BMC Bioinformatics

RESEARCH Open Access

Maximum expected accuracy structural neighbors of an RNA secondary structure

Peter Clote[1,2,3*], Feng Lou[2*], William A Lorenz[4]

From First IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011)
Orlando, FL, USA. 3-5 February 2011

Abstract

Background: Since RNA molecules regulate genes and control alternative splicing by *allostery*, it is important to develop algorithms to predict RNA *conformational switches*. Some tools, such as `paRNAss, RNAshapes and RNAbor`, can be used to predict potential conformational switches; nevertheless, no existent tool can detect general (i.e., not family specific) *entire* riboswitches (both aptamer and expression platform) with accuracy. Thus, the development of additional algorithms to detect conformational switches seems important, especially since the difference in free energy between the two metastable secondary structures may be as large as 15-20 kcal/mol. It has recently emerged that RNA secondary structure can be more accurately predicted by computing the *maximum expected accuracy* (MEA) structure, rather than the *minimum free energy* (MFE) structure.

Results: Given an arbitrary RNA secondary structure S_0 for an RNA nucleotide sequence $a = a_1,..., a_n$, we say that another secondary structure S of a is a k-neighbor of S_0, if the base pair distance between S_0 and S is k. In this paper, we prove that the Boltzmann probability of all k-neighbors of the minimum free energy structure S_0 can be approximated with accuracy ε and confidence 1 - p, simultaneously for all $0 \leq k < K$, by a relative frequency count over N sampled structures, provided that $N > N(\varepsilon, p, K) = \frac{\Phi^{-1}\left(\frac{p}{2K}\right)^2}{4\varepsilon^2}$, where $\Phi(z)$ is the cumulative distribution function (CDF) for the standard normal distribution. We go on to describe the algorithm `RNAborMEA`, which for an arbitrary initial structure S_0 and for all values $0 \leq k < K$, computes the secondary structure $MEA(k)$, having *maximum expected accuracy* over all k-neighbors of S_0. Computation time is $O(n^3 \cdot K^2)$, and memory requirements are $O(n^2 \cdot K)$. We analyze a sample TPP riboswitch, and apply our algorithm to the class of *purine riboswitches*.

Conclusions: The approximation of `RNAbor` by sampling, with rigorous bound on accuracy, together with the computation of maximum expected accuracy k-neighbors by `RNAborMEA`, provide additional tools toward conformational switch detection. Results from `RNAborMEA` are quite distinct from other tools, such as `RNAbor, RNAshapes and paRNAss`, hence may provide orthogonal information when looking for suboptimal structures or conformational switches. Source code for `RNAborMEA` can be downloaded from http://sourceforge.net/projects/rnabormea/ or http://bioinformatics.bc.edu/clotelab/RNAborMEA/.

* Correspondence: clote@bc.edu; lou@lri.fr
[1]Department of Biology, Boston College, Chestnut Hill, MA 02467, USA
[2]Laboratoire de Recherche en Informatique (LRI), Université Paris-Sud XI, 91405 Orsay cedex, France
Full list of author information is available at the end of the article

Background

RNA secondary structure conformational switches play an essential role in a number of biological processes, such as regulation of viral replication [1] and of viroid replication [2], regulation of R1 plasmid copy number in *E. coli* by *hok/sok* system [3], transcriptional and translational gene regulation in prokaryotes by riboswitches [4], regulation of alternative splicing in eukaryotes [5], and stress-responsive gene regulation in humans [6], etc. Due to the biological importance of conformational switches, several groups have developed algorithms that attempt to recognize switches - in particular, thermodynamics-based methods such as `paRNAss`[7], `RNAshapes`[8], `RNAbor`[9], as well as an approach using the second eigenvalue of the Laplacian matrix [10].

Riboswitches are portions of the 5' untranslated region (UTR) of messenger RNAs, experimentally known to regulate genes in bacteria by *allostery* [4], and to regulate alternative splicing of the gene NMT1 in the eukaryote *Neurospora crassa* [5]. Riboswitches are composed of a 5' *aptamer* and a 3' *expression platform.* Since the aptamer binds to a specific ligand with high affinity (K_D ≈ 5 nM), thus triggering the conformational change of the expression platform upon ligand binding [11], its sequence and secondary structure tend to be highly conserved. In contrast, there is lower sequence for the expression platform, which forms a bistable switch, effecting gene regulation by premature abortion of transcription (as in guanine riboswitches [12]), or by sequestering the Shine-Dalgarno sequence (as in thiamine pyrophosphate riboswitches [13]). Due to the conserved sequence and secondary structure within the aptamer, all existent algorithms (to the best of our knowledge), such as [14-16], attempt only to detect riboswitch *aptamers*, without the expression platform. In addition to these specific algorithmic approaches, more general computational tools that rely on *stochastic context free grammars*, such as `Infernal`[17] and `CMFinder`[18], have been trained to recognize riboswitch aptamers; in particular, `Infernal` was used to create the Rfam database [19], which includes 14 families of riboswitch aptamers.

Since current riboswitch detection algorithms do not attempt to predict the location of the expression platform, we have developed tools, `RNAbor-Sample` and `RNAborMEA`, described in this paper, which yield information concerning alternative, or suboptimal, structures of a given RNA sequence. These tools can suggest the presence of a conformational switch; however, much more work must be done to actually produce a riboswitch gene finder, part of the difficulty due to the fact that riboswitch aptamers contain *pseudoknots* that cannot be captured by secondary structure.

In previous work [20,21], we described a novel program `RNAbor` to predict RNA conformational switches. For a given secondary structure S of a given RNA sequence $\mathbf{s}$, the secondary structure T of $\mathbf{s}$ is said to be a k-neighbor of S, if the base pair distance between S and T is k. (Base pair distance is the minimum number of base pairs that must be either added or removed, in order to transform the structure S into T.) Given an arbitrary initial structure S_0, for all values $0 \le k < K$, the program `RNAbor`[20], computes the secondary structure $MFE(k)$, having minimum free energy over all k-neighbors of S_0. (Note that $K \le 2 \cdot n$, since the base pair distance between any two secondary structures of a length n RNA sequence is at most $2 \cdot n$.) As well, `RNAbor` computes for each value $0 \le k \le K$, the Boltzmann probability $p_k = \frac{Z(k)}{Z}$, where $Z(k)$ is the sum of all Boltzmann factors $\exp(-E(S)/RT)$ of all structures S having base pair distance k from an initially given structure S_0, and where the *partition function* Z is the sum of all Boltzmann factors of all secondary structures of the given RNA sequence. Here $E(S)$ is the free energy of secondary structure S, with respect to the Turner energy model [22,23], $R = 0.001987$ kcal mol^{-1} K^{-1} is the universal gas constant, and T is absolute temperature. In the case that S_0 is the minimum free energy structure, the existence of one or more 'peaks', or values $k \gg 0$, where p_k is relatively large, suggests that there are two or more low energy structures having large base pair distance k from S_0 - i.e., a potentially distinct metastable structure, as shown in Figure 1.

In [24], Do et al. introduced the notion of *maximum expected accuracy* (MEA) secondary structure, determined as follows: *(i)* compute base pairing probabilities $p(i, j)$ using a trained stochastic context free grammar; *(ii)* compute probabilities $q(i) = 1 - \sum_{i<j} p(i,j) - \sum_{j<i} p(j,i)$ that position i does not pair; *(iii)* using a dynamic programming algorithm similar to that of Nussinov and Jacobson [25], determine that secondary structure S having maximum score $\sum_{(i,j)\in S} 2\alpha \cdot p(i,j) + \sum_{i\text{unpaired}} \beta q_i$, where the first sum is over paired positions (i, j) of S and the second sum is over positions i located in loop regions of S, and where α, β >0 are parameters with default values 1. Subsequently Kiryu et al. [26] computed the MEA structure by replacing the stochastic context free grammar computation of base pairs in *(i)* by using McCaskill's algorithm [27], which computes the Boltzmann base pairing probabilities

$$p(i,j) = \frac{\sum_{\{S:(i,j)\in S\}} \exp(-E(S)/RT)}{\sum_S \exp(-E(S)/RT)} \qquad (1)$$

The sum in the numerator is taken over all secondary structures of the given RNA sequence, that contain base

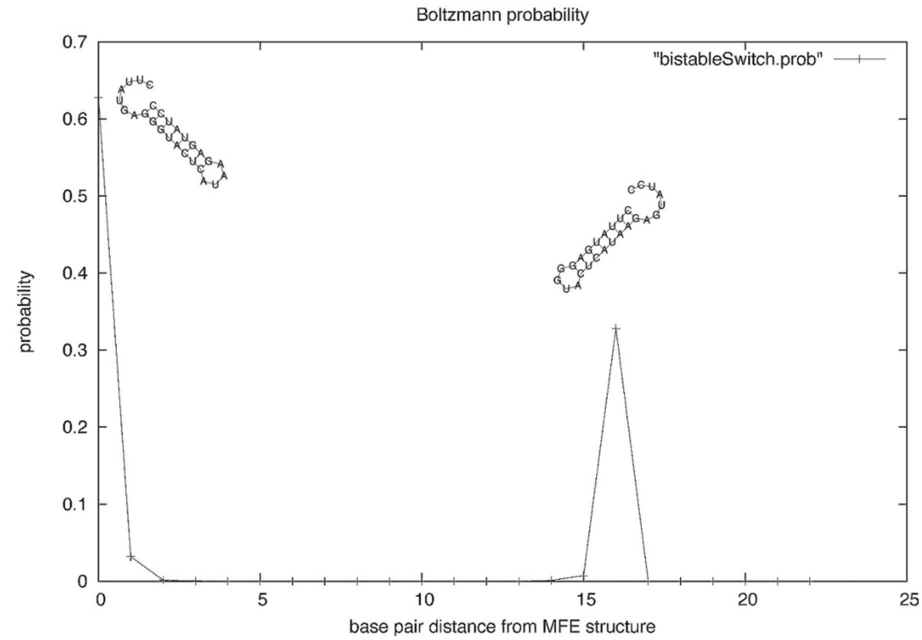


Figure 1 Output of `RNAbor`**on the 27 nt bistable switch with nucleotide sequence CUUAUGAGGG UACUCAUAAG AGUAUCC and initial structure S_0, the minimum free energy (MFE) structure....... ((((((((....)))))))) with free energy -10.3 kcal/mol**. The 16-neighbor of S_0 is the metastable structure ((((((((....))))))))....... with free energy -9.9 kcal/mol. The MFE structure appears above the leftmost peak, while the *MFE*(16) structure appears above the rightmost peak. The output of `RNAbor` includes a graph of the Boltzmann probabilities $p_k = \frac{Z_k}{Z}$, and *MFE*(*k*) structures, for all $0 \leq k \leq 2n$. The existence of distinct 'peaks' suggests the presence of a conformational switch.

pair (*i*, *j*), while the sum in the denominator is taken over all secondary structures of the given RNA sequence. Thus *p*(*i*, *j*) is the sum of the Boltzmann factors of all secondary structures that contain the fixed base pair (*i*, *j*), divided by the partition function, which latter is the sum of Boltzmann factors of all secondary structures. In fact, Kiryu et al. [26] describe an algorithm to compute the MEA structure common to all RNAs in a given alignment. Later, Lu et al. [28] rediscovered Kiryu's method; in addition, Lu et al. computed suboptimal MEA structures by implementing an analogue of Zuker's method [29].

Our motivation in developing both `RNAbor-Sample` and `RNAborMEA`, was to simplify and improve our previous software, `RNAbor`, in detecting conformational switches. Since `RNAbor` computes the minimum free energy structure, *MFE*(*k*), over all structures having base pair distance *k* from an initially given structure S_0, a complex portion of the code in `RNAbor` concerns the retrieval of free energy parameters from the Turner model [22,23]. The idea of `RNAborMEA` was to compute the base pairing probabilities *p*(*i*, *j*) - see equation (1) - by McCaskill's algorithm using `RNAfold`, then to compute the maximum expected accuracy structure, *MEA* (*k*), which needs no retrieval of energy parameters, and which we hoped would be very similar to the *MFE*(*k*) structure, in light of previously mentioned results [26,28]. Surprisingly, it turns out that *MEA*(*k*) structures are quite different from *MFE*(*k*) structures, as shown later in one of the figures.

In this paper, we begin by showing rigorously how to approximate the output of `RNAbor` by frequency counts from sampling, using `Sfold`[30]. We then extend the MEA technique to compute the maximum expected accuracy *k-neighbor* of a given RNA secondary structure S_0; i.e., that secondary structure which has maximum expected accuracy over all structures that differ from S_0 by exactly *k* base pairs. By analyzing the family of purine riboswitches, obtained by retrieving full riboswitch sequences (aptamer and expression platform) from corresponding EMBL genomic data, by extending the aptamers from the seed alignment of Rfam family RF00167 [31], we show that our software `RNAborMEA` produces strikingly different results from other software that produce suboptimal structures (`RNAbor, RNAbor-Sample, RNAlocopt, RNAshapes, UNAFold`).

Since the detection of computational switches remains an open problem, despite the success of some tools such as `RNAshapes` and `RNAbor`, we feel the addition of the tool `RNAborMEA` could prove useful, since it appears to be orthogonal to all other methods of generating suboptimal secondary structures.

Results and discussion

In this paper, we describe the following new results, discussed in the 'Methods' section in greater detail with attendant definitions of unexplained concepts.

1. We describe a Python script `RNAbor-Sample` that approximates the output $p_k = \frac{Z_k}{Z}$ of `RNAbor` by frequency counts $\hat{p}_k$ from sampled structures, for all $0 \leq k \leq 2n$, using `Sfold`[30], or `RNAsubopt -p` [32].
2. We prove that for any desired accuracy $0 < \varepsilon$ and probability $0 < \alpha < 1$, if at least

$$N(\varepsilon, p, K) = \frac{\Phi^{-1}\left(\frac{p}{2K}\right)^2}{4\varepsilon^2} \tag{2}$$

structures are sampled, then

$$P(|p_k - \hat{p}_k| < \varepsilon) > 1 - \alpha \tag{3}$$

for all $0 \leq k < K$; i.e., `RNAbor-Sample` furnishes estimates $\hat{p}_k$ of p_k, for all $0 \leq k < K$, which with confidence $1 - \alpha$ are within ε of the actual values p_k. Here, $\Phi(z)$ is the cumulative distribution function (CDF) for the standard normal distribution.
3. We develop an algorithm, `RNAborMEA`, running in time $O(n^3 \cdot K^2)$ and space $O(n^2 \cdot K)$, which computes simultaneously for all $0 \leq k \leq K$, the *maximum expected accuracy k-neighbors* of a given RNA secondary structure S_0; i.e., for each $0 \leq k \leq K$, `RNAborMEA` computes that structure S_k which has maximum expected accuracy over all structures that differ from S_0 by exactly k base pairs. The algorithm `RNAborMEA` additionally computes, for each $0 \leq k \leq K$, the *pseudo* partition function

$$\widetilde{Z}_k = \sum_{\{S: d_{\mathrm{BP}}(S,S_0)=k\}} \exp(MEA(S)/RT).$$

Moreover, `RNAborMEA` allows the user to stipulate (partial) hard constraints, that stipulate whether particular nucleotides are unpaired, or base-pair with certain other nucleotides. The implementation of hard constraints follows ideas from Mathews [33], albeit suitably modified to simultaneously consider all k-neighbors, for $0 \leq k \leq K$.

We now describe the 13 figures and 4 tables, corresponding to computational experiments performed with `RNAbor-Sample` and `RNAborMEA`. These tables and figures are only briefly described, and we refer the reader to the captions of the figures and tables, which explain the results in greater detail.

Figure 1 illustrates the presence of two peaks, corresponding to the Boltzmann probability of each of the metastable structures for a 27 nt bistable switch previously considered by Hofacker et al. Figure 2 displays the Boltzmann probabilities p_k from `RNAbor`, Boltzmann probabilities estimates $\hat{p}_k$ from `RNAbor-Sample` for the SAM riboswitch aptamer with GenBank accession code AP004597.1/11894-11904. Clearly, probability estimates $\hat{p}_k$ are close to actual values p_k. The figure additionally shows probabilities r_k from our software `RNAlocopt`[34], computed by $r_k = \frac{Z_k(LO)}{Z(LO)}$, where $Z(LO)$ is the sum of Boltzmann factors of all *locally optimal* secondary structures, and $Z_k(LO)$ is the sum of all locally optimal k-neighbors of S_0. A secondary structure S is said to be *locally optimal*, if its energy does not decrease by the addition or removal of a single (valid) base pair; i.e., $E(S \cup \{(x, y)\}) \geq E(S)$, and $E(S - \{(x, y)\}) \geq E(S)$. Figure 3 displays the experimentally determined GENE ON and GENE OFF structures of an XPT guanine riboswitch from *B. subtilis*, taken from [35]. Figure 4 shows the outputs of `RNAborMEA`, `RNAbor`, and `RNAshapes`, which are most similar to the GENE ON structure from the previous Figure 3. Figures 5 and 6 determine the structural simlarity, as measured by the program `NestedAlign`[36], between that structure output by `RNAborMEA` (as well as structures output by `RNAbor`, `RNAbor-Sample`, `RNAlocopt`, `RNAshapes`, and `UNAFold`), which are most similar to the XPT purine riboswitch, displayed in Figure 3. Figure 5 determines the structural similarity to the GENE ON structure (left panel of Figure 3), while Figure 6 determines the structural similarity to the GENE OFF structure (right panel of Figure 3). None of the structural neighbors, or sampled structures, are identical to the GENE ON or GENE OFF structures; however, there are some candidates that bear some resemblance to those structures. At this point, we can say that `RNAbor-Sample` and `RNAborMEA` are methods that generate suboptimal structures, some of which may be similar to the metastable structures of a conformational switch; however, much additional work is necessary before a robust method can be developed to detect conformational switches.

Figure 7 shows that the $MEA(k)$ structural neighbors, as computed by `RNAborMEA`, are very different than the $MFE(k)$ structural neighbors, as computed by `RNAbor`. At present, such computational experiments show `RNAborMEA` computes suboptimal structures, which seem to share (chimeric) similarities between parts of low energy structures, but which themselves do not have

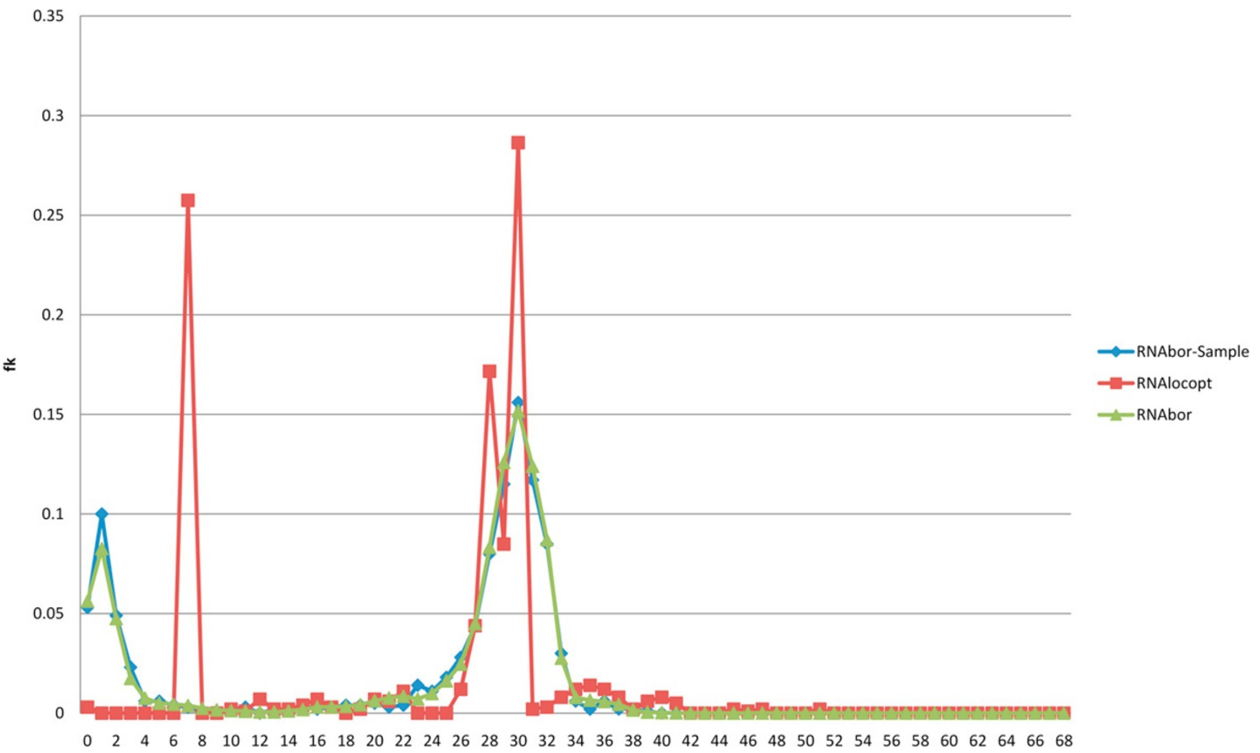


Figure 2 Boltzmann density plot for `RNAbor`**, along with approximating relative frequency plots for** `RNAborMEA`**and**`RNAlocopt`**for the 101 nt RNA sequence UACUUAUCAA GAGAGGUGGA GGGACUGGCC CGCUGAAACC UCAGCAACAG AACGCAUCUG UCUGUGCUAA AUCCUGCAAG CAAUAGCUUG AAAGAUAAGU U for the SAM riboswitch aptamter with GenBank accession code** AP004597.1/118941-119041. The program `RNAbor` computes the Boltzmann probability $p_k = \frac{Z_k}{Z}$, where $Z_k = \sum_{\{S:d_{BP}(S,S_0)=k\}} \exp(-E(S)/RT)$, where S_0 is the initial structure (taken as the minimum free energy here). The script `RNAbor-Sample` calls `Sfold` on 1000 structures, in order to compute a relative frequence $f_k \approx p_k$ of all k-neighbors of S_0. Finally, we compute relative frequency of `RNAlocopt`[34], a program that samples only *locally optimal* secondary structures, having the property that one cannot obtain a lower energy structure by adding or removing a single base pair.

very low energies. Such suboptimal structures appear to be 'orthogonal' to those output by all other methods, such as `Sfold`, `RNAbor`, `RNAbor-Sample`, `RNAlocopt`, `RNAshapes`, `UNAFold`). Figure 8 displays the output of `RNAborMEA`, given the sequence of a TPP riboswitch with EMBL accession code AF269819/1811-1669. In this instance, `RNAborMEA` found two low energy structures having large base pair distance from each other. (Other computational experiments did not yield such a good example.) Figure 9 displays the free energy and maximum expected accuracy scores, for each of the k-neighbors of the given TPP riboswitch sequence, just described in Figure 8. Figures 10 and 11 present the pseudocode for the `RNAborMEA` algorithm, which given an RNA sequence $a_1, .. .,a_n$ and initial structure S_0, computes the $MEA(k)$ structure and pseudo partition function $\tilde{Z}_k$, for each $0 \le k \le K$ in time $O(n^3 \cdot K^2)$ and space $O(n^2 \cdot K)$. Figure 12 presents pseudocode for the $O(n^2)$ algorithm to sample structures from the ensemble of structures having high MEA scores - a maximum expected accuracy analogue of the sampling algorithm `Sfold`[30]. Figure 13 displays the pseudo-Boltzmann probabilities $\tilde{p}_k = \frac{\tilde{Z}_k}{Z}$ for two small RNA sequences. While temperature T has a natural significance, when computing Boltzmann probabilities $p_k = \frac{Z_k}{Z}$, there is no natural meaning of temperature T, when computing pseudo Boltzmann factors $\exp(MEA(S)/RT)$, and indeed very different curves can be obtained with different temperatures.

We now briefly describe Tables 1, 2, 3, 4. Table 1 provides some sample sizes N, computed by the formula from equation (2), for an ε approximation of Boltzmann probabilities p_k, $0 \le k < K$, with $1 - \alpha$ confidence level. Tables 2 and 3 provide the numerical values for the earlier described Figures 5 and 6, where the `NestedAlign` structural similarity is computed for the most similar k-neighbor, determined by `RNAborMEA`, `RNAbor-Sample` and `RNAlocopt`. Table 4 presents the number of times that each of the methods `RNAborMEA`, `RNAbor`, `RNAbor-Sample`, `RNAlocopt`, `RNAshapes`, `UNAFold` output the most similar structure to the GENE ON resp. GENE OFF structure for the

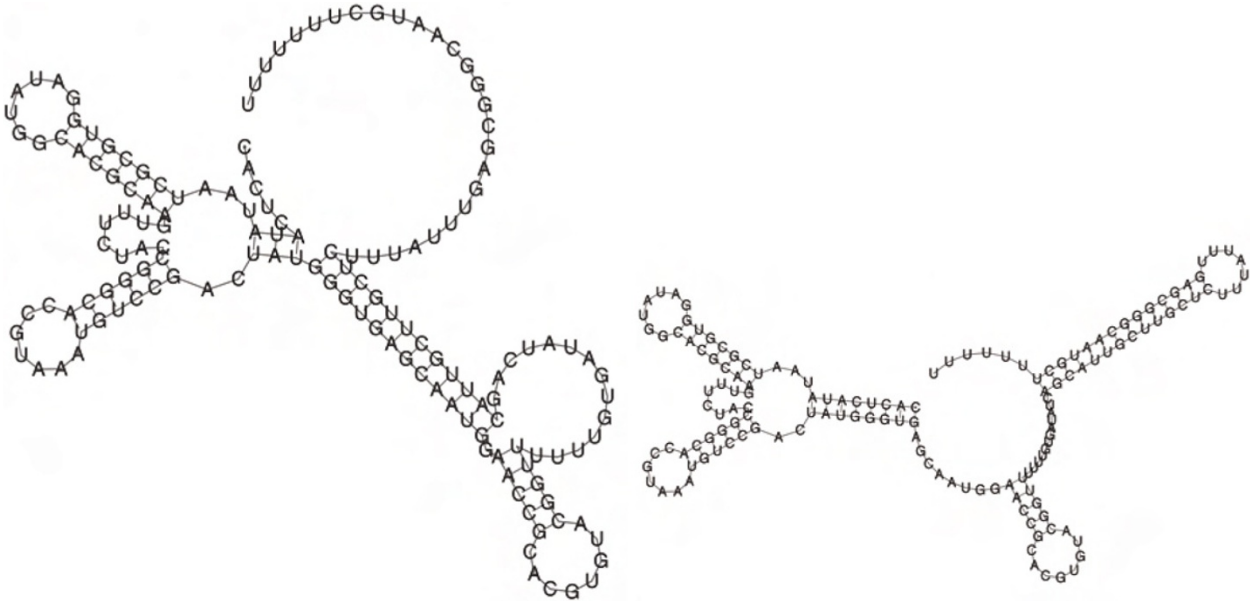


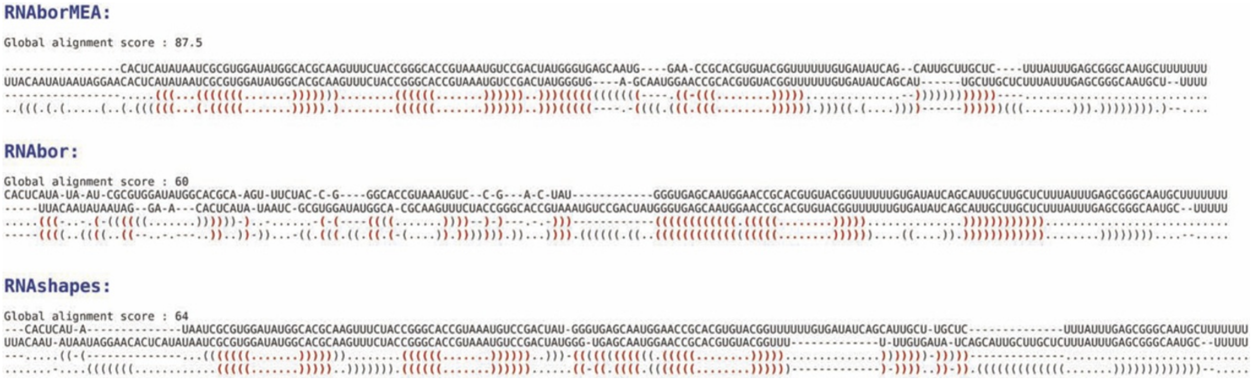
Figure 3 GENE ON (left) and GENE OFF (right) secondary structures for the 148 nt. XPT guanine riboswitch from *B. subtilis* with sequence CACUCAUAUA AUCGCGUGGA UAUGGCACGC AAGUUUCUAC CGGGCACCGU AAAUGUCCGA CUAUGGGUGA GCAAUGGAAC CGCACGUGUA CGGUUUUUUG UGAUAUCAGC AUUGCUUGCU CUUUAUUUGA GCGGGCAAUG CUUUUUUU. Sequence and secondary structure taken from [35]. The free energy of GENE ON resp. GENE OFF secondary structrure is -16.46 kcal/mol resp. -22.6 kcal/mol. Free energies were determined using RNAeval and figures produced using RNAplot, both from the Vienna RNA Package [40].

XPT purine riboswitch described in Figure 3. This computational experiment was performed for all RNA sequences in the seed alignment of the Rfam purine riboswitch family RF00167 [31]. This table shows that RNAborMEA and RNAbor both outperform any other method in determining structures similar to the GENE OFF XPT structure; however, RNAborMEA uniquely outperforms all methods, including RNAbor, in determining structures similar to the GENE ON XPT structure. One of the reasons for this excellent result is that unlike other methods, RNAborMEA does *not* look for low energy structures, but rather for maximum expected accuracy structures.

The figures and tables show, in summary, that RNAborMEA provides useful suboptimal structures, which may be closer to metastable structures of a conformational switch than more traditional methods, which rely on searching for low energy structures.

Conclusions

We have applied the notion of *maximum expected accuracy* within the context of *structural* neighbors of a



RNAborMEA:

```
Global alignment score : 87.5

-----------------CACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGGGUGAGCAAUG----GAA-CCGCACGUGUACGGUUUUUUGUGAUAUCAG--CAUUGCUUGCUC----UUUAUUUGAGCGGGCAAUGCUUUUUUU
UUACAAUAUAAUAGGAACACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGGGUG----A-GCAAUGGAACCGCACGUGUACGGUUUUUUGUGAUAUCAGCAU------UGCUUGCUCUUUAUUUGAGCGGGCAAUGCU--UUUU
-----------------.....(((...(((((((.......)))))))........((((((.......))))))..)))((((((((((((----.((-(((.........)))))..............--))))))))))))---...........................
..(((.(.(.....(..(.((((((...(.(((((.......))))).)........((((((.......))))))..)))(((((----.-((((.(((.(((.........)))))).)))((.(....))))------))))))(((.......))).))))))))).)--....
```

RNAbor:

```
Global alignment score : 60
CACUCAUA-UA-AU-CGCGUGGAUAUGGCACGCA-AGU-UUCUAC-C-G----GGCACCGUAAAUGUC--C-G---A-C-UAU-------------GGGUGAGCAAUGGAACCGCACGUGUACGGUUUUUUGUGAUAUCAGCAUUGCUUGCUCUUUAUUUGAGCGGGCAAUGCUUUUUUU
-----UUACAAUAUAAUAG--GA-A---CACUCAUA-UAAUC-GCGUGGAUAUGGCA-CGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGGGUGAGCAAUGGAACCGCACGUGUACGGUUUUUUGUGAUAUCAGCAUUGCUUGCUCUUUAUUUGAGCGGGCAAUGC--UUUUU
.....(((-...-.(-(((((((.......)))))))-)..-.......-(-(----((((.......))))--)-)---.-.-)))-------------((((((((((((.(((((.........)))))..................))))))))))))..............................
-----((((..((((..((--..-.---..))..))-))...-((.(((.((.((.(-(....)).)))))))).))...)))).((((((.((..(((((((((((((((((((.........)))))))......((....)).))))))))))))).........))))))))....--.....
```

RNAshapes:

```
Global alignment score : 64
---CACUCAU-A---------------UAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAU-GGGUGAGCAAUGGAACCGCACGUGUACGGUUUUUUGUGAUAUCAGCAUUGCU-UGCUC--------------UUUAUUUGAGCGGGCAAUGCUUUUUUU
UUACAAU-AUAAUAGGAACACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGGG-UGAGCAAUGGAACCGCACGUGUACGGUUU--------------U-UUGUGAUA-UCAGCAUUGCUUGCUCUUUAUUUGAGCGGGCAAUGC--UUUUU
---.....((-(---------------...(((((((.......)))))))........((((((.......))))))..)))-((((((((((((.(((((.........)))))...............)))))))-))))))--------------................................
.......-....((((((((.............(((((.......)))))..)))))))).((((((.......))))))......((-((.((((.((((((.........))))))-------------)-)))))..))-)).(((((((((((((.......)))))))))))))--.....
```

Figure 4 Given riboswitch sequence X83878/168-267 and initial structure S_0, the minimum free energy structure, a structure output by RNAborMEAis most structurally similar to the XPT GENE ON**structure, as measured by** NestedAlign[36]. The NestedAlign score for RNAborMEA is 87.5, while optimal score for RNAbor is 60.0, and for RNAshapes is 64.0.

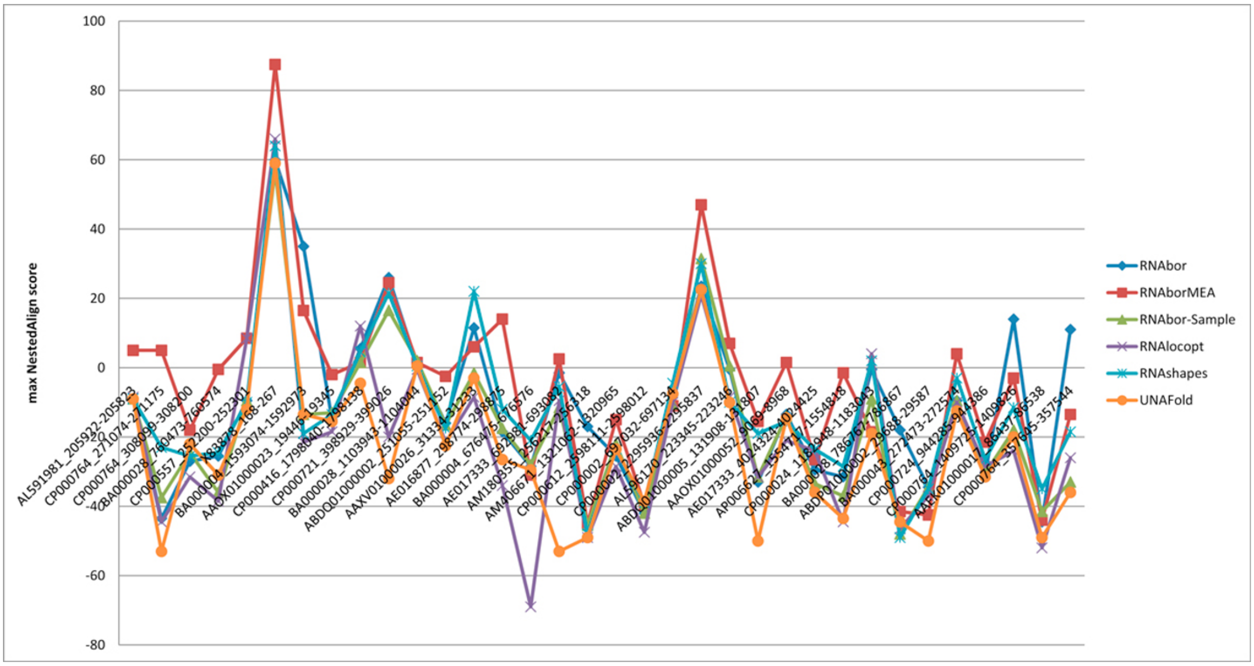


Figure 5 For each RNA sequence in the seed alignment from Rfam family RF00167 of purine riboswitch *aptamers*, we retrieved downstream flanking residues from the appropriate EMBL files, in order to ensure likelihood that the expression platform was included. Then the following six programs were run: RNAbor, RNAborMEA, RNAbor-Sample, RNAlocopt, RNAshapes, UNAFold. Each program outputs a number of *near-optimal* secondary structures, each according to different criteria. Taking RNAbor and RNAborMEA as examples, the programs RNAbor and RNAborMEA were run, in order to compute the *MFE*(k) structure and the *MEA*(k) structure, which have *minimum free energy* resp. *maximum expected accuracy* among all k-neighbors of the intial minimum free energy structures S_0. Subsequently, we applied the program NestedAlign described in [36] to compute the *structural similarity* between the experimentally determined GENE ON structure for XPT guanine riboswitch of *B. subtilis*; i.e. the left panel of Figure 3. (Similar structures have positive scores; dissimilar structures have negative scores.) For each RNA in the seed alignment of RF00167, we determined the value k_1, such the *MEA*(k_1) structure for that RNA has the greatest structural similarity with the XPT GENE ON structure, as determined by NestedAlign. (See the left panel of Figure 3 for the experimentally determined GENE ON structure of XPT.) As earlier explained, we performed similar computations for the programs RNAshapes [39] and UNAFold [41], the programs RNAborMEA and RNAbor-Sample, described in this paper, and programs RNAbor[9] and RNAlocopt [34], developed by our lab. In 21 out of 34 instances, RNAborMEA produced the secondary structure most structurally similar to the experimentally determined XPT GENE OFF structure, as measured by NestedAlign.

given RNA sequence $a_1, .. ., a_n$ and structure S_0. Our software RNAborMEA not only computes the structures $MEA(k)$ having maximum expected accuracy over all structures S, whose base pair distance $d_{\mathrm{BP}}(S_0, S)$ is equal to k. In addition, RNAborMEA allows the user to enter *structural constraints*, which specify partial secondary structures required of all $MEA(k)$ structures, if so desired. Additionally, RNAborMEA computes an analogue of the temperature-dependent partition function, defined by

$$\widetilde{Z}_k(T) = \sum_{\{S:d_{\mathrm{BP}}(S_0,S)=k\}} \exp(\sigma(S))/RT$$

and

$$\widetilde{Z(T)} = \sum_k \widetilde{Z}_k = \sum_S \exp(\sigma(S))/RT.$$

Here, the expected accuracy score σ is defined by

$$\sigma(S) = 2 \cdot \sum_{(i,j)\in S} p_{i,j} + \sum_{i\text{unpaired}} q_i$$

where first sum is taken over all base pairs (i, j) belonging to S, and the second sum is taken over all unpaired positions in S, and where $p_{i,j}$ [resp. q_i] is the probability that i, j are paired [resp. i is unpaired] in the ensemble of low energy structures, as computed by McCaskill's algorithm [27]. Finally, RNAborMEA allows the user to sample structures from the maximum expected accuracy ensemble, in a fashion analogous to Ding-Lawrence sampling from the low energy Boltzmann ensemble, as in Sfold[30].

Our preliminary investigations have not indicated a clear application of the partition function analogue, though it may be construed to provide a representation

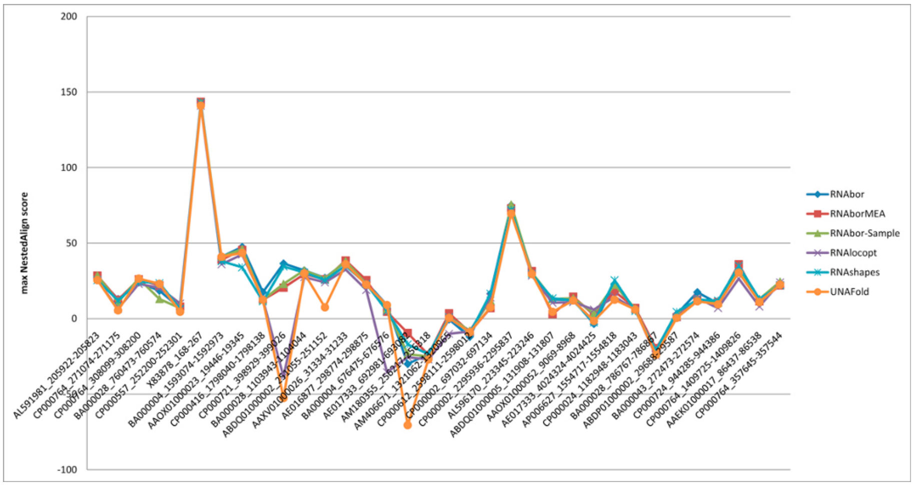


Figure 6 For each RNA sequence in the seed alignment from Rfam family RF00167 of purine riboswitch *aptamers*, we retrieved downstream flanking residues from the appropriate EMBL files, in order to ensure likelihood that the expression platform was included. Then the following six programs were run: RNAbor, RNAborMEA, RNAbor-Sample, RNAlocopt, RNAshapes, UNAFold. Each program outputs a number of *near-optimal* secondary structures, each according to different criteria. Taking RNAbor and RNAborMEA as examples, the programs RNAbor and RNAborMEA were run, in order to compute the *MFE*(k) structure and the *MEA*(k) structure, which have *minimum free energy* resp. *maximum expected accuracy* among all k-neighbors of the intial minimum free energy structures S_0. Subsequently, we applied the program NestedAlign described in [36] to compute the *structural similarity* between the experimentally determined GENE OFF structure for XPT guanine riboswitch of *B. subtilis*; i.e. the right panel of Figure 3. (Similar structures have positive scores; dissimilar structures have negative scores.) For each RNA of the seed alignment of RF00167, we determined the value k_1, such the *MEA*(k_1) structure for that RNA has the greatest structural similarity with the XPT GENE OFF structure, as determined by NestedAlign. (See the right panel of Figure 3 for the experimentally determined GENE OFF structure of XPT.) As earlier explained, we performed similar computations for the programs RNAshapes [39] and UNAFold [41], the programs RNAborMEA and RNAbor-Sample, described in this paper, and RNAbor[9] and RNAlocopt[34]. In 22 out of 34 instances, RNAborMEA produced the secondary structure most structurally similar to the experimentally determined XPT GENE OFF structure, as measured by NestedAlign.

of the temperature-dependent *mixing* of various structures having large score σ. On the other hand, in computational experiments reported in the Results Section, it appears that RNAborMEA produces near-optimal structures that are closer to the biologically functional structures, in the case of conformational switches that are difficult to predict by any method.

Indeed, in 18 [resp. 11] out of 34 instances, RNAborMEA produced the secondary structure most structurally similar to the experimentally determined XPT GENE ON [resp. GENE OFF] structure, as measured by NestedAlign[36]. See Table 4. Since there appears to be little to no correlation between the structures *MFE*(k) output by RNAbor[20] and the structures *MEA*(k) output by our current program RNAborMEA, it appears that RNAborMEA yields a signal that is orthogonal and complementary to that provided by state-of-the-art thermodynamics software, such as UNAFold, RNAfold, RNAstructure, Sfold, RNAshapes, RNAbor, etc. For these reasons, we feel that RNAborMEA has a certain value, along with the programs UNAFold, RNAfold, RNAstructure, Sfold, RNAshapes, RNAbor, etc. when producing suboptimal structures. RNAborMEA is written in C and available at http://sourceforge.net/projects/rnabormea/ and http://bioinformatics.bc.edu/clotelab/RNAborMEA/.

Methods

Preliminaries

Recall the definition of RNA secondary structure.

Definition 1 *A secondary structure S on RNA sequence* $a_1, .. ., a_n$ *is defined to be a set of ordered pairs* (i, j), *such that* $1 \leq i < j \leq n$ *and the following are satisfied.*

1. Watson-Crick or GU wobble pairs: *If* (i, j) *belongs to S, then pair* (a_i, a_j) *must be one of the following canonical base pairs:* (A, U), (U, A), (G, C), (C, G), (G, U), (U, G).
2. Threshold requirement: *If* (i, j) *belongs to S, then* $j - i > \theta$, *where* θ, *generally taken to be equal to* 3, *is the minimum number of unpaired bases in a hairpin loop; i.e., there must be at least* θ *unpaired bases in a hairpin loop.*
3. Nonexistence of pseudoknots: *If* (i, j) *and* (k, ℓ) *belong to S, then it is not the case that* $i < k < j < \ell$.
4. No base triples: *If* (i, j) *and* (i, k) *belong to S, then* $j = k$; *if* (i, j) *and* (k, j) *belong to S, then* $i = k$.

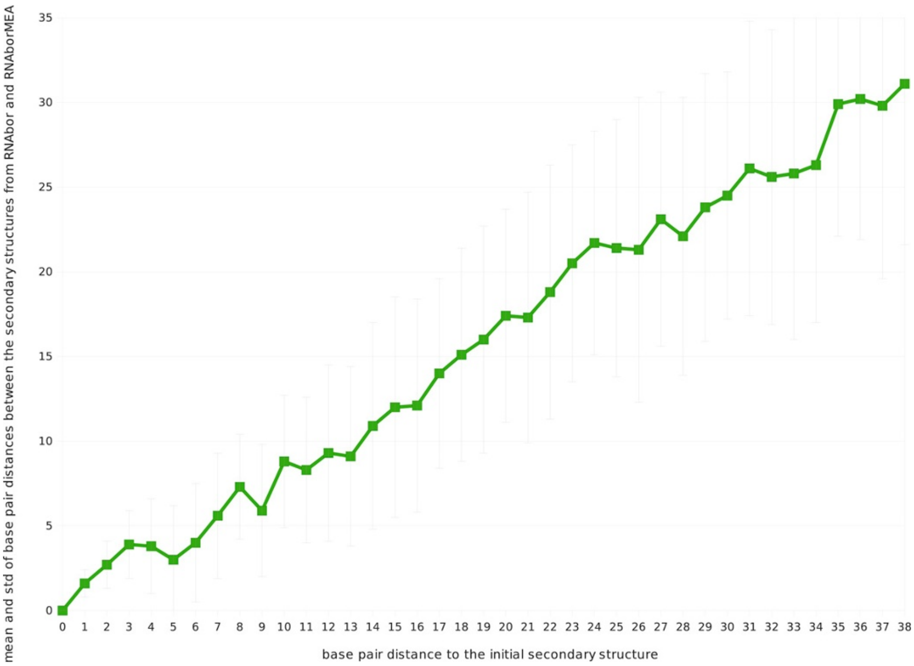


Figure 7 Figure depicting the increasing divergence between `RNAbor`**and** `RNAborMEA`. For each RNA sequence in the seed alignment from Rfam family RF00066 of U7 small nuclear RNAs, both `RNAbor` and `RNAborMEA` were run, in order to compute the *MFE*(*k*) structure and the *MEA*(*k*) structure, which have *minimum free energy* resp. *maximum expected accuracy* among all *k*-neighbors of the intial minimum free energy structures S_0. We computed the base pair distance between the *MFE*(*k*) structure and the *MEA*(*k*) structure over all sequences in the seed alignment of RF00066. The figure displays the average ± one standard deviation of base pair distance.

The preceding definition provides for an inductive construction of the set of all secondary structures for a given RNA sequence $a_1, .. ., a_n$. For all values of $d = 0, .. ., n$ and all values of $i = 1, .. ., n - d$, the collection $\mathbb{S}_{i,i+d}$ of all secondary structures for $a_i, .. ., a_{i+d}$ is defined as follows. If $0 \leq d \leq \theta$, then $\mathbb{S}_{i,i+d} = \{\emptyset\}$; i.e., the only secondary structure for $a_i, .. ., a_{i+d}$ is the empty structure containing no base pairs (due to the requirement that all hairpins contain at least θ unpaired bases). If $d > \theta$ and $\mathbb{S}_{i,j}$ has been defined by recursion for all $i \leq j <$

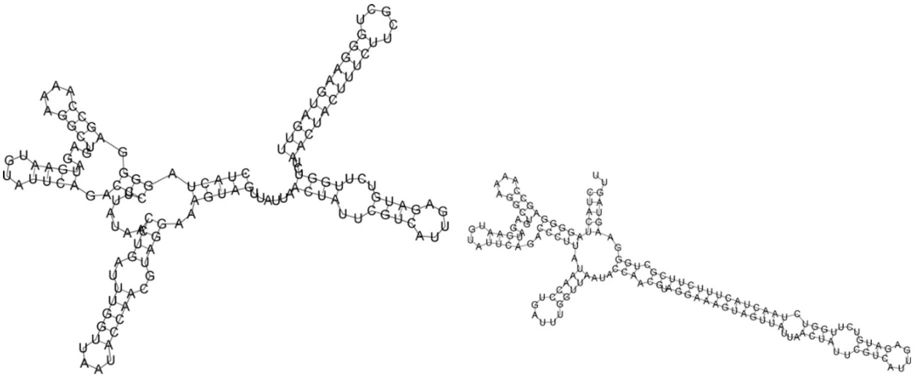


Figure 8 Sample outputs from `RNAborMEA`**on a 143 nt TPP-riboswitch, AF269819/1811-1669 with sequence CUACUAGGGG AGCCAAAAGG CUGAGAUGAA UGUAUUCAGA CCCUUAUAAC CUGAUUUGGU UAAUACCAAC GUAGGAAAGU AGUUAUUAAC UAUUCGUCAU UGAGAUGUCU UGGUCUAACU ACUUUCUUCG CUGGGAAGUA GUU**. We took the TPP riboswitch aptamer from the Rfam database [19], then extracted right-flanking nucleotides from the corresponding EMBL file, in order to include the expression platform. Displayed from left to right are the structures *MEA*(0) and *MEA*(61) (the structure *MEA*(52) is similar to that of *MEA*(61) and corresponds to a free energy local minimum in the left figure.) The structure *MEA*(61) had the highest MEA score over all structural neighbors, including the original structure S_0 = *MEA*(0), and had free energy, -46.0 kcal/mol, that was equal to that of the initial structure S_0 = *MEA*(0), which is the minimum free energy structure for the given sequence.

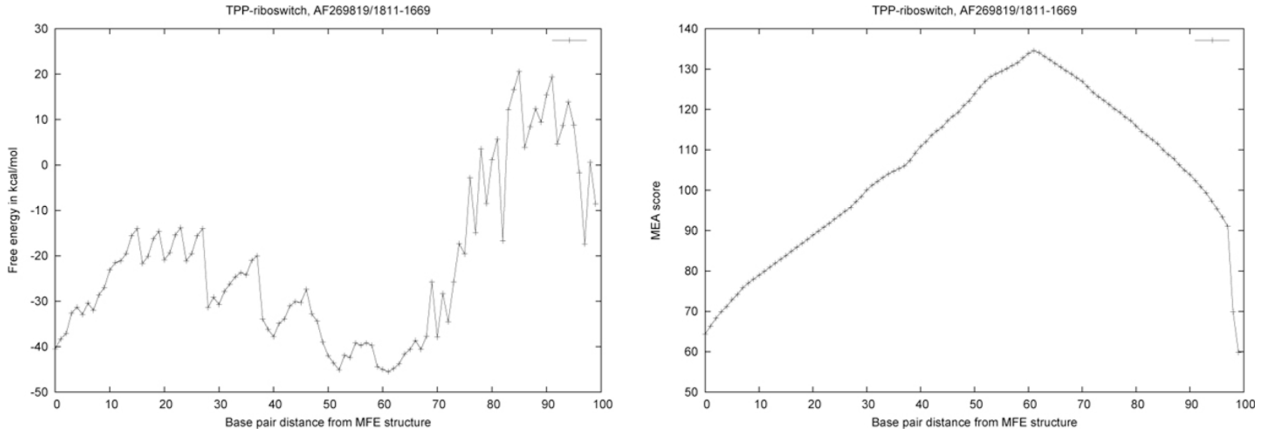
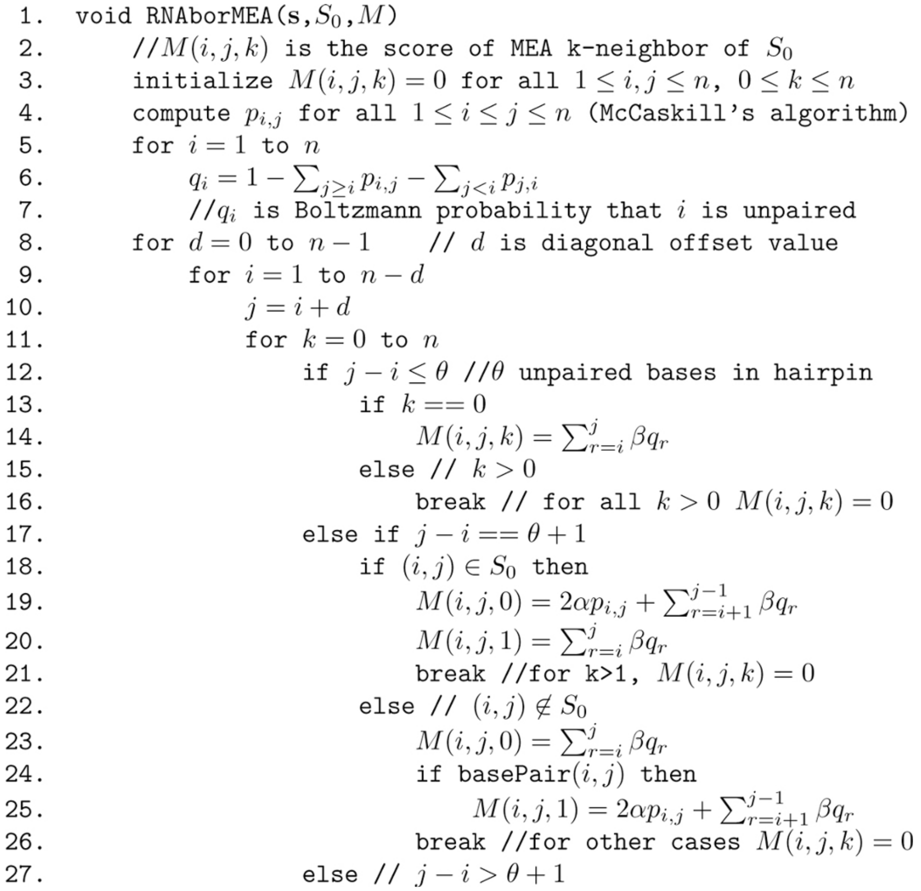


Figure 9 ***(Left)*** **Free energy for all *MEA*(*k*) structural neighbors, 0 ≤ *k* ≤ 99, of the TPP-riboswitch, AF269819/1811-1669, described in the previous figure**. Clearly, *MEA*(0) and *MEA*(61) have the least energy, - 46.0 kcal/mol, and *MEA*(61) has the largest MEA score, 134.555, of all secondary structures for the given RNA sequence. It is more common that the free energy of the *MEA*(*k*) structure is monotonically increasing as a function of *k*. *(Right)* MEA score for all *MEA*(*k*) structural neighbors, 0 ≤ *k* ≤ 99, of the TPP-riboswitch, AF269819/1811-1669, described in the previous figure. Clearly, *MEA*(61) has the largest MEA score, 134.555, of all secondary structures for the given RNA sequence.



```
void RNAborMEA(s,$S_0$,$M$)
    //$M(i,j,k)$ is the score of MEA k-neighbor of $S_0$
    initialize $M(i,j,k) = 0$ for all $1 \le i,j \le n$, $0 \le k \le n$
    compute $p_{i,j}$ for all $1 \le i \le j \le n$ (McCaskill's algorithm)
    for $i = 1$ to $n$
        $q_i = 1 - \sum_{j \ge i} p_{i,j} - \sum_{j<i} p_{j,i}$
        //$q_i$ is Boltzmann probability that $i$ is unpaired
    for $d = 0$ to $n-1$     // $d$ is diagonal offset value
        for $i = 1$ to $n - d$
            $j = i + d$
            for $k = 0$ to $n$
                if $j - i \le \theta$ //$\theta$ unpaired bases in hairpin
                    if $k == 0$
                        $M(i,j,k) = \sum_{r=i}^{j} \beta q_r$
                    else // $k > 0$
                        break // for all $k > 0$ $M(i,j,k) = 0$
                else if $j - i == \theta + 1$
                    if $(i,j) \in S_0$ then
                        $M(i,j,0) = 2\alpha p_{i,j} + \sum_{r=i+1}^{j-1} \beta q_r$
                        $M(i,j,1) = \sum_{r=i}^{j} \beta q_r$
                        break //for k>1, $M(i,j,k) = 0$
                    else // $(i,j) \notin S_0$
                        $M(i,j,0) = \sum_{r=i}^{j} \beta q_r$
                        if basePair$(i,j)$ then
                            $M(i,j,1) = 2\alpha p_{i,j} + \sum_{r=i+1}^{j-1} \beta q_r$
                        break //for other cases $M(i,j,k) = 0$
                else // $j - i > \theta + 1$
```

Figure 10 Initial portion of pseudocode for `RNAborMEA`**algorithm, which continues in Figure 11**. Given RNA sequence **s** = s_1, .. .,s_n of length *n*, initial secondary structure S_0 of **s**, `RNAborMEA` computes for all values of 0 ≤ *k* ≤ *n* that structure *S* with base pair distance *k* from S_0, which maximizes the value $M(i,j,k) = \sum_{(i,j)\in S} 2\alpha p_{i,j} + \sum_{i\text{unpaired in } s} \beta q_i$. The pseudocode actually computes only values *M*(*i*, *j*, *k*) for all *i*, *j*, *k*; the MEA structures are obtained by backtracing. This algorithm clearly runs in $O(n^5)$ time with $O(n^3)$ space.

```
else // $j - i > \theta + 1$
    $\max = 0$ // $M(i,j,k) = \max$ of following
    // Case 1:  $j$ unpaired in $S[i,j]$
    $b_0 = d_{BP}(S_0[i,j-1], S_0[i,j])$
    //$b_0 = 1$ if $j$ paired in $S_0[i,j]$, else $0$
    $val = M(i,j-1,k-b_0) + \beta q_j$
    if $val > \max$ then
        $\max = val$
        index $= (0,0,0)$
        //backtracking:  $j$ unpaired
    // Case 2:  $(i,j) \in S$
    if $basePair(i,j)$ //check if $i,j$ can pair
        $b_1 = d_{BP}(S_0[i+1,j-1] \cup \{(i,j)\}, S_0[i,j])$
        $val = M(i+1,j-1,k-b_1) + 2\alpha p_{i,j}$
        if $val > \max$ then
            $\max = val$
            index $= (i, k-b_1, 0)$
            //backtracking:  $(i,j) \in S$
    // Case 3:  $(r,j) \in S$ for some $i < r < j$
    for $r = i+1$ to $j - \theta - 1$
        if $basePair(r,j)$
            $b_2 = d_{BP}(S_0[i,r-1] \cup S_0[r+1,j-1] \cup \{(r,j)\}, S_0[i,j])$
            for $k_0 = 0$ to $k - b_2$
                $k_1 = k - b_2 - k_0$  //$k_0 + k_1 + b_2 = k$
                $val = M(i,r-1,k_0) + M(r+1,j-1,k_1) + 2\alpha p_{r,j}$
                if $val > \max$ then
                    $\max = val$
                    index $= (r, k_0, k_1)$
                    //backtracking:  $(r,j) \in S$
    $M(i,j,k) = \max$
    $M(j,i,k) =$ index
```

Figure 11 Pseudocode for `RNAborMEA`**algorithm**. Given RNA sequence **s** = $s_1, .. ., s_n$ of length *n*, initial secondary structure S_0 of **s**, `RNAborMEA` computes for all values of $0 \leq k \leq n$ that structure *S* with base pair distance *k* from S_0, which maximizes the value $M(i,j,k) = \sum_{(i,j)\in S} 2\alpha p_{i,j} + \sum_{i\text{unpaired in } s} \beta q_i$. The pseudocode actually computes only values *M*(*i*, *j*, *k*) for all *i*, *j*, *k*; the MEA structures are obtained by backtracing. This algorithm clearly runs in $O(n^3)$ time with $O(n^3)$ space.

```
void $traceback(i,j,k,M,paren)$
    //perform traceback on $[i,j]$ for $k$-neighbors of $S_0[i,j]$
    if $j - i > \theta$ and $M(i,j,k) > 0$
        $(r, k_0, k_1) = M(j,i,k)$
        if $r > 0$ //$j$ pairs with $r$ in $[i,j]$
            $paren[r] =$ '('
     //note that paren has dummy char '$' at position 0
            $paren[j] =$ ')'
            $traceback(r+1, j-1, k1, M, paren)$
            $traceback(i, r-1, k_0, M, paren)$
        else //$r = 0$, so $j$ not paired in $[i,j]$
            $traceback(i, j-1, k_0, M, paren)$
    return
```

Figure 12 Pseudocode for the $O(n^2)$ traceback computed by our `RNAborMEA`**algorithm**. Note that run time could be reduced to *O*(*n* ln *n*) by applying the *boustrephedonic* method described in [42].

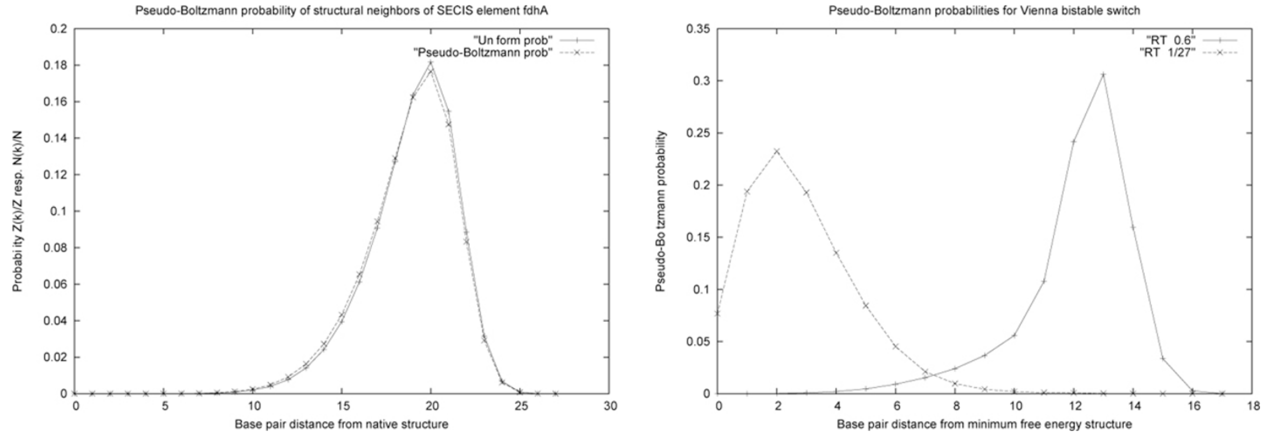


Figure 13 ***(Left)*** **Pseudo-Boltzmann and uniform probabilities of structural neighbors *MEA*(*k*) for the 49 nt SECIS sequence fdhA, with nucleotide sequence CGCCACCCUG CGAACCCAAU AAUAAAAUAU ACAAGGGAGC AAGGUGGCG and where S_0 is (((((((.(((... (((................))).))).)))))))**. Here, the (formal) parameter *RT* taken to be 49 (length of sequence), in order to uniformize MEA scores to range between 0 and 1. The pseudo-Boltzmann probability is defined by $P_b(k) = \frac{Z^{(k)}}{Z}$, where *(i)* $Z^{(k)} = \Sigma\exp(MEA(S)/RT)$, the sum being taken over all *S* such that $d_{BP}(S_0, S) = k$, and *(ii)* $Z = \Sigma_k Z^{(k)}$. The uniform probability is defined by $P_u(k) = \frac{N^{(k)}}{N}$, where $N^{(k)}$ is the number of *k*-neighbors of S_0 and *N* is the total number of secondary structures. *(Right)* Pseudo-Boltzmann probabilities for *MEA*(*k*) structural neighbors of the 27 nt Vienna bistable switch with nucleotide sequence CUUAUGAGGG UACUCAUAAG AGUAUCC and initial (minimum free energy) structure.......((((((((....)))))))). The left curve is when *RT* = 0.6, the approximate value obtained by multiplying the universal gas constant 0.00198 kcal/mol times 310 Kelvin. In contrast, the right curve is when *RT* = 27 (length of sequence). Though not shown in this graph, the pseudo-Boltzmann distribution is identical with the uniform distribution, when *RT* = *n*, where *n* is sequence length.

$i + d$, then

- Any secondary structure of $a_i, .. ., a_{i+d-1}$ is a secondary structure for $a_i, .. ., a_{i+d}$, in which a_{i+d} is unpaired.
- If a_i, a_j can form a Watson-Crick or wobble base pair, then for any secondary structure *S* for $a_{i+1}, .. ., a_{i+d-1}$, the structure $S \cup \{(i, j)\}$ is a secondary structure for $a_i, ..., a_{i+d}$.
- For any intermediate value $i + 1 \leq r \leq j - \theta - 1$, if a_r, a_j can form a Watson-Crick or wobble base pair, then for any secondary structure *S* for $a_i, .. ., a_{r-1}$ and any secondary structure *T* for $a_{r+1}, ..., a_{j-1}$, the structure $S \cup T \cup \{(r, j)\}$ is a secondary structure for $a_i, .. ., a_{i+d}$.

Table 1 Number of samples needed for high-confidence approximation of Boltzmann probabilities

P	*K*	*ε*	*z*	*N*
0.05	1	0.01	1.45	9506
0.05	100	0.01	3.48	30276
0.05	1000000	0.01	5.45	74256
0.001	100	0.01	3.89	37830
0.000001	100	0.01	5.73	82082
0.05	1	0.001	1.45	950600
0.05	100	0.001	3.48	3027600

The number $N = N(\varepsilon, p, K) = \frac{\Phi^{-1}\left(\frac{p}{2K}\right)^2}{4\varepsilon^2}$ of samples sufficient to guarantee that $|f_k - p_k| < \varepsilon$ with confidence 1 - *p*, for all $0 \leq k < K$, in the application `RNAbor-Sample`. Here $p_k = \frac{Z_k}{Z}$, the Boltzmann probability, as computed exactly by `RNAbor`, for a *k*-neighbor of S_0, and f_k is the relative frequency of *k*-neighbors among *N* sampled structures.

Given two secondary structures *S*, *T*, we define the *base pair distance* between *S*, *T*, denoted by $d_{BP}(S, T)$, to be the cardinality of the symmetric difference of *S*, *T*; i.e., $d_{BP}(S, T) = |(S - T) \cup (T - S)|$.

RNAbor-Sample

In this section, we describe how sampling from the Boltzmann ensemble, using `Sfold`[30], leads to a simple and fast approximation of RNAbor computations, provided that the input consists of an RNA sequence and the minimum free energy (MFE) secondary structure for that sequence. The work of this section is inspired by sampling approximations of the number of structural motifs, such as hairpins, multiloops, etc. of Ding and Lawrence [30], as well as the sampling approximation used in `RNAshapes`[8] for large sequences. A novelty of our work is that we provide a rigorous justification for the accuracy of the approximation, depending on sample size.

Let `RNAbor-Sample` denote the protocol, where we apply `Sfold`[30] to sample *N* secondary structures *S* of an input RNA sequence $a_1, .. ., a_n$, then subsequently compute the *relative frequencies* f_k for $0 \leq k < K$, where $f_k = \frac{N_k}{N}$ is defined to be the number N_k of sampled structures *S*, whose base pair distance with S_0 is *k*,

Table 2 Comparison of `NestedAlign` similarity scores for the GENE ON structure of the XPT guanine riboswitch

index	EMBL	RNAbor	RNAborMEA	RNAbor-Sample	RNAlocopt	RNAshapes	UNAFold
0	AL591981/205922-205823	-9.0	5.0	-9.0	-8.5	-9.0	-9.0
1	CP000764/271074-271175	-43.5	5.0	-37.5	-44.5	-23.0	-53.0
2	CP000764/308099-308200	-27.0	-18.0	-24.5	-31.5	-25.5	-22.0
3	BA000028/760473-760574	-25.5	-0.5	-36.0	-38.5	-24.5	-31.0
4	CP000557/252200-252301	-9.5	8.5	-9.5	8.5	-10.0	-12.0
5	X83878/168-267	60.0	87.5	57.0	66.0	64.0	59.0
6	BA000004/1593074-1592973	35.0	16.5	-13.5	-21.5	-19.0	-13.5
7	AAOX01000023/19446-19345	-15.0	-2.0	-13.0	-18.5	-13.5	-15.5
8	CP000416/1798040-1798138	5.5	1.5	1.5	12.0	4.5	-4.5
9	CP000721/398929-399026	26.0	24.5	16.5	-20.0	21.5	-32.0
10	BA000028/1103943-1104044	1.0	1.5	2.0	-0.5	0.5	0.5
11	ABDQ01000002/251055-251152	-16.0	-2.5	-16.5	-21.5	-17.5	-22.5
12	AAXV01000026/31334-31233	11.5	6.0	-1.5	-8.5	22.0	-3.0
13	AE016877/298774-298875	-18.5	14.0	-17.5	-34.0	-12.0	-26.5
14	BA000004/676475-676576	-28.5	-31.0	-28.0	-69.0	-21.0	-29.5
15	AE017333/692981-693082	-1.5	2.5	-11.5	-9.5	-5.5	-53.0
16	AM180355/256217-256318	-17.0	-45.0	-45.5	-49.0	-48.0	-49.0
17	AM406671/1321062-1320965	-25.5	-15.0	-22.0	-28.5	-23.5	-23.5
18	CP000612/2598111-2598012	-42.0	-39.5	-42.0	-47.5	-39.0	-38.5
19	CP000002/697032-697134	-8.0	-11.0	-10.5	-10.0	-4.5	-7.5
20	CP000002/2295936-2295837	23.5	47.0	31.5	21.0	30.0	22.5
21	AL596170/223345-223246	-0.5	7.0	0.5	-8.5	-10.0	-10.0
22	ABDQ01000005/131908-131807	-33.0	-15.5	-31.5	-31.5	-19.0	-50.0
23	AAOX01000052/9069-8968	-13.5	1.5	-14.0	-21.0	-15.5	-14.5
24	AE017333/4024324-4024425	-29.5	-26.5	-33.5	-24.0	-23.5	-36.0
25	AP006627/1554717-1554818	-31.5	-1.5	-37.0	-44.5	-28.5	-43.5
26	CP000024/1182948-1183043	-0.5	-18.5	-9.0	4.0	2.0	-19.0
27	BA000028/786767-786867	-18.0	-41.5	-48.0	-46.5	-49.0	-44.5
28	ABDP01000002/29688-29587	-34.5	-42.5	-34.5	-37.0	-35.0	-50.0
29	BA000043/272473-272574	-9.5	4.0	-9.5	-10.0	-3.0	-12.5
30	CP000724/944285-944386	-30.5	-21.5	-30.5	-28.5	-26.5	-31.5
31	CP000764/1409725-1409826	14.0	-3.0	-18.0	-24.0	-11.5	-20.0
32	AAEK01000017/86437-86538	-44.5	-44.0	-41.5	-52.0	-35.0	-49.0
33	CP000764/357645-357544	11.0	-13.5	-33.0	-26.0	-18.5	-36.0

`NestedAlign` similarity scores between the GENE ON structure of the XPT guanine riboswitch of *B. subtilis*, experimentally determined using in-line probing (see [35]), and the structurally most similar secondary structure among near-optimal structures generated by each of the following six methods: `RNAbor`, `RNAborMEA`, `RNAbor-Sample`, `RNAlocopt`, `RNAshapes`, `UNAFold`. These values are plotted in Figure 5, where more details on the computational experiment are given.

divided by N. Since `Sfold` appears to be only available as a web server, where the user can not stipulate the number of sampled structures, we instead use the Vienna RNA Package implementation of `Sfold`, given in `RNAsubopt -p`[32].

Let $a_1, \ldots, a_n$ be an arbitrary RNA sequence having MFE structure of S_0. Following [9], let Z_k denote the sum of Boltzmann factors of all k-neighbors of S_0; i.e.,

$$Z_k = \sum_{d_{\mathrm{BP}}(S_0,S)=k} \exp(-E(S)/RT).$$

As usual, let Z denote the partition function, representing the sum of Boltzmann factors of all secondary structures of $a_1, \ldots, a_n$; i.e.,

$$Z = \sum_{S} \exp(-E(S)/RT)$$

and let $p_k = \frac{Z_k}{Z}$ denote the Boltzmann probability of all k-neighbors.

Given a desired approximation accuracy of ε, a probability p, and an upper bound K, we wish to compute a value $N = N(\varepsilon, p, K)$, such that whenever we sample at

Table 3 Comparison of NestedAlign similarity scores for the GENE OFF structure of the XPT guanine riboswitch

Index	EMBL	RNAbor	RNAborMEA	RNAbor-Sample	RNAlocopt	RNAshapes	UNAFold
0	AL591981/205922-205823	27.5	28.5	28.5	25.5	25.5	25.5
1	CP000764/271074-271175	13.0	12.5	11.0	6.5	12.0	5.5
2	CP000764/308099-308200	24.0	26.0	26.5	23.0	24.5	26.5
3	BA000028/760473-760574	18.5	22.0	13.0	20.5	23.5	23.0
4	CP000557/252200-252301	7.0	8.0	7.0	10.0	6.5	4.5
5	X83878/168-267	143.0	143.5	143.0	141.0	143.0	141.0
6	BA000004/1593074-1592973	41.0	39.0	41.0	36.0	38.0	41.0
7	AAOX01000023/19446-19345	47.5	45.5	46.0	42.5	34.0	43.5
8	CP000416/1798040-1798138	17.5	12.5	12.5	13.0	11.5	12.5
9	CP000721/398929-399026	36.5	20.5	23.0	-38.5	34.5	-52.5
10	BA000028/1103943-1104044	32.0	29.5	32.0	27.5	30.5	30.0
11	ABDQ01000002/251055-251152	27.0	26.0	26.5	24.0	25.5	7.5
12	AAXV01000026/31334-31233	37.5	38.5	38.0	32.5	35.0	36.0
13	AE016877/298774-298875	24.0	25.5	23.0	19.0	23.0	22.5
14	BA000004/676475-676576	9.0	4.5	6.5	-35.5	5.0	9.0
15	AE017333/692981-693082	-30.0	-9.5	-23.5	-25.5	-17.0	-70.5
16	AM180355/256217-256318	-23.5	-24.0	-25.0	-27.0	-23.5	-27.0
17	AM406671/1321062-1320965	-0.5	3.5	1.0	-10.0	1.0	0.5
18	CP000612/2598111-2598012	-12.0	-9.0	-8.0	-8.5	-9.5	-9.0
19	CP000002/697032-697134	16.5	7.0	12.0	14.0	16.5	7.5
20	CP000002/2295936-2295837	75.0	73.0	75.5	71.0	72.0	69.5
21	AL596170/223345-223246	30.5	31.5	30.5	28.5	29.5	29.5
22	ABDQ01000005/131908-131807	12.5	3.0	13.0	10.5	13.5	4.5
23	AAOX01000052/9069-8968	12.5	14.5	13.5	11.0	12.0	12.0
24	AE017333/4024324-4024425	-3.5	2.5	3.5	6.0	-2.5	-1.5
25	AP006627/1554717-1554818	22.5	18.0	22.5	14.5	25.5	12.5
26	CP000024/1182948-1183043	6.0	7.0	6.5	6.0	5.0	6.0
27	BA000028/786767-786867	-23.5	-19.5	-23.0	-24.5	-21.0	-24.0
28	ABDP01000002/29688-29587	3.0	1.0	2.5	1.0	4.5	0.5
29	BA000043/272473-272574	17.5	12.5	12.5	13.5	12.5	11.5
30	CP000724/944285-944386	10.0	11.0	10.5	7.0	12.0	9.5
31	CP000764/1409725-1409826	32.5	36.0	32.0	26.5	35.0	30.5
32	AAEK01000017/86437-86538	11.5	11.5	13.0	8.0	13.0	11.0
33	CP000764/357645-357544	23.5	22.0	24.5	24.0	22.0	22.5

NestedAlign similarity scores between the GENE OFF structure of the XPT guanine riboswitch of *B. subtilis*, experimentally determined using in-line probing (see [35]), and the structurally most similar secondary structure among near-optimal structures generated by each of the following six methods: RNAbor, RNAborMEA, RNAbor-Sample, RNAlocopt, RNAshapes, UNAFold. These values are plotted in Figure 6, where more details on the computational experiment are given.

Table 4 Number of times that the most similar structure was produced

Method	greatest similarity to gene on	greatest similarity to gene off
RNAborMEA	18	11
RNAbor	7	11
RNAbor-Sample	2	8
RNAlocopt	3	2
RNAshapes	5	8
UNAFold	1	3

Number of times that the most similar structure to the GENE ON resp. GENE OFF structure of the *B. subtilis* XPT riboswitch was produced by each of the six methods investigated. Although the test was made with 34 sequences from the seed alignment of Rfam family RF00167 [31], the sums of each column may exceed 34; this is because If two or more methods produced the maximum score, then each was counted. Structural similarity was measured using the NestedAlign structural alignment algorithm [36]. While the GENE OFF structure involves a *terminator loop*, that is often correctly found by thermodynamics-based software, the GENE ON secondary structure, having higher free energy (hence less stable thermodynamically) is less likely to be found using thermodynamics-based approaches.

least N secondary structures from the Boltzmann ensemble using `Sfold`, the relative frequency f_k of k-neighbors sampled is within ε of the probability p_k of k-neighbors, for all $0 \leq k < K$, with confidence level of (1 - p). Formally, this means that for each $0 \leq k < K$,

$$P(|f_k - p_k| < \varepsilon) \geq 1 - p. \tag{4}$$

Consider the value k as *bin number*. For a fixed bin k, let us denote the exact value of $\frac{Z_k}{Z}$ by p_k. If we sample N structures, each falling in the kth bin with probability p_k, then the number of structures in the kth bin is given by the random variable X_k having binomial distribution with mean $N \cdot p_k$ and variance $N \cdot p_k(1 - p_k)$. It follows that the *proportion* $\frac{X_k}{N}$ of structures in the kth bin has mean $\mu = p_k$ and standard deviation $\sigma = \sqrt{\frac{p_k(1-p_k)}{N}} < \frac{1}{2\sqrt{N}}$. To determine minimum sample size sufficient to ensure a certain approximation accuracy with certain confidence interval, we adapt a standard argument from statistics [37] (see equation (24.35) on p. 529), by approximating the binomial distribution by the standard normal distribution using Z-scores.

Before starting, we mention that it will suffice for our intended application of `RNAbor-Sample` to have a precise approximation of those p_k which exceed some modest lower bound, such as $\delta = 0.01$ or $\delta = 0.0001$. Thus we intend to prove that for all $0 \leq k < K$, *if* $p_k \geq \delta$, *then* Equation (4) holds.

Temporarily, we fix k. Let X be a Bernouilli random variable with success probability p_k, corresponding to the indicator random variable that returns 1 if a single sampled secondary structure is a k-neighbor of S_0. Provided that we sample a number N of structures, which satisfies $N \cdot p_k \geq 30$, then the standard normal distribution can be used to approximate the left and right tail of the distribution of Z-scores of sampled *proportions* $f_k = \frac{\sum_{i=1}^{N} X_k}{N}$, defined by

$$z = \frac{x - \mu}{\sigma} = \frac{f_k - p_k}{\sqrt{\frac{p_k(1-p_k)}{N}}} = \frac{\sqrt{N}(f_k - p_k)}{\sqrt{p_k(1 - p_k)}}. \tag{5}$$

Let $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-x^2/2)dx$ denote the cumulative distribution function (CDF) for the standard normal distribution. Given desired confidence interval of $C = 1 - \alpha$, recall that the *critical value* $z_{\alpha/2}$ is defined by

$$z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) = |\Phi^{-1}(\alpha/2)|.$$

If ε is the *margin of error* in the left tail ($-\infty$, $-z_{\alpha/2}$) and right tail ($z_{\alpha/2}$, $+\infty$) for the normal approximation of the binomial distribution, then by a well-known argument (e.g. equation (24.35) on p. 529 of [37]), we have

$$\varepsilon = z_{\alpha/2} \cdot \sqrt{\frac{p_k(1 - p_k)}{N}}.$$

It follows that

$$N = N(\alpha, \varepsilon) = \frac{z_{\alpha/2}^2}{4\varepsilon^2} \geq \frac{z_{\alpha/2}^2}{\varepsilon^2} \cdot p_k(1 - p_k)$$

provides a sufficient lower bound on number of samples necessary to guarantee margin of error ε. Let $\alpha = \frac{p}{K}$ and define

$$N = N(\varepsilon, p, K) = \frac{\Phi^{-1}\left(\frac{p}{2K}\right)^2}{4\varepsilon^2} = \frac{Z^2_{\frac{p}{2K}}}{4\varepsilon^2}. \tag{6}$$

We have just shown that for $N \geq N(\varepsilon, p, K)$, $P(|z| > |\Phi^{-1}\left(\frac{P}{2K}\right)|) < \frac{p}{K}$, hence

$$P\left(\frac{|f_k - p_k|}{\sqrt{\frac{p_k(1-p_k)}{N}}} > |\Phi^{-1}\left(\frac{p}{2K}|\right)\right) < \frac{p}{K}.$$

The following is now a key step. If we have K bins, and we desire to have a small probability p that we are off by more than ε in our estimate of the probability of any bin (in our intended application, the kth bin, for $0 \leq k < K$, is the collection of k-neighbors of S_0, i.e., those structures S, whose base pair distance with S_0 is k) then it suffices that we have a probability $\frac{p}{K}$ that we are off by more than ε in any single bin. Indeed, let Y_k denote the indicator random variable, with value 1 provided that $|f_k - p_k| > \varepsilon$, where f_k is the relative frequency of sampling a k-neighbor of S_0, after sampling N secondary structures, where by Equation (5), N is chosen so that

$$P(|z| > \varepsilon) = P\left(\frac{\sqrt{N}(|f_k - p_k|)}{\sqrt{p_k(1 - p_k)}} > \varepsilon\right) < \frac{p}{K}$$

then

$$P(Y_0 \vee \cdots \vee Y_{K-1}) < K \cdot p/K = p.$$

Putting everything together, we have shown that for given ε, p, K, we can define by defining N

$$N = N(\varepsilon, p, K) = \frac{\Phi^{-1}\left(\frac{p}{2k}\right)^2}{4\varepsilon^2} \tag{7}$$

we have

$$P\left([\exists 0 \leq k < K]\left[\frac{|f_k - p_k|}{\sqrt{\frac{p_k(1-p_k)}{N}}} > \Phi^{-1}\left(\frac{p}{2K}\right)\right]\right) < p$$

We have completed a more rigorous argument using Chernoff bounds, but prefer the exposition given here for simplicity. Note that the same argument, given *verbatim*, can be used for *any binning* procedure. In particular, this approach provides information on sufficient number of samples in order to approximate the result of RNAshapes [8,38,39] by means of sampling.

We can make some basic conclusions from the above analysis. The number of samples sufficient to ensure that $|f_k - p_k| < \varepsilon$ for $0 \leq k < K$ with confidence $1 - p$ is reasonable, and only slightly increases for a higher number K of bins or to ensure greater confidence $1 - p$. However, the number of samples increases greatly when higher precision estimates (smaller ε values) are needed, even for one bin.

In the case of one bin, it is important to remember that the value N is a conservative estimate, sufficient to ensure our conclusion. This estimate uses the worst case of 50% probability of being in a bin, which maximizes the standard deviation. For bins with small probability, one can re-estimate what is needed. However, for bins with smaller probability, higher precision is needed as well, unless all that is required is to verify that the bin has small probability. Also, clearly if a bin has probability of 10^{-6}, then at least on the order of one million samples are needed, just for a reasonable probability of sampling the bin once.

Algorithm description

Given an RNA sequence $a = a_1, .. ., a_n$, a secondary structure S_0 of a, and a maximum desired value $Kmax \leq n$, the RNAborMEA algorithm computes, for each $1 \leq i < j \leq n$ and each $0 \leq k < Kmax \leq n$, the maximum *score* $M(i, j, k)$

$$\sum_{(i,j)\in S} 2\alpha p_{i,j} + \sum_{i \text{unpaired}} \beta q_i$$

where the first sum is taken over all base pairs (i, j) belonging to S, the second sum is taken over all unpaired positions in S, and where $p_{i,j}$ [resp. q_i] is the probability that i, j are paired [resp. i is unpaired] in the ensemble of low energy structures, and $\alpha, \beta > 0$ are weights. Our computational experiments, as in Figure 9, were carried out with default values of 1 for α, β. (See Equation 1 for the formal definition of Boltzmann base pairing probability $p_{i,j}$.)

The dynamic programming computation of $M(i, j, k)$ is performed by recursion on increasing values of $j - i$ for all values $1 \leq i \leq j \leq n$ and $0 \leq k \leq Kmax$. The value of $M(i, j, k)$, stored in the upper triangular portion of matrix M, will involve taking the maximum over three cases, which correspond to the inductive construction of all secondary structures on $a_i, .. ., a_j$, as described in the previous section. At the same time, the value $M(j, i, k)$, stored in the lower triangular portion of matrix M, will consist of a triple r, k_0, k_1 of numbers, such that the following *approximately* holds (in this section, we provide the motivating idea; the actual algorithm description, which deviates slightly from the description here, is given in the next section and in Figures 10 and 11). *(i)* If $r = 0$ then $M(i, j, k)$ is maximized by a k-neighbor S of $S_0[i, j]$ for the subsequence $a_i, .. ., a_j$ in which a_j is unpaired. In this case, $k_0 = k$ and $k_1 = 0$. *(ii)* If $r = i$, then $M(i, j, k)$ is maximized by a k-neighbor S of $S_0[i, j]$ for the subsequence $a_i, ...,a_j$ in which base pair $(i, j) \in S$. In this case, $k_0 = 0$ and $k_1 = k - 1$. *(i)* If $i < r \leq j - \theta - 1$ then $M(i, j, k)$ is maximized by a k-neighbor S of $S_0[i, j]$ for the subsequence $a_i, .. .,a_j$ in which base pair $(r, j) \in S$. The left portion of S, which is $S[i, r - 1]$ will be a k_0 neighbor of $S[i, r - 1]$, while the right portion of S, which is $S[r, j]$ must contain the base pair (r, j) and itself be a k_1 neighbor of $S[r, j]$. In summary, the values r, k_0, k_1 will be used in computing the traceback, where the maximum expected accurate structure that is a k-neighbor of $S[i, j]$ will be constructed by one of the following: *(i)* MEA k-neighbor of $S[i, j - 1]$, in the event that a_j is unpaired in $[i, j]$; *(ii)* MEA $k - 1$-neighbor of $S[i + 1, j - 1]$, in the event that a_i, a_j form a base pair; *(iii)* MEA k_0-neighbor of $S[i, r - 1]$ and the MEA k_1-neighbor of $S[r, j]$, where $k_0 + k_1 = k$, in the event that a_r, a_j form a base pair.

Pseudocode for the algorithm RNAborMEA is given in Figures 10 and 11. An array M of size $n \times n \times Kmax$ is required to store the MEA scores in $M(i, j, k)$ for all subsequences $[i, j]$ and all base pair distances $0 \leq k \leq Kmax$ between structures $S[i, j]$ and initially given structure $S_0[i, j]$. For $1 \leq i \leq j \leq n$ and all $0 \leq k \leq Kmax$, the pseudocode in Figure 11 stores a value of the form (x, y, z) in the lower triangular portion, $M(j, i, k)$, of the array. Here, $x = 0$ indicates that the optimal structure on $[i, j]$, i.e., having maximum MEA score over all k-neighbors of $S_0[i, j]$, is obtained by not pairing j with any nucleotide in $[i, j]$; for values $x > 0$, hence $x \in [i, j - \theta - 1]$, the optimal k-neighbor of $S_0[i, j]$ is obtained by pairing x with j. The values y, z correspond to the values k_0, k_1, such that: *(i)* if $x = 0$, then the optimal k-neighbor of $S_0[i, j]$ is obtained by first computing the optimal k_0-neighbor of $S_0[i, j - 1]$, where $k_0 = k - b_0$, then leaving j unpaired; *(ii)* if $x = i$, then the optimal k-neighbor of $S_0[i, j]$ is obtained by first computing the optimal k_1-neighbor of $S_0[i + 1, j - 1]$, where $k_1 = k - b_1$, then adding the enclosing base pair (i, j); *(iii)* if $x = r \in [i + 1, j - \theta - 1]$, then the optimal k-neighbor of $S_0[i, j]$ is obtained by first computing the optimal k_0-neighbor of $S_0[i, r - 1]$ as well as the optimal k_1-neighbor of $S_0[r + 1, j - 1]$, then adding the base pair (r, j). This last calculation must be done over all values k_0, k_1 such that $k_0 + k_1 = k$. Using the values $M(j, i, k) = (x, y, z)$, the traceback can be easily computed by recursion; see Figure 12 for pseudocode of traceback.

In a manner similar to the pseudocode of Figures 10 and 11 (essentially, one replaces the operation of taking the *maximum* by the a summation, and one replaces the MEA score by the pseudo-Boltzmann factor exp($MEA(S)/RT$)), RNAborMEA also computes the *pseudo-Boltzmann* partition function values

$$Z^{(k)}_{i,j} = \sum_{\{S \in \mathbb{S}_{i,j}: d_{BP}(S_0,S)=k\}} \exp(\mathrm{MEA(S/RT)}).$$

We have graphed the Boltzmann probabilities $\frac{Z^{(k)}_{1,n}}{Z_{1,n}}$ as well as the uniform probabilities $\frac{N^{(k)}_{1,n}}{N_{1,n}}$, where $N^{(k)}_{1,n}$ is the number of k-neighbors, and $N_{1,n}$ is the total number of secondary structures. When $RT = n$, which normalizes the MEA score to a maximum of 1, it appears that the Boltzmann distribution is the *same* as the uniform distribution, as shown in Figure 13.

Acknowledgements

Funding for the research of P. Clote and F. Lou was provided by the Digiteo Foundation for the project *RNAomics*. Additional funding was provided to P. Clote by National Science Foundation grants DMS-1016618 and DMS-0817971. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
This article has been published as part of *BMC Bioinformatics* Volume 13 Supplement 5, 2012: Selected articles from the First IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/13/S5.

Author details

[1]Department of Biology, Boston College, Chestnut Hill, MA 02467, USA. [2]Laboratoire de Recherche en Informatique (LRI), Université Paris-Sud XI, 91405 Orsay cedex, France. [3]Laboratoire d'Informatique (LIX), Ecole Polytechnique, 91128 Palaiseau, France. [4]Department of Mathematics and Computer Science, Denison University, Granville, OH 43023-0810, USA.

Authors' contributions

RNAbor-Sample was conceived by W.A.L., who provided a proof for sample size sufficient to ensure a desired degree of accuracy with a desired level of confidence. RNAborMEA was conceived by P.C., then designed and programmed by P.C. (prototype implementation in Python) and F.L. (implementation in C). Computational experiments were carried out by F.L., figures were created by F.L. and P.C. and the manuscript was written by P.C.

Competing interests

The authors declare that they have no competing interests.

Published: 12 April 2012

References

1. Olsthoorn R, Mertens S, Brederode F, Bol J: **A conformational switch at the 3' end of a plant virus RNA regulates viral replication.** *EMBO J* 1999, **18**:4856-4864.
2. Repsilber D, Wiese S, Rachen M, Schroder A, Riesner D, Steger G: **Formation of metastable RNA structures by sequential folding during transcription: time-resolved structural analysis of potato spindle tuber viroid (-)-stranded RNA by temperature-gradient gel.** *RNA* 1999, **5**:574-584.
3. Franch T, Gultyaev AP, Gerdes K: **Programmed Cell Death by hok/sok of Plasmid R1: Processing at the hok mRNA 3H-end Triggers Structural Rearrangements that Allow Translation and Antisense RNA Binding.** *J Mol Biol* 1997, **273**:38-51.
4. Mandal M, Boese B, Barrick J, Winkler W, Breaker R: **Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria.** *Cell* 2003, **113(5)**:577-586.
5. 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.
6. Ray PS, Jia J, Yao P, Majumder M, Hatzoglou M, Fox PL: **A stress-responsive RNA switch regulates VEGFA expression.** *Nature* 2009, **457(7231)**:915-919.
7. Voss B, Meyer C, Giegerich R: **Evaluating the predictability of conformational switching in RNA.** *Bioinformatics* 2004, **20(10)**:1573-1582.
8. Voss B, Giegerich R, Rehmsmeier M: **Complete probabilistic analysis of RNA shapes.** *BMC Biol* 2006, **4(5)**.
9. Freyhult E, Moulton V, Clote P: **Boltzmann probability of RNA structural neighbors and riboswitch detection.** *Bioinformatics* 2007, **23(16)**:2054-2062, Doi: 10.1093/bioinformatics/btm314.
10. Barash D: **Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation.** *Bioinformatics* 2004, **20(12)**:1861-1869.
11. Mandal M, Breaker RR: **Adenine riboswitches and gene activation by disruption of a transcription terminator.** *Nat Struct Mol Biol* 2004, **11**:29-35.
12. Serganov A, Yuan Y, Pikovskaya O, Polonskaia A, Malinina L, Phan A, Hobartner C, Micura R, Breaker R, Patel D: **Structural Basis for Discriminative Regulation of Gene Expression by Adenine- and Guanine-Sensing mRNAs.** *Chem Biol* 2004, **11(12)**:1729-1741.
13. Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ: **Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch.** *Nature* 2006, **441(7097)**:1167-1171.
14. Abreu-Goodger C, Merino E: **RibEx: A web server for locating riboswitches and other conserved bacterial regulatory elements.** *Nucleic Acids Res* 2005, **33**:W690-W692.
15. Bengert P, Dandekar T: **Riboswitch finder - A tool for identification of riboswitch RNAs.** *Nucleic Acids Res* 2004, **32**:W154-W159.
16. Chang TH, Huang HD, Wu LC, Yeh CT, Liu BJ, Horng JT: **Computational identification of riboswitches based on RNA conserved functional sequences and conformations.** *RNA* 2009, **15(7)**.
17. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25(10)**:1335-1337.
18. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, Neph S, Tompa M, Ruzzo WL, Breaker RR: **Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline.** *Nucleic Acids Res* 2007, **35(14)**:4809-4819.
19. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A: **Rfam: updates to the RNA families database.** *Nucleic Acids Res* 2009, **37(Database)**:D136-D140.
20. Freyhult E, Moulton V, Clote P: **Boltzmann probability of RNA structural neighbors and riboswitch detection.** *Bioinformatics* 2007, **23(16)**:2054-2062.
21. Freyhult E, Moulton V, Clote P: **RNAbor: a web server for RNA structural neighbors.** *Nucleic Acids Res* 2007, **35(Web)**:W305-W309.
22. Matthews D, Sabina J, Zuker M, Turner D: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288**:911-940.
23. Xia T, SantaLucia JJ, Burkard M, Kierzek R, Schroeder S, Jiao X, Cox C, Turner D: **Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs.** *Biochemistry* 1999, **37**:14719-35.
24. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment.** *Genome Res* 2005, **15(2)**:330-340.
25. Nussinov R, Jacobson AB: **Fast Algorithm for Predicting the Secondary Structure of Single Stranded RNA.** *Proceedings of the National Academy of Sciences, USA* 1980, **77(11)**:6309-6313.
26. Kiryu H, Kin T, Asai K: **Robust prediction of consensus secondary structures using averaged base pairing probability matrices.** *Bioinformatics* 2007, **23(4)**:434-441.
27. McCaskill J: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29**:1105-1119.

28. Lu ZJ, Gloor JW, Mathews DH: **Improved RNA secondary structure prediction by maximizing expected pair accuracy.** *RNA* 2009, **15(10)**:1805-1813.
29. Zuker M: **On finding all suboptimal foldings of an RNA molecule.** *Science* 1989, **244(7)**:48-52.
30. Ding Y, Lawrence CE: **A statistical sampling algorithm for RNA secondary structure prediction.** *Nucleic Acids Res* 2003, **31**:7280-7301.
31. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A: **Rfam: Wikipedia, clans and the "decimal" release.** *Nucleic Acids Res* 2011, **39(Database)**:D141-D145.
32. Gruber A, Lorenz R, Bernhart S, Neubock R, Hofacker I: **The Vienna RNA websuite.** *Nucleic Acids Research* 2008, **36**:70-74.
33. 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.** *Proc Natl Acad Sci USA* 2004, **101(19)**:7287-7292.
34. Lorenz W, Clote P: **Computing the partition function for kinetically trapped RNA secondary structures.** *Public Library of Science One (PLoS ONE)* 2011, **6**:316178, Doi:10.1371/journal.pone.0016178.
35. Wakeman CA, Winkler WC, Dann C: **Structural features of metabolite-sensing riboswitches.** *Trends Biochem Sci* 2007, **32(9)**:415-424.
36. Blin G, Denise A, Dulucq S, Herrbach C, Touz H: **Alignments of RNA structures.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2010.
37. Zar J: *Biostatistical Analysis* Prentice-Hall, Inc; 1999.
38. Giegerich R, Voss B, Rehmsmeier M: **Abstract shapes of RNA.** *Nucleic Acids Res* 2004, **32(16)**:4843-4851.
39. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R: **RNAshapes: an integrated RNA analysis package based on abstract shapes.** *Bioinformatics* 2006, **22(4)**:500-503.
40. Hofacker I: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31**:3429-3431.
41. Markham NR, Zuker M: **UNAFold: software for nucleic acid folding and hybridization.** *Methods Mol Biol* 2008, **453**:3-31.
42. Ponty Y: **Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: The boustrophedon method.** *J Math Biol* 2008, **56(1-2)**:107-127.

doi:10.1186/1471-2105-13-S5-S6
Cite this article as: Clote *et al.*: Maximum expected accuracy structural neighbors of an RNA secondary structure. *BMC Bioinformatics* 2012 13 (Suppl 5):S6.