τ varies the severity with which metastable substructures are excluded from predictions. Although available data is insufficient to optimize τ methodically, our default choice $\tau = 5$ was effective enough to render CP at least comparable to SC. Moreover, our tests suggest that varying τ offers a novel and computationally efficient method of traversing the energy landscapes to find metastable structures. We also showed that another important CP parameter, the seed, could detect functional substructures in riboswitches with regulatory functions as well as being able to classify riboswitches with similar performance to the more computationally costly sampling and clustering of the RNA energy landscape.

Availability

Our CondAlt source code for predicting alternative structures is publicly available for download at https://go.usa.gov/xRu79. The data used in the first dataset (Barsacchi) is available in S2 File and also publicly available in [54]. The data used in the second dataset (Purine aptamers) were taken from Rfam. Please refer to Material and Methods for further details.

Supporting information

S1 Table. Barsacchi riboswitch dataset. The 20 Riboswitches and their corresponding lengths.
(PDF)

S2 Table. Sensitivity and Positive-Predictive-Values (PPV) of the Barsacchi Structural riboswitch set.
(PDF)

S1 File. The relationship between Relaxed Base Pair score and dissimilarity threshold.
(PDF)

S2 File. Barsacchi datasets.
(PDF)

Acknowledgments

We would like to thank Prof. Michael Zuker for his help with computations using version 4.0.0 of the UNAFold Software. We would also like to thank Eric Nawrocki for some critical comments on our manuscript.

Author Contributions

Conceptualization: Amirhossein Manzourolajdad, John L. Spouge.

Formal analysis: Amirhossein Manzourolajdad.

Investigation: Amirhossein Manzourolajdad.

Methodology: Amirhossein Manzourolajdad, John L. Spouge.

Software: Amirhossein Manzourolajdad.

Supervision: John L. Spouge.

Writing – original draft: Amirhossein Manzourolajdad.

Writing – review & editing: John L. Spouge.

References

1. Grundy FJ, Henkin TM. From ribosome to riboswitch: control of gene expression in bacteria by RNA structural rearrangements. Critical reviews in biochemistry and molecular biology. 2006; 41(6):329–38. Epub 2006/11/10. https://doi.org/10.1080/10409230600914294 PMID: 17092822.

2. Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. Cell. 2003; 113(5):577–86. Epub 2003/06/06. PMID: 12787499.

3. Cheah MT, Wachter A, Sudarsan N, Breaker RR. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. Nature. 2007; 447(7143):497–500. Epub 2007/05/01. https://doi.org/10.1038/nature05769 PMID: 17468745.

4. Huthoff H, Berkhout B. Two alternating structures of the HIV-1 leader RNA. RNA. 2001; 7(1):143–57. PMID: 11214176

5. Huthoff H, Berkhout B. Multiple secondary structure rearrangements during HIV-1 RNA dimerization. Biochemistry. 2002; 41(33):10439–45. PMID: 12173930.

6. Legiewicz M, Badorrek CS, Turner KB, Fabris D, Hamm TE, Rekosh D, et al. Resistance to RevM10 inhibition reflects a conformational switch in the HIV-1 Rev response element. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(38):14365–70. https://doi.org/10.1073/pnas.0804461105 PMID: 18776047.

7. Olsthoorn RC, Mertens S, Brederode FT, Bol JF. A conformational switch at the 3' end of a plant virus RNA regulates viral replication. The EMBO journal. 1999; 18(17):4856–64. Epub 1999/09/02. https://doi.org/10.1093/emboj/18.17.4856 PMID: 10469663

8. Richter SN, Bélanger F, Zheng P, Rana TM. Dynamics of nascent mRNA folding and RNA–protein interactions: an alternative TAR RNA structure is involved in the control of HIV-1 mRNA transcription. Nucleic Acids Research. 2006; 34(15):4278–92. https://doi.org/10.1093/nar/gkl499

9. Sherpa C, Rausch JW, Le Grice SF, Hammarskjold ML, Rekosh D. The HIV-1 Rev response element (RRE) adopts alternative conformations that promote different rates of virus replication. Nucleic Acids Res. 2015; 43(9):4676–86. Epub 2015/04/10. https://doi.org/10.1093/nar/gkv313 PMID: 25855816.

10. Ancel LW, Fontana W. Plasticity, evolvability, and modularity in RNA. The Journal of experimental zoology. 2000; 288(3):242–83. PMID: 11069142.

11. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, et al. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Nucleic Acids Res. 2007; 35 (14):4809–19. Epub 2007/07/11. https://doi.org/10.1093/nar/gkm487 PMID: 17621584.

12. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, et al. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome biology. 2010; 11 (3):R31. Epub 2010/03/17. https://doi.org/10.1186/gb-2010-11-3-r31 PMID: 20230605.

13. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. 2015; 43(Database issue):D130–7. https://doi.org/10.1093/nar/gku1063 PMID: 25392425.

14. Serganov A, Nudler E. A decade of riboswitches. Cell. 2013; 152(1–2):17–24. Epub 2013/01/22. https://doi.org/10.1016/j.cell.2012.12.024 PMID: 23332744.

15. Henkin TM. Riboswitch RNAs: using RNA to sense cellular metabolism. Genes & development. 2008; 22(24):3383–90. Epub 2009/01/15. https://doi.org/10.1101/gad.1747308 PMID: 19141470.

16. McCown PJ, Corbino KA, Stav S, Sherlock ME, Breaker RR. Riboswitch diversity and distribution. RNA. 2017; 23(7):995–1011. https://doi.org/10.1261/rna.061234.117 PMID: 28396576.

17. Breaker RR. Riboswitches and the RNA world. Cold Spring Harbor perspectives in biology. 2012; 4(2). Epub 2010/11/26. https://doi.org/10.1101/cshperspect.a003566 PMID: 21106649.

18. Quarta G, Kim N, Izzo JA, Schlick T. Analysis of riboswitch structure and function by an energy landscape framework. Journal of molecular biology. 2009; 393(4):993–1003. Epub 2009/09/08. https://doi.org/10.1016/j.jmb.2009.08.062 PMID: 19733179.

19. Roy S, Hennelly SP, Lammert H, Onuchic José N, Sanbonmatsu KY. Magnesium controls aptamer-expression platform switching in the SAM-I riboswitch. Nucleic Acids Research. 2019; 47(6):3158–70. https://doi.org/10.1093/nar/gky1311%J Nucleic Acids Research. PMID: 30605518

20. Dann CE 3rd, Wakeman CA, Sieling CL, Baker SC, Irnov I, Winkler WC. Structure and mechanism of a metal-sensing regulatory RNA. Cell. 2007; 130(5):878–92. https://doi.org/10.1016/j.cell.2007.06.051 PMID: 17803910.

21. Zuker M. On finding all suboptimal foldings of an RNA molecule. Science (New York, NY). 1989; 244 (4900):48–52. Epub 1989/04/07. PMID: 2468181.

22. Zuker M. use of dynamic programming algorithms in RNA secondary structure prediction. 1989.

23. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31(13):3406–15. Epub 2003/06/26. https://doi.org/10.1093/nar/gkg595 PMID: 12824337.

24. Clote P. Computational Prediction of Riboswitches. Methods in Enzymology. 2015; 553:287–312. http://dx.doi.org/10.1016/bs.mie.2014.10.063. PMID: 25726470

25. Antunes D, Jorge NAN, Caffarena ER, Passetti F. Using RNA Sequence and Structure for the Prediction of Riboswitch Aptamer: A Comprehensive Review of Available Software and Tools. Frontiers in genetics. 2017; 8:231. Epub 2018/02/07. https://doi.org/10.3389/fgene.2017.00231 PMID: 29403526.

26. Manzourolajdad A, Wang Y, Shaw TI, Malmberg RL. Information-theoretic uncertainty of SCFG-modeled folding space of the non-coding RNA. Journal of theoretical biology. 2013; 318:140–63. https://doi.org/10.1016/j.jtbi.2012.10.023 PMID: 23160142.

27. Garcia-Martin JA, Clote P. RNA Thermodynamic Structural Entropy. PloS one. 2015; 10(11):e0137859. Epub 2015/11/12. https://doi.org/10.1371/journal.pone.0137859 PMID: 26555444.

28. Saffarian A, Giraud M, Touzet H. Modeling alternate RNA structures in genomic sequences. Journal of computational biology: a journal of computational molecular cell biology. 2015; 22(3):190–204. Epub 2015/03/15. https://doi.org/10.1089/cmb.2014.0272 PMID: 25768235.

29. Jager S, Schiller B, Babel P, Blumenroth M, Strufe T, Hamacher K. StreAM-Tg: algorithms for analyzing coarse grained RNA dynamics based on Markov models of connectivity-graphs. Algorithms for Molecular Biology. 2017; 12:2–16. https://doi.org/10.1186/s13015-017-0091-2

30. Voss B, Giegerich R, Rehmsmeier M. Complete probabilistic analysis of RNA shapes. BMC biology. 2006; 4:5. Epub 2006/02/17. https://doi.org/10.1186/1741-7007-4-5 PMID: 16480488.

31. Shapiro BA. An algorithm for comparing multiple RNA secondary structures. Computer applications in the biosciences: CABIOS. 1988; 4(3):387–93. Epub 1988/08/01. PMID: 2458170.

32. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics (Oxford, England). 2006; 22(4):500–3. Epub 2005/12/17. https://doi.org/10.1093/bioinformatics/btk010 PMID: 16357029.

33. Giegerich R, Voss B, Rehmsmeier M. Abstract shapes of RNA. Nucleic Acids Res. 2004; 32(16):4843–51. Epub 2004/09/17. https://doi.org/10.1093/nar/gkh779 PMID: 15371549.

34. Janssen S, Giegerich R. Faster computation of exact RNA shape probabilities. Bioinformatics (Oxford, England). 2010; 26(5):632–9. Epub 2010/01/19. https://doi.org/10.1093/bioinformatics/btq014 PMID: 20080511.

35. Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. 1999; 49(2):145–65. Epub 1999/03/10. https://doi.org/10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G PMID: 10070264.

36. Freyhult E, Moulton V, Clote P. RNAbor: a web server for RNA structural neighbors. Nucleic Acids Res. 2007; 35(Web Server issue):W305–9. https://doi.org/10.1093/nar/gkm255 PMID: 17526527.

37. Clote P, Lou F, Lorenz WA. Maximum expected accuracy structural neighbors of an RNA secondary structure. BMC bioinformatics. 2012; 13 Suppl 5:S6. Epub 2012/05/02. https://doi.org/10.1186/1471-2105-13-s5-s6 PMID: 22537010.

38. Michálik J, Touzet H, Ponty Y. Efficient approximations of RNA kinetics landscape using non-redundant sampling. Bioinformatics (Oxford, England). 2017; 33(14):i283–i92. https://doi.org/10.1093/bioinformatics/btx269 PMID: 28882001

39. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(19):7287–92. Epub 2004/05/05. https://doi.org/10.1073/pnas.0401799101 PMID: 15123812.

40. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of molecular biology. 1999; 288(5):911–40. Epub 1999/05/18. https://doi.org/10.1006/jmbi.1999.2700 PMID: 10329189.

41. Serra MJ, Turner DH. Predicting thermodynamic properties of RNA. Methods Enzymol. 1995; 259:242–61. PMID: 8538457.

42. Walter AE, Turner DH, Kim J, Lyttle MH, Muller P, Mathews DH, et al. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proceedings of the National Academy of Sciences of the United States of America. 1994; 91(20):9218–22. https://doi.org/10.1073/pnas.91.20.9218 PMID: 7524072.

43. Xia T, SantaLucia J Jr., Burkard ME, Kierzek R, Schroeder SJ, Jiao X, et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998; 37(42):14719–35. https://doi.org/10.1021/bi9809425 PMID: 9778347.

44. Quarta G, Sin K, Schlick T. Dynamic Energy Landscapes of Riboswitches Help Interpret Conformational Rearrangements and Function. PLoS Comput Biol. 2012; 8(2):e1002368. https://doi.org/10.1371/journal.pcbi.1002368 PMID: 22359488

45. Ritz J, Martin JS, Laederach A. Evolutionary Evidence for Alternative Structure in RNA Sequence Co-variation. PLOS Computational Biology. 2013; 9(7):e1003152. https://doi.org/10.1371/journal.pcbi.1003152 PMID: 23935473

46. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Research. 2003; 31(24):7280–301. https://doi.org/10.1093/nar/gkg938 PMID: 14654704

47. Tinoco I Jr., Uhlenbeck OC, Levine MD. Estimation of secondary structure in ribonucleic acids. Nature. 1971; 230(5293):362–7. PMID: 4927725.

48. McCaskill JS. The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure. Biopolymers. 1990; 29. https://doi.org/10.1002/bip.360290621 PMID: 1695107

49. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures (the Vienna RNA package). Monatshefte für Chemie / Chemical Monthly. 1994; 125(2):167–88.

50. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. Methods in molecular biology (Clifton, NJ). 2008; 453:3–31. Epub 2008/08/21. https://doi.org/10.1007/978-1-60327-429-6_1 PMID: 18712296.

51. Jaeger JA, Turner DH, Zuker M. Predicting optimal and suboptimal secondary structure for RNA. Methods Enzymol. 1990; 183:281–306. PMID: 1690335.

52. Zuker M. Prediction of RNA secondary structure by energy minimization. Methods in molecular biology (Clifton, NJ). 1994; 25:267–94. Epub 1994/01/01. https://doi.org/10.1385/0-89603-276-0:267 PMID: 7516239.

53. Zuker M, Mathews DH, Turner DH. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In: Barciszewski J, Clark BFC, editors. RNA Biochemistry and Biotechnology. Dordrecht: Springer Netherlands; 1999. p. 11–43.

54. Barsacchi M, Novoa EM, Kellis M, Bechini A. SwiSpot: modeling riboswitches by spotting out switching sequences. Bioinformatics (Oxford, England). 2016; 32(21):3252–9. Epub 2016/10/30. https://doi.org/10.1093/bioinformatics/btw401 PMID: 27378291.

55. Ames TD, Rodionov DA, Weinberg Z, Breaker RR. A eubacterial riboswitch class that senses the coenzyme tetrahydrofolate. Chemistry & biology. 2010; 17(7):681–5. Epub 2010/07/28. https://doi.org/10.1016/j.chembiol.2010.05.020 PMID: 20659680.

56. Baker JL, Sudarsan N, Weinberg Z, Roth A, Stockbridge RB, Breaker RR. Widespread genetic switches and toxicity resistance proteins for fluoride. Science (New York, NY). 2012; 335(6065):233–5. https://doi.org/10.1126/science.1215063 PMID: 22194412.

57. Batey RT, Gilbert SD, Montange RK. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature. 2004; 432(7015):411–5. Epub 2004/11/19. https://doi.org/10.1038/nature03037 PMID: 15549109.

58. Block KF, Hammond MC, Breaker RR. Evidence for widespread gene control function by the ydaO riboswitch candidate. Journal of bacteriology. 2010; 192(15):3983–9. Epub 2010/06/01. https://doi.org/10.1128/JB.00450-10 PMID: 20511502.

59. Corbino KA, Barrick JE, Lim J, Welz R, Tucker BJ, Puskarz I, et al. Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. Genome biology. 2005; 6(8):R70. Epub 2005/08/10. https://doi.org/10.1186/gb-2005-6-8-r70 PMID: 16086852.

60. Freyhult E, Moulton V, Clote P. Boltzmann probability of RNA structural neighbors and riboswitch detection. Bioinformatics (Oxford, England). 2007; 23(16):2054–62. Epub 2007/06/19. https://doi.org/10.1093/bioinformatics/btm314 PMID: 17573364.

61. Nahvi A, Barrick JE, Breaker RR. Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. Nucleic Acids Res. 2004; 32(1):143–50. https://doi.org/10.1093/nar/gkh167 PMID: 14704351.

62. Nechooshtan G, Elgrably-Weiss M, Sheaffer A, Westhof E, Altuvia S. A pH-responsive riboregulator. Genes & development. 2009; 23(22):2650–62. https://doi.org/10.1101/gad.552209 PMID: 19933154.

63. Ray PS, Jia J, Yao P, Majumder M, Hatzoglou M, Fox PL. A stress-responsive RNA switch regulates VEGFA expression. Nature. 2009; 457(7231):915–9. https://doi.org/10.1038/nature07598 PMID: 19098893.

64. Rieder R, Lang K, Graber D, Micura R. Ligand-induced folding of the adenosine deaminase A-riboswitch and implications on riboswitch translational control. Chembiochem: a European journal of chemical biology. 2007; 8(8):896–902. https://doi.org/10.1002/cbic.200700057 PMID: 17440909.

65. Sudarsan N, Barrick JE, Breaker RR. Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA. 2003; 9(6):644–7. https://doi.org/10.1261/rna.5090103 PMID: 12756322.

66. Sudarsan N, Cohen-Chalamish S, Nakamura S, Emilsson GM, Breaker RR. Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. Chemistry & biology. 2005; 12 (12):1325–35. https://doi.org/10.1016/j.chembiol.2005.10.007 PMID: 16356850.

67. Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, Breaker RR. An mRNA structure in bacteria that controls gene expression by binding lysine. Genes & development. 2003; 17(21):2688–97. Epub 2003/11/05. https://doi.org/10.1101/gad.1140003 PMID: 14597663.

68. Tomsic J, McDaniel BA, Grundy FJ, Henkin TM. Natural variability in S-adenosylmethionine (SAM)-dependent riboswitches: S-box elements in bacillus subtilis exhibit differential sensitivity to SAM In vivo and in vitro. Journal of bacteriology. 2008; 190(3):823–33. Epub 2007/11/28. https://doi.org/10.1128/JB.01034-07 PMID: 18039762.

69. Wachter A, Tunc-Ozdemir M, Grove BC, Green PJ, Shintani DK, Breaker RR. Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs. The Plant cell. 2007; 19(11):3437–50. Epub 2007/11/13. https://doi.org/10.1105/tpc.107.053645 PMID: 17993623.

70. Wang JX, Lee ER, Morales DR, Lim J, Breaker RR. Riboswitches that sense S-adenosylhomocysteine and activate genes involved in coenzyme recycling. Molecular cell. 2008; 29(6):691–702. Epub 2008/04/01. https://doi.org/10.1016/j.molcel.2008.01.012 PMID: 18374645.

71. Winkler W, Nahvi A, Breaker RR. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature. 2002; 419(6910):952–6. https://doi.org/10.1038/nature01145 PMID: 12410317.

72. Darty K, Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics (Oxford, England). 2009;25. https://doi.org/10.1093/bioinformatics/btp250 PMID: 19398448

73. Agius P, Bennett KP, Zuker M. Comparing RNA secondary structures using a relaxed base-pair score. RNA. 2010; 16(5):865–78. https://doi.org/10.1261/rna.903510 PMID: 20360393

74. Rivas E, Lang R, Eddy SR. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. RNA. 2012; 18(2):193–212. https://doi.org/10.1261/rna.030049.111 PMID: 22194308

75. Rijsbergen CJV. Information Retrieval: Butterworth-Heinemann; 1979. 208 p.

76. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747 PMID: 7063747

77. Serganov A, Yuan Y-R, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, et al. Structural Basis for Discriminative Regulation of Gene Expression by Adenine- and Guanine-Sensing mRNAs. Chemistry & biology. 2004; 11(12):1729–41. https://doi.org/10.1016/j.chembiol.2004.11.018 PMID: 15610857