BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 16 2007, pages 2054–2062
doi:10.1093/bioinformatics/btm314

Structural bioinformatics

Boltzmann probability of RNA structural neighbors and riboswitch detection

Eva Freyhult[1], Vincent Moulton[2] and Peter Clote[3,*]

[1]Linnaeus Centre for Bioinformatics, Uppsala University, 75124 Uppsala, Sweden, [2]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK and [3]Department of Biology, Boston College, Chestnut Hill, MA 02467, USA

Received and revised on April 30, 2007; accepted on June 5, 2007

Advance Access publication June 14, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: We describe algorithms implemented in a new software package, RNAbor, to investigate structures in a neighborhood of an input secondary structure $\mathcal{S}$ of an RNA sequence **s**. The input structure could be the minimum free energy structure, the secondary structure obtained by analysis of the X-ray structure or by comparative sequence analysis, or an arbitrary intermediate structure.

Results: A secondary structure $\mathcal{T}$ of **s** is called a δ-neighbor of $\mathcal{S}$ if $\mathcal{T}$ and $\mathcal{S}$ differ by exactly δ base pairs. RNAbor computes the number ($\mathbf{N}^\delta$), the Boltzmann partition function ($\mathbf{Z}^\delta$) and the minimum free energy (MFE^δ) and corresponding structure over the collection of all δ-neighbors of $\mathcal{S}$. This computation is done simultaneously for all $\delta \leq m$, in run time $O(mn^3)$ and memory $O(mn^2)$, where n is the sequence length. We apply RNAbor for the detection of possible RNA conformational switches, and compare RNAbor with the switch detection method paRNAss. We also provide examples of how RNAbor can at times improve the accuracy of secondary structure prediction.

Availability: http://bioinformatics.bc.edu/clotelab/RNAbor/

Contact: clote@bc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In the last few years, there has been intense interest in RNA due to the surprising, previously unsuspected roles played by ribonucleic acid in what until now has been a predominantly protein-centric view of molecular biology. Apart from its roles as messenger RNA and transfer RNA, ribonucleic acid molecules play a catalytic role in the peptidyltransferase reaction in peptide bond formation (Nissen *et al.*, 2000; Weinger *et al.*, 2004) and in intron splicing (Vicens and Cech, 2006), both examples of enzymatic RNAs now termed ribonucleic enzymes or *ribozymes* (Doudna and Cech, 2002). RNA plays a role in post-transcriptional gene regulation due to the hybridization of mRNA by small interfering RNAs (siRNA) (Harborth *et al.*, 2003; Tuschl, 2003) and microRNAs (miRNA) (Lim *et al.*, 2003). By completely different means, RNA performs transcriptional and translational gene regulation by allostery, where a portion of the 5′ untranslated region (5′ UTR) of mRNA known as a *riboswitch* (Penchovsky and Breaker, 2005; Winkler *et al.*, 2002) can undergo a conformational change upon binding a specific ligand, such as adenine, guanine or lysine. RNA is known to play critical roles in various other cellular mechanisms, such as dosage compensation (Brown *et al.*, 1992), protein shuttling (Walter and Blobel, 1982), expansion of the genetic code such as selenocysteine insertion (Commans and Böck, 1999), and ribosomal frameshift (Bekaert *et al.*, 2003; Moon *et al.*, 2004). Illustrative of the growing recognition for the importance of RNA, the 2006 Nobel Prize in Physiology or Medicine was awarded to A.Z. Fire and C.C. Mello for their discovery of RNA interference and gene silencing by double-stranded RNA.

In this article, we develop novel and efficient algorithms to investigate structures in a neighborhood of a given secondary structure $\mathcal{S}$ of an RNA sequence **s**. We call another secondary structure $\mathcal{T}$ of s a δ-neighbor of $\mathcal{S}$, if $\mathcal{T}$ and $\mathcal{S}$ differ by exactly δ base pairs (see Methods section for more details). We develop algorithms to compute the number, $\mathbf{N}^\delta = \mathbf{N}^\delta(\mathbf{s}, \mathcal{S})$, the partition function, $\mathbf{Z}^\delta = \mathbf{Z}^\delta(\mathbf{s}, \mathcal{S}) = \sum_{\mathcal{T}} \exp(-E(\mathcal{T})/RT)$ and the minimum free energy $\mathrm{MFE}^\delta = \mathrm{MFE}^\delta(\mathbf{s}, \mathcal{S})$ structure over the collection of all δ-neighbors $\mathcal{T}$ of $\mathcal{S}$, where $E(\mathcal{T})$ denotes the energy of $\mathcal{T}$ with respect to the Turner nearest neighbor energy model (Xia *et al.*, 1999), R is the universal gas constant and T is temperature in Kelvin. Our software, called RNAbor (RNA neighbor), additionally computes graphs of the probability density function $p^\delta = \mathbf{Z}^\delta/\mathbf{Z}$ as a function of δ.

RNAbor was motivated by Moulton *et al.* (2000), who suggested that the stability of a secondary structure might depend on the number of structural neighbors at varying distances from the given structure—for instance from the minimum free energy (MFE) structure. It turns out that the number of structural neighbors at varying distances is not sufficient to distinguish between structural RNA and random RNA having the same MFE structure; see Figure 1. However, we do see a distinction when computing a *weighted* count of structural neighbors, where low energy structures are more heavily weighted. Formally, this is the Boltzmann partition function with respect to all structural neighbors at a given base pair distance δ. Figure 1 displays a density plot produced by RNAbor which clearly suggests that precursor miRNA

*To whom correspondence should be addressed.

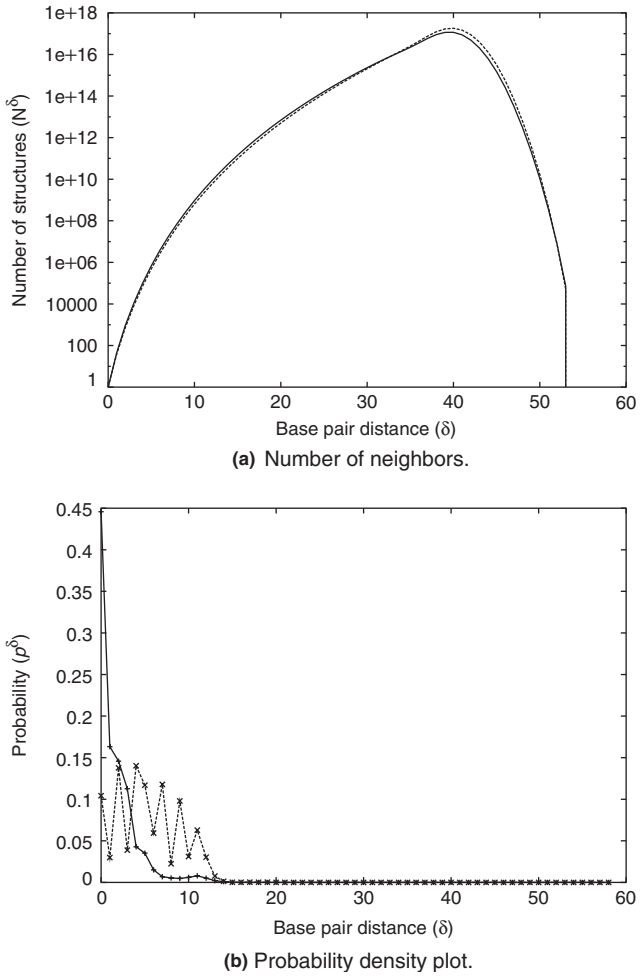


Fig. 1. The number and probability of neighboring structures at varying distances from the MFE structure of the precursor miRNA `dme-mir-1 (AE003667.3/4813-4888)` from *Drosophila melanogaster* (Adams *et al.*, 2000) (solid line) and a random RNA having the same secondary structure (dashed line). The random RNA was obtained by running `RNAinverse` (Hofacker *et al.*, 1994) with the MFE structure of `dme-mir-1` as input structure.

`dme-mir-1` from *D.melanogaster* (AE003667) has a single, well-defined native structure, whereas a random sequence with the same MFE structure has several alternate low energy secondary structures with different topologies.

We show that the probability density plot can be used to detect RNA conformational switches and compare `RNAbor` with the conformational switch detection program `paRNAss` (Giegerich *et al.*, 1999). In `paRNAss`, a structural RNA switch is predicted by means of studying properties of the energy landscape of the RNA. Secondary structures are sampled from the structure space using `RNAsubopt` (Wuchty *et al.*, 1999) or `mfold` (Zuker, 1994). Pairwise distances are calculated between the sampled structures using two different distance measures (e.g. pairs energy barrier, morphological, tree alignment or string edit distance). Using a standard clustering method the structures are clustered into two clusters based on the distance measures. If the RNA is a conformational switch, then it has two stable structures and hence two clusters are expected (in a multi-switch, more than two stable structures are expected). As an additional test, the consensus structure of the clusters are computed and for each sample structure the distances to the two consensus structures are plotted against each other. If the RNA is really a conformational switch, then `paRNAss` output should display two clouds of points—one near the x-axis and one near the y-axis.

Note that since `paRNAss` calls `RNAsubopt` from the Vienna RNA Package program (Hofacker, 2003), it requires a user-defined energy bound, E, in order to generate all secondary structures within E kcal/mol of the MFE.

The plan for the rest of this article is as follows. In the Methods section, we describe the algorithms for computing the number $\mathbf{N}^\delta$, the Boltzmann partition function $\mathbf{Z}^\delta$ and the minimum free energy MFE^δ structure over the collection of all δ-neighbors. In the Results section, we present graphs of the number $\mathbf{N}^\delta$ and the Boltzmann probability density $p^\delta = \mathbf{Z}^\delta/\mathbf{Z}$ of structural neighbors, which differ by δ base pairs from a given secondary structure. In addition, we compare the output of `RNAbor` with the conformational switch detection program `paRNAss` (Giegerich *et al.*, 1999). In the Discussion section, we conclude by presenting some possible future applications of our algorithms. Pseudocode for our implementation is presented in the Supplementary Material. Additional data obtained by running `RNAbor` on all SAM riboswitches from Rfam (Griffiths-Jones *et al.*, 2003) is available in the Supplementary Material at http://bioinformatics.bc.edu/clotelab/RNAbor/webSupplement.

2 MATERIALS AND METHODS

Given an RNA nucleotide sequence **s**, consider a fixed secondary structure $\mathcal{S}$ of **s**. In this section, we describe how to efficiently compute the number $\mathbf{N}^\delta$ of δ-neighbors of $\mathcal{S}$, the partition function $\mathbf{Z}^\delta$ for δ-neighbors, and the minimum free energy MFE^δ together with the corresponding structure over all δ-neighbors. $\mathbf{N}^\delta$, $\mathbf{Z}^\delta$ and MFE^δ are all computed for a fixed temperature. The temperature is set to 37°C by default, but can be changed by the user.

2.1 The number of δ-neighbors of a fixed secondary structure

Let $\mathbf{s} = s_1, \ldots, s_n$ denote an RNA sequence, i.e. a sequence of letters in the alphabet of nucleotides $\{A,C,G,U\}$. A secondary structure $\mathcal{S}$ on **s** is a set of base pairs (i,j), where $1 \leq i \leq i + \theta < j \leq n$ and $\theta \geq 0$ is an integer (corresponding to minimum hairpin loop size, which we usually set to 3), such that if (k,l) is a base pair, then $k{=}i \iff l{=}j$ (a nucleotide is involved in at most one base pair) and $i < k < j \iff i < l < j$ (no pseudoknots). We say that $\mathcal{S}$ is compatible with **s** if for every base pair (i,j) in $\mathcal{S}$ the pair $s_i\, s_j$ is contained in the set $\mathbb{B} = \{AU, UA, GC, CG, GU, UG\}$ (i.e. the set of Watson–Crick base pairs together with wobbles). Given two secondary structures $\mathcal{S}$, $\mathcal{T}$ on **s**, we define the base pair distance d_{BP} between $\mathcal{S}$ and $\mathcal{T}$ to be the number of base pairs that they have that are not in common, i.e.

$$d_{\mathrm{BP}}(\mathcal{S},\mathcal{T}) = |\mathcal{S} \cup \mathcal{T}| - |\mathcal{S} \cap \mathcal{T}|. \quad (1)$$

For the rest of this section, we consider both **s** as well as the secondary structure $\mathcal{S}$ on **s** to be fixed. We now provide recursions for determining the number of secondary structures $\mathcal{T}$ compatible with **s** that are at precisely base pair distance δ to $\mathcal{S}$.

Let $\mathcal{S}_{[i,j]}$ denote the restriction of $\mathcal{S}$ to interval $[i,j]$ of $\mathbf{s}$, i.e. the set of base pairs $\mathcal{S}_{[i,j]} = \{(k,l) : i \leq k < l \leq j, (k,l) \in \mathcal{S}\}$. A secondary structure $\mathcal{T}_{[i,j]}$ on $\mathbf{s}$ is a δ-*neighbor* of $\mathcal{S}_{[i,j]}$ if $d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{T}_{[i,j]}) = \delta$. For all $0 \leq \delta \leq m$, and all $1 \leq i \leq j \leq n$, let $\mathbf{N}^{\delta}_{i,j}(\mathbf{s}, \mathcal{S})$ denote the number of secondary structures $\mathcal{T}_{[i,j]}$ compatible with $\mathbf{s}$ such that $d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{T}_{[i,j]}) = \delta$. In the following, we may omit the sequence s and secondary structure $\mathcal{S}$ in our notation since these are fixed. In particular, we put $\mathbf{N}^{\delta}_{i,j} = \mathbf{N}^{\delta}_{i,j}(\mathbf{s}, \mathcal{S})$.

$\mathbf{N}^{\delta}_{i,j}$ is computed recursively. The initial conditions for computing $\mathbf{N}^{\delta}_{i,j}$ are given by

$$\mathbf{N}^{0}_{i,j} = 1, \text{ for } i \leq j, \tag{2}$$

since the only 0-neighbor to a structure is the structure itself, and

$$\mathbf{N}^{\delta}_{i,j} = 0, \text{ for } \delta > 0, i \leq j \leq i + \theta, \tag{3}$$

since the empty structure is the only possible structure for a sequence shorter than $\theta + 2$ nucleotides, and so there are no δ-neighbors for $\delta > 0$. The recursion used to compute $\mathbf{N}^{\delta}_{i,j}$ for $\delta > 0$ and $j > i + \theta$ is

$$\mathbf{N}^{\delta}_{i,j} = \mathbf{N}^{\delta - b_0}_{i,j-1} + \sum_{\substack{s_k s_j \in \mathbb{B}, \\ i \leq k < j}} \sum_{w + w' = \delta - b} \mathbf{N}^{w}_{i,k-1} \mathbf{N}^{w'}_{k+1,j-1}, \tag{4}$$

where $b_0 = 1$ if j is base paired in $\mathcal{S}_{[i,j]}$ and 0 otherwise, and $b = d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{S}_{[i,k-1]} \cup \mathcal{S}_{[k+1,j-1]} \cup \{(k,j)\})$. This holds since in a secondary structure $\mathcal{T}_{[i,j]}$ on $[i,j]$ that is a δ-neighbor of $\mathcal{S}_{[i,j]}$, either nucleotide j is unpaired in $[i,j]$ or it is paired to a nucleotide k such that $i \leq k < j$. In this latter case it is enough to study the smaller sequence segments $[i,k-1]$ and $[k+1,j-1]$ noting that, except for (k,j), base pairs outside of these regions are not allowed. In addition, for $d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{T}_{[i,j]}) = \delta$ to be fulfilled it is necessary for $w + w' = \delta - b$ to hold, where $w = d_{\mathrm{BP}}(\mathcal{S}_{[i,k-1]}, \mathcal{T}_{[i,k-1]})$ and $w' = d_{\mathrm{BP}}(\mathcal{S}_{[k+1,j-1]}, \mathcal{T}_{[k+1,j-1]})$, since b is the number of base pairs that differ between $\mathcal{S}_{[i,j]}$ and a structure $\mathcal{T}_{[i,j]}$, due to the introduction of the base pair (k,j).

Pseudocode for computing $\mathbf{N}^{\delta}_{i,j}$ for values of δ between 0 and m is given in the Supplementary Material. The algorithm runs in time $O(mn^3)$ and space $O(mn^2)$ where, as defined above, n is the length of $\mathbf{s}$ and m is the maximum value of δ.

2.2 Probability analog

In this section, we explain how to extend our approach of computing $\mathbf{N}^{\delta}_{i,j}$ to compute the partition function contribution of the set of structures compatible with a given RNA sequence $\mathbf{s}$ at a fixed base pair distance δ from an RNA structure $\mathcal{S}$ compatible with $\mathbf{s}$. This allows us to compute the probability of the set of structures compatible with $\mathbf{s}$ at distance δ from $\mathcal{S}$.

It is straight–forward to extend the previous approach to compute partition functions for the Nussinov–Jacobson energy model (Nussinov and Jacobson, 1980). In particular, by simply replacing recursion (4) with

$$\mathbf{N}^{\delta}_{i,j} = \mathbf{N}^{\delta - b_0}_{i,j-1} + \sum_{\substack{s_k s_j \in \mathbb{B}, \\ i \leq k < j}} \left(e^{\frac{-E_{(k,j)}}{RT}} \sum_{w + w' = \delta - b} \mathbf{N}^{w}_{i,k-1} \mathbf{N}^{w'}_{k+1,j-1} \right), \tag{5}$$

where $E_{(k,j)}$ is the energy of the base pair (k,j), R is the gas constant and T is the temperature, we can compute the partition function contribution of structures at a given base pair distance δ. The base pair energy $E_{(k,j)}$ takes the value -1 if $s_k\, s_j \in \mathbb{B}$ and 0 otherwise. Note that the energy contribution can be altered for different base pairs e.g. -3 for GC, -2 for AU and -1 for GU are weights used in Nussinov and Jacobson, (1980).

Employing a substantially more complicated algorithm, similar to the dynamic programming calculation of the partition function described in McCaskill (1990), the partition function contributions can also be computed according to the Turner energy model. In the Turner energy model a secondary structure is decomposed into loops, as described in Zuker and Sankoff (1984), and the energy is computed as a sum of the energy contributions of the loops. A k-loop consists of $k-1$ base pairs (excluding the closing base pair) and u unpaired bases. The energies of 1-loops (hairpins) and 2-loops (stacks if u=0, bulges or interior loops if $u>0$) are based on experimental data (Mathews and Turner, 2002; Mathews *et al.*, 1999) and are dependent on k and u as well as the RNA sequence. In the Turner model, the energies for multi-loops ($k>2$) are generally determined by the approximate linear model $E_{\mathrm{M}} = a + b(k-1) + cu$, where a, b and c are constants.

As before, from now on we regard $\mathbf{s}$ and $\mathcal{S}$ to be a fixed RNA sequence with compatible secondary structure $\mathcal{S}$. The partition function for s is then defined as $\mathbf{Z} = \sum_{\mathcal{T}} e^{-E_{\mathcal{T}}/RT}$, where the sum is taken over all structures $\mathcal{T}$ compatible with $\mathbf{s}$, and $E_{\mathcal{T}}$ is the energy of the structure $\mathcal{T}$. We aim to compute the restriction $\mathbf{Z}^{\delta} = \mathbf{Z}^{\delta}_{1,n} = \mathbf{Z}^{\delta}_{1,n}(\mathbf{s}, \mathcal{S})$, i.e. the sum of $e^{-E_{\mathcal{T}}/RT}$ taken over all structures $\mathcal{T}$ that are compatible with $\mathbf{s}$ and at base pair distance δ from $\mathcal{S}$. The probability for finding a structure at a distance δ from $\mathcal{S}$ is then given by $p^{\delta} = \mathbf{Z}^{\delta}/\mathbf{Z}$.

As with the usual McCaskill partition function calculations (McCaskill, 1990), in the dynamic programming we use three matrices $\mathbf{Z}$, $\mathbf{ZB}$ and $\mathbf{ZM}$ for recursively computing $\mathbf{Z}^{\delta}$ instead of the single matrix $\mathbf{N}$ used for computing $\mathbf{N}^{\delta}$ in the previous section. In particular, for the sequence segment $[i,j]$ of $\mathbf{s}$, define $\mathbf{Z}^{\delta}_{i,j} = \sum e^{-E_{\mathcal{T}_{[i,j]}}/RT}$, where the sum is over all structures $\mathcal{T}_{[i,j]}$ compatible with $\mathbf{s}$ and such that $d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{T}_{[i,j]}) = \delta$. Also, define the restricted partition function $\mathbf{ZB}^{\delta}_{i,j}$ as the sum of $e^{-E_{\mathcal{T}_{[i,j]}}/RT}$ taken over all structures $\mathcal{T}_{[i,j]}$ such that $(i,j) \in \mathcal{T}_{[i,j]}$, and $\mathbf{ZM}^{\delta}_{i,j}$, which is the partition function contribution if the sequence segment $[i,j]$ is part of a multi-loop. The matrices $\mathbf{Z}$, $\mathbf{ZB}$ and $\mathbf{ZM}$ are filled using the following three recursions.

To compute $\mathbf{Z}$ we use the recursions

$$\mathbf{Z}^{\delta}_{i,j} = \mathbf{Z}^{\delta - b_0}_{i,j-1} + \sum_{\substack{s_k\, s_j \in \mathbb{B}, \\ i \leq k < j}} \left(e^{-E_d/RT} \sum_{w + w' = \delta - d_1} \mathbf{Z}^{w}_{i,k-1} \mathbf{ZB}^{w'}_{k,j} \right), \tag{6}$$

where E_d is the energy contribution due to dangling ends (energy contributions from single bases stacking on adjacent base pairs) and closing AU base pairs (since a non-GC base pair closing a stem has a destabilizing effect), and $d_1 = d_{BP}(\mathcal{S}_{[i,j]}, \mathcal{S}_{[i,k-1]} \cup \mathcal{S}_{[k,j]})$. Note that the first term of this recursion corresponds to the case where j is unpaired (and hence has no energy contribution) in $[i,j]$. The second term includes all other structures on $[i,j]$. The sum is taken over all possible base pairs (k,j) with $i \leq k < j$. If (k,j) is a base pair the partition function for $[k,j]$ is given by $\mathbf{ZB}^{w'}_{k,j}$, the partition function for $[i,k-1]$ is given by $\mathbf{Z}^{w}_{i,k-1}$.

We compute $\mathbf{ZB}$ using the recursion

$$\begin{aligned} \mathbf{ZB}^{\delta}_{i,j} = {} & \Delta(d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \{(i,j)\}) - \delta) e^{-E(i,j)/RT} \\ & + \sum_{\substack{s_k s_j \in \mathbb{B}, \\ i < k < l < j}} \mathbf{ZB}^{\delta - d_2}_{k,l} e^{-E(i,j,k,l)/RT} \\ & + \sum_{\substack{s_k s_j \in \mathbb{B}, \\ i < k < l < j}} \left(e^{-(a + b + c(j-l-1))/RT} \sum_{w + w' = \delta - d_3} \mathbf{ZM}^{w}_{i+1,k-1} \mathbf{ZB}^{w'}_{k,l} \right), \end{aligned} \tag{7}$$

where $E(i,j)$ is the energy of the hairpin loop with closing base pair (i,j), $E(i,j,k,l)$ is the energy of the stack, bulge or interior loop with the closing base pair (i,j) and the interior base pair (k,l), $d_2 = d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{S}_{[k,l]} \cup \{(i,j)\})$, and $d_3 = d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{S}_{[i+1,k-1]} \cup \mathcal{S}_{[k,l]} \cup \{(i,j)\})$. Here, $\Delta(x,y)$ is the Kronecker function, which equals 1 if x=y, and else 0. Note that since the above equation computes $\mathbf{ZB}^{\delta}_{i,j}$, it follows that (i,j) forms a base pair in the neighboring structures $\mathcal{T}_{[i,j]}$ (if this is not possible, then $\mathbf{ZB}^{\delta}_{i,j} = 0$). The first term in the recursion takes care of the case where (i,j) is the only base pair in $[i,j]$, i.e. (i,j) closes a hairpin loop. The second term handles the case where there is an interior loop (or a bulge or a stack) closed by (i,j) and (k,l). The third term takes care of

all the structures where (i,j) closes a multi-loop. To reduce complexity of the algorithm, the interior and bulge loop size can be limited to a maximum size of L, by requiring that $l > j-L$ in the above recursion.

The final recursion, for computing **ZM**, is

$$\begin{aligned}\mathbf{ZM}^{\delta}_{i,j} = \mathbf{ZM}^{\delta-b_0}_{i,j-1}e^{-c/RT} \\ + \sum_{\substack{s_k s_j\in\mathbb{B},\\ i\le k<j}} \left(\mathbf{ZB}^{\delta-d_4}_{k,j}e^{-(b+c(k-i))/RT}\right. \\ \left. + \sum_{w+w'=\delta-d_5} \mathbf{ZM}^{w}_{i,k-1}\mathbf{ZB}^{w'}_{k,j}e^{-b/RT}\right),\end{aligned} \tag{8}$$

where $d_4 = d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{S}_{[k,j]})$ and $d_5 = d_{\mathrm{BP}}(\mathcal{S}_{[i,j]}, \mathcal{S}_{[i,k-1]} \cup \mathcal{S}_{[k,j]})$. Note that since $\mathbf{ZM}^{\delta}_{i,j}$ computes the partition function contribution under the assumption that $[i,j]$ is part of a multi-loop, there will be exactly one stem-loop structure in this region (the **ZB** term) or more than one (the **ZB**-**ZM** term).

Note that the recursions for computing the number of δ-neighbors and the partition function analogues are non-redundant in that each structure is counted once and only once.

Pseudocode for computing $\mathbf{Z}^{\delta}$ is given in the Supplementary Material. The complexity is the same as for computing the number of δ-neighbors, $O(mn^2)$ in space and $O(mn^3)$ in time, if the size of internal loops and bulges are limited to a fixed length such as 30, following the convention of Vienna RNA Package.

2.3 Minimum free energy δ-neighbors

Given an RNA nucleotide sequence **s** and secondary structure $\mathcal{S}$, the *minimum free energy δ-neighbor* is that secondary structure $\mathcal{T}$ of **s**, which has base pair distance δ with $\mathcal{S}$, and which has least free energy MFE^{δ} among all structures having base pair distance δ with $\mathcal{S}$. Free energy is measured according to the Turner energy model (Mathews *et al.*, 1999; Xia *et al.*, 1999), where our treatment of dangles follows that of Vienna RNA package with `-d2` option.

In this section, we describe a novel algorithm capable of computing the MFE^{δ} structures, for all δ. As in our partition function computation, the run time [resp. space requirement] to compute all MFE^{δ} structures for $\delta \le m$ is $O(mn^3)$ [resp. $O(mn^2)$]. This algorithm is obtained from the algorithm in Section 2.2 essentially by replacing Boltzmann factor $e^{-E(\mathcal{T})/RT}$ by free energy $E(\mathcal{T})$ and by replacing the operations of addition [resp. multiplication] by minimization [resp. addition]. In future work, we plan to analyze the structure morphological changes in proceeding from $\mathcal{S}$ to MFE^0, MFE^1, MFE^2, etc. As indicated in the Results section, such an analysis could prove useful in conformational switch detection and other applications.

Fix RNA nucleotide sequence $\mathbf{s} = s_1,\ldots,s_n$ and secondary structure $\mathcal{S}$ of **s**. To compute MFE^{δ} we use the following recursions:

$$\begin{aligned}\mathrm{MFE}^{\delta}_{i,j} = \min\Bigg\{&\mathrm{MFE}^{\delta-b_0}_{i,j-1}, \\ &\min_{\substack{s_k s_j\in\mathbb{B},\\ i\le k<j}} \min_{w+w'=\delta-d_1} \mathrm{MFE}^{w}_{i,k-1} + \mathrm{MFEB}^{w'}_{k,j} + E_d\Bigg\}\end{aligned} \tag{9}$$

$$\begin{aligned}\mathrm{MFEB}^{\delta}_{i,j} = \min\Bigg\{&\Delta(d_{\mathrm{BP}}(\mathcal{S}_{[i,j]},\{(i,j)\})-\delta)E(i,j), \\ &\min_{\substack{s_k s_l\in\mathbb{B},\\ i<k<l<j}} \mathrm{MFEB}^{\delta-d_2}_{k,l} + E(i,j,k,l), \\ &\min_{\substack{s_k s_l\in\mathbb{B},\\ i<k<l<j}} \min_{w+w'=\delta-d_3} \mathrm{MFEM}^{w}_{i+1,k-1} + \mathrm{MFEB}^{w'}_{k,l} + a + b + c(j-l-1)\Bigg\},\end{aligned} \tag{10}$$

$$\begin{aligned}\mathrm{MFEM}^{\delta}_{i,j} = \min\Bigg\{&\mathrm{MFEM}^{\delta-b_0}_{i,j-1} + c, \\ &\min_{\substack{s_k s_j\in\mathbb{B},\\ i\le k<j}} \Big\{\mathrm{MFEB}^{\delta-d_4}_{k,j} + b + c(k-i), \\ &\min_{w+w'=\delta-d_5} \mathrm{MFEM}^{w}_{i,k-1} + \mathrm{MFEB}^{w'}_{k,j} + b\Big\}\Bigg\}.\end{aligned} \tag{11}$$

Once the minimum free energy of δ-neighbors (MFE^{δ}) is computed the corresponding MFE structures can be computed by a simple traceback for each MFE^{δ}.

For reasons of space, the pseudocode for computing MFE^{δ} is not presented; given our previous description of MFE^{δ} and the pseudocode for computing the partition function $\mathbf{Z}^{\delta}$, appearing in the Supplementary Material, the reader will have no difficulty to reconstruct the pseudocode for MFE^{δ}.

3 RESULTS

In this section, we present probability density graphs for a variety of conformational switches and for some non-switches. Additional data is provided for all SAM riboswitches in our Supplementary Material. We also compare the output of `RNAbor` with distance plots generated by the web server `paRNAss` (Giegerich *et al.*, 1999), that uses a heuristic to determine whether there appear to be two or more clusters of distinct secondary structures for a given RNA sequence. Some example are also presented that indicate that `RNAbor` can be used to provide improved secondary structure predictions as compared with the MFE structure.

3.1 Detecting conformational switches

In this section, we define a *conformational switch* to be an RNA sequence which has exactly two distinct low energy secondary structures. By *multi-switch* we mean an RNA sequence which can adopt two or more distinct low energy secondary structures. For a given RNA sequence **s** and secondary structure $\mathcal{S}$ of **s**, we use `RNAbor` to compute $p^{\delta} = \mathbf{Z}^{\delta}/\mathbf{Z}$. Taking $\mathcal{S}$ to be the MFE structure, or alternatively the structure determined by comparative sequence alignment (Cannone *et al.*, 2002), our intuition is that a conformational switch should display a bi-modal probability density graph.

To illustrate the behavior of a typical conformational switch, we present examples of `RNAbor` output for known switches. Consider for instance the 105 nt SAM riboswitch with EMBL accession number `AE016750.1/132874-132778` and sequence `AACUUAUCAA GAGAAGUGGA GGGACUGGCC CAAAGAAGCU UCGGCAACAU UGUAUCAUGU GCCAAUUCCA GUAACCGAGA AGGUUAGAAG AUAAGGU`. Figure 2 displays three secondary structures: the MFE structure, the MFE^{24} structure and the native structure inferred by comparative sequence analysis of the SAM riboswitch seed multiple sequence alignment from Rfam (Griffiths-Jones *et al.*, 2003). As computed by `RNAbor`, the MFE structure has free energy of −28.1 kcal/mol and the Boltzmann probability p^0 is 0.11, while the MFE^{24} structure has free energy of −26.7 kcal/mol and p^{24} is 0.05. Note the similarity of the MFE^{24} structure with the native structure; in particular, the apical loop regions are correctly computed. There is a second MFE structure–the

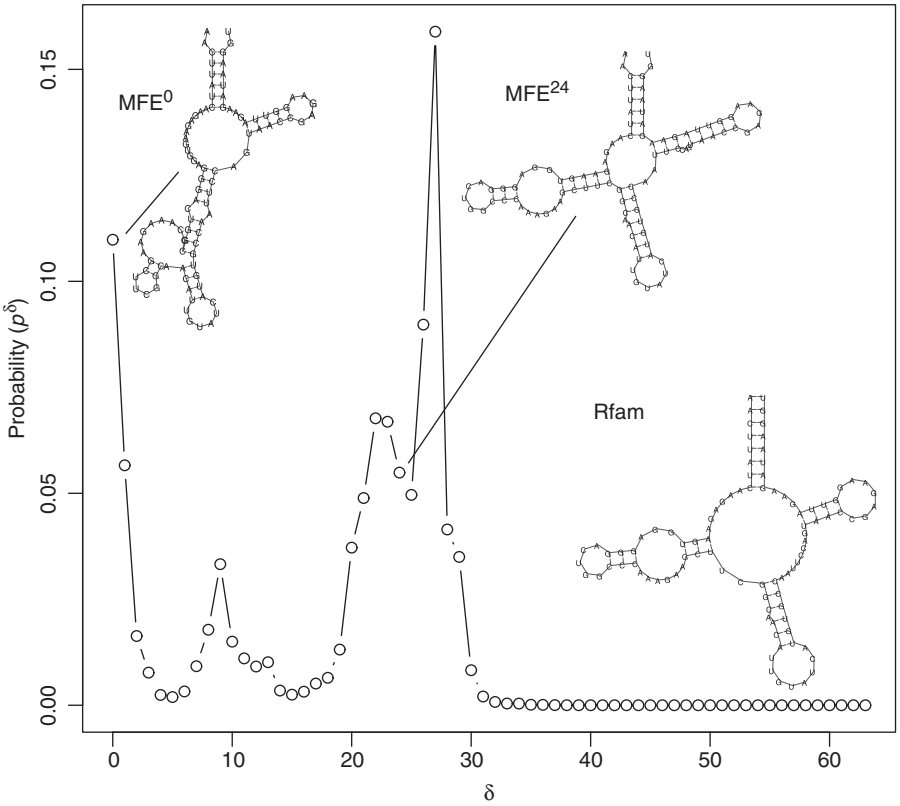


Fig. 2. Boltzmann probability density plot for the 105 nt SAM riboswitch with EMBL accession number AE016750.1/132874-132778. The curve shows the probability, $p^{\delta} = \mathbf{Z}^{\delta}(\mathbf{s}, \mathcal{S})/Z(\mathbf{s}, \mathcal{S})$, for all secondary structures of RNA sequence **s** having base pair distance δ from the MFE structure $\mathcal{S}$. Alternative secondary structures for the riboswitch are shown in the figure. In the upper left is the MFE structure with free energy −28.10 kcal/mol shown, in the middle the MFE^{24} structure with free energy −26.7 kcal/mol and in the lower right the consensus Rfam structure. Sorted in decreasing order, the most significant Boltzmann probabilities are 0.16, 0.11, 0.090, 0.068, 0.067, 0.057, 0.055, respectively for values of δ = 27,0,26,22,23,1,24.

MFE^{27} structure, with free energy −28.1 kcal/mol. Although the Boltzmann probability p^{27} is 0.16, the maximum of p^{δ} over all δ, the MFE^{27} structure is rather different than the native structure (data not shown).

Figure 2 shows the probability density plot, i.e. the probability $p^{\delta} = \mathbf{Z}^{\delta}/\mathbf{Z}$ of finding a structure at distance δ from the input structure, which in this example is the MFE structure.

It is not the case that all conformational switches display a bi-modal or multi-modal Boltzmann probability density curve. In particular, the probability density curve is uni-modal for the 101 nt switch (Schlax *et al.*, 2001) with EMBL accession number AE0140031/5850-5961. This mRNA has a pseudoknotted structure, which is responsible for the translational repression of the alpha operon by an entrapment mechanism. Since the algorithm RNAbor, like its predecessors mfold and RNAfold, considers only non-pseudoknotted secondary structures, there is no reason to expect that RNAbor display a multi-modal probability density curve for this conformational switch.

Figure 3a depicts a bi-modal density graph for the artificially engineered bistable switch CUUAUGAGGG UACUCAUAAG AGUAUCC of Flamm *et al.* (2001). Figure 3b displays the probability density function p^{δ} for the 76 nt conformational switch (Ke *et al.*, 2004), which controls hepatitis delta virus ribozyme catalysis (PDB ID 1SJ3:R). Both these examples display bi-modal Boltzmann probability curves.

3.2 Comparison with paRNAss

We now compare the ability of RNAbor and paRNAss to predict RNA conformational switches. We have chosen to display the paRNAss distance plot of energy barrier versus morphological distance (Voss *et al.*, 2004), but in all the below examples the distance plots using tree alignment or string edit distance showed similar results.

The *Escherichia coli hok* (*ho*st *k*illing) mRNA folds into two different conformations (Franch *et al.*, 1997). The full length mRNA folds into a stable structure involving a long-range interaction between the 5′ and 3′-end. Degradation of the 3′-end leads to a conformational change as the stabilizing long-range interaction is broken. Here, we have investigated the part of the mRNA that undergoes a conformational change (as provided on the paRNAss web server http://bibiserv.techfak.uni-bielefeld.de/parnass).

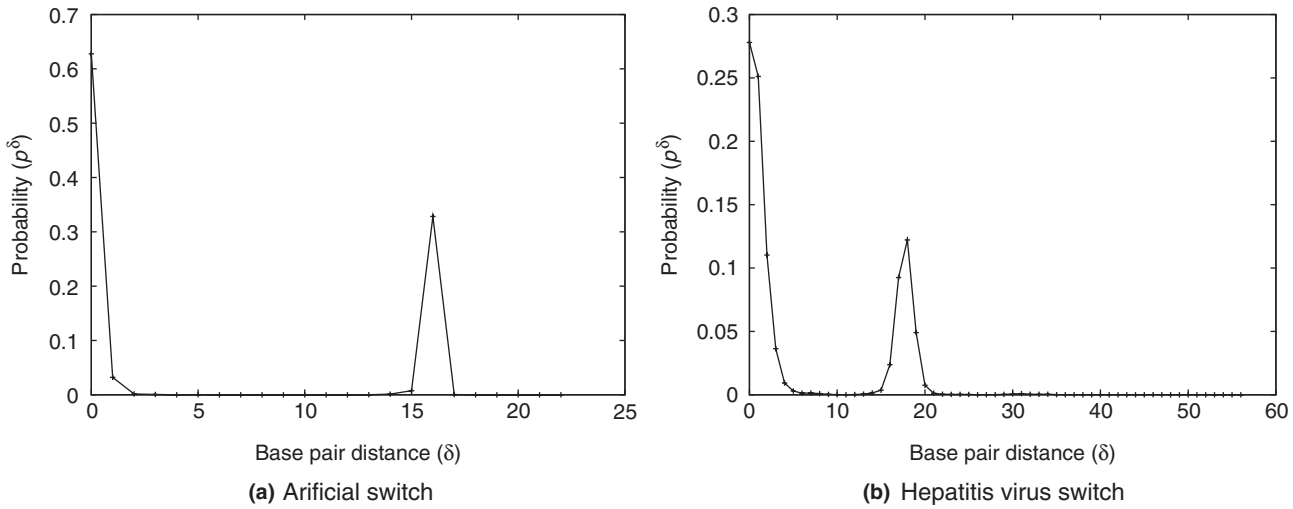


Fig. 3. (**a**) Boltzmann probability density plot for the 29 nt bistable switch artificially engineered by Flamm *et al.* (2001) and having sequence CUUAUGAGGG UACUCAUAAG AGUAUCC. The graph shows the Boltzmann probability, $p^\delta = \mathbf{Z}^\delta(\mathbf{s}, \mathcal{S})/\mathbf{Z}(\mathbf{s}, \mathcal{S})$, of all δ-neighbors, for all values of δ bounded by sequence length. (**b**) Boltzmann probability density plot of δ-neighbors for the 76 nt conformational switch which controls hepatitis delta virus ribozyme catalysis (PDB code `1SJ3:R`)(Ke *et al.*, 2004).

For this RNA, both `RNAbor` and `paRNAss` detect the conformational switch, the `RNAbor` probability plot shows two distinct peaks suggesting two alternative stable structures and the `paRNAss` plot shows two clearly separated clusters, both suggesting that all the reasonably stable structures fall into one out of two conformations, see Figure 4.

Although both `RNAbor` and `paRNAss` suggest that the *hok* gene has two alternative structures, there are some uncertainties in the result. In the `RNAbor` density plot there are actually three peaks (even though the third peak is significantly smaller than the other two), indicating that there might be more than two alternative structures.

The 5′-untranslated (UTR) region of *E.coli thiM* mRNA undergoes a change in structure, that is important for regulation (Winkler *et al.*, 2002). Both `RNAbor` and `paRNAss` indicate more than one single stable structure for the *thiM*-leader. As can be seen from Figure 5, there actually seem to be more than two alternative structures. However, the third structure seems to be less important (lower probability), and hence this RNA is predicted as a conformational switch by `RNAbor`.

3.3 Improving on the minimum free energy structure

In this section, we discuss several examples where a MFE^δ structure is closer to the native secondary structure, as extracted from the 3D X-ray structure, than is the MFE structure. This phenomenon generally occurs when the probability density graph indicates a second peak, although sometimes that peak may be modest.

Figure 6 presents the Boltzmann probability density plot and alternate secondary structure models for the S-adenosylmethionine riboswitch mRNA regulatory element with PDB code `2GIS` (Montange and Batey, 2006). The figure shows the native secondary structure for `2GIS`, determined by extraction from the 3D X-ray structure[1], the MFE^{24} structure, which clearly resembles the native state, and the rather different looking MFE structure. Figure 6 graphs the probability density, where a large second peak is present, centered at $\delta = 23$.

Improvement over the MFE structure is not restricted to riboswitches. Indeed, Figure 7 displays a very unrealistic linear MFE structure for Ile-tRNA from *E.coli*—accession number DI1660 from Sprinzl's tRNA database (Sprinzl *et al.*, 1998), as well as alternative structures predicted by `RNAbor`. Note the correctly predicted anti-codon GAU.

4 DISCUSSION

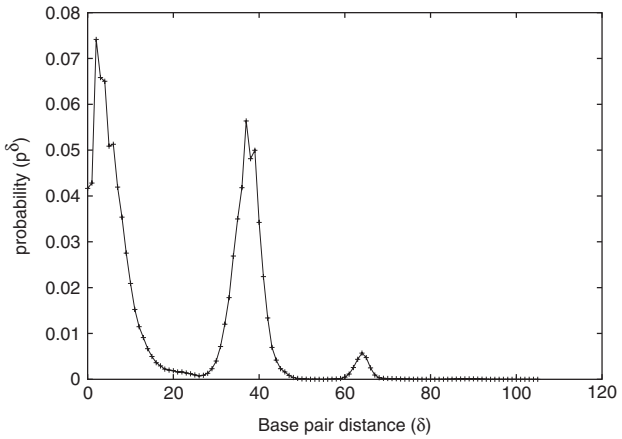
In this article we present some novel algorithms for efficiently computing the number ($\mathbf{N}^\delta$), the Boltzmann partition function ($\mathbf{Z}^\delta$) and the minimum free energy (MFE^δ) and corresponding structure over the collection of all δ-neighbors of a secondary structure of a fixed RNA sequence, all of which are implemented in the webserver `RNAbor`.

We find that the output of `RNAbor` gives useful insights into the landscape of foldings for a single RNA sequence. In addition, we observe that the output of `RNAbor` compares well with that of the conformational switch detection program `paRNAss`. In future work, we will make a more extensive comparison of `RNAbor` with other RNA folding landscape analysis programs, such as `RNAshapes` (Giegerich *et al.*, 2004; Steffen *et al.*, 2006; Voss *et al.*, 2006) and `sfold` (Ding and Lawrence, 2003; Ding *et al.*, 2004).

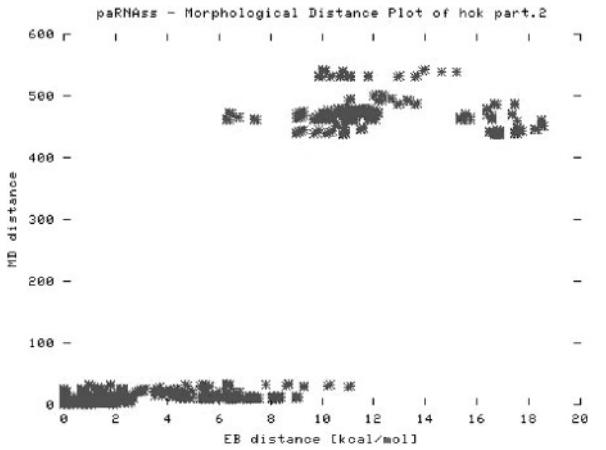
Potential applications of `RNAbor` which will be pursued in future work include the following.

- Since `RNAbor` allows one to distinguish whether the given RNA nucleotide sequence **s** has a single pronounced *well of*

[1]Using the software `rnaview` (Yang *et al.*, 2003), we first obtained a list of all Watson–Crick *cis* base pairs.

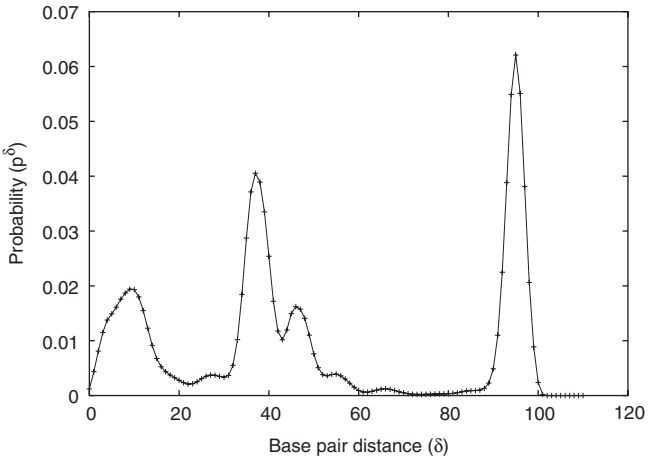


(a) RNAbor density plot.

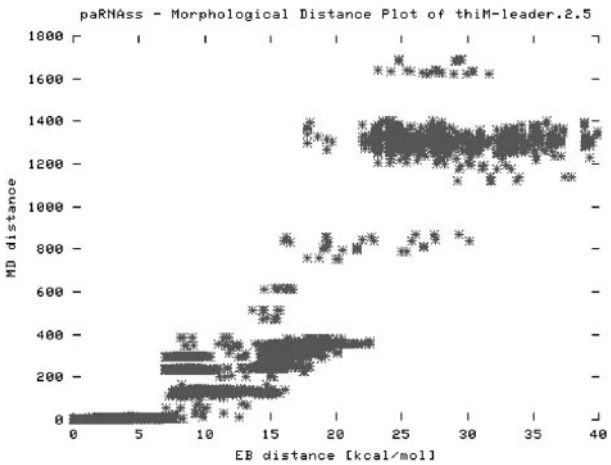


(b) paRNAss distance plot.

Fig. 4. The *E.coli hok* (*ho*st *k*illing) mRNA folds into two different conformations (Franch *et al.*, 1997). (**a**) The RNAbor density plot shows two distinct peaks indicating that the *hok* mRNA has two alternative folds. The paRNAss (**b**) distance plot also indicate that this RNA sequence can fold into two alternative secondary structures.



(a) RNAbor density plot.



(b) The paRNAss distance plot.

Fig. 5. *E.coli thiM*-leader. (**a**) RNAbor density plot and (**b**) the paRNAss distance plot.

attraction around a given secondary structure $\mathcal{S}$ of **s**, it may be possible to use RNAbor to detect situations where the native secondary structure, as determined by X-ray crystallography, is different than that proposed by mfold and RNAfold (e.g. see Section 3.3). The idea would be to determine if there is no peak around $\delta = 0$, when $\mathcal{S}$ is taken to be the MFE structure.

- Figure 5a shows an interesting example where the RNA seems to have more than one alternative structure. Does this RNA have more than two alternative structures? Is it the case that the MFE structure is not biologically functional? (In this example, the other two alternative structures seem to be probable.) Using RNAbor, we can determine the minimum free energy structures over all δ-neighbors, and subsequently focus on MFE^{δ} structures where the Boltzmann probability p^{δ} is high. Ultimately, chemical probing experiments might determine whether these MFE^{δ} structures are the preferred biologically active structure.
- RNAbor is a useful complement to already existing tools for detecting putative conformational switches. Unlike paRNAss, the number of structures to be analyzed and the maximum allowable free energy difference from the MFE structure need not be decided in advance. (These can change the paRNAss result quite dramatically.) Depending on the number of structures to be analyzed and the energy bound, paRNAss can take an exponential amount of time, in contrast to $O(mn^3)$ time for RNAbor to compute $\mathbf{N}^{\delta}$, $\mathbf{Z}^{\delta}$ and MFE^{δ}.
- As shown in Figures 6 and 7, RNAbor can sometimes predict a secondary structure, which is closer to the real secondary structure than is the MFE structure, as determined from X-ray structures or comparative sequence analysis.
- As for any bioinformatics software, it will be necessary to perform experimental validation of predictions made by RNAbor. In future work, we intend to include user-defined

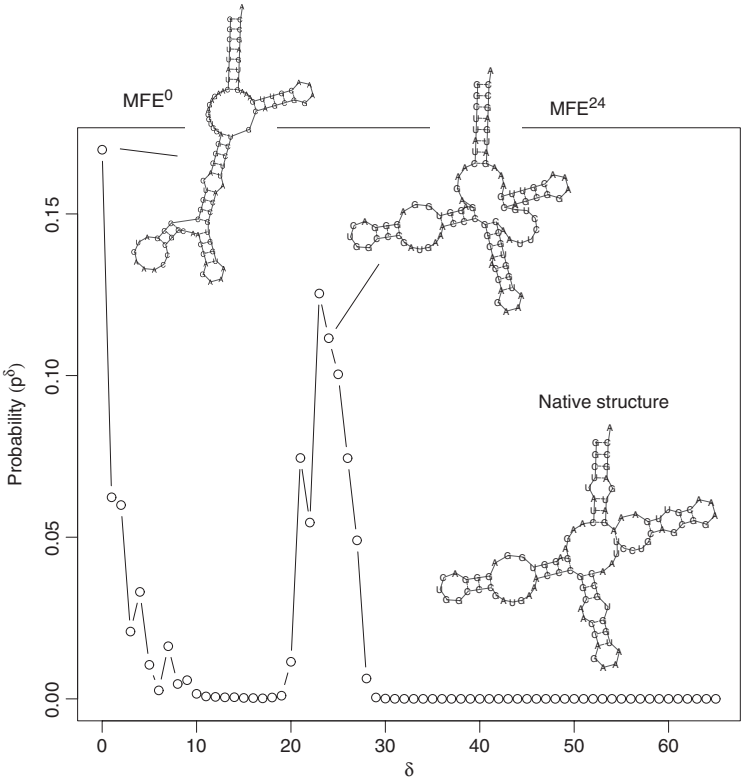


Fig. 6. Boltzmann probability density plot for the S-adenosylmethionine riboswitch mRNA regulatory element, with sequence GGCUUAUCAA GAGAGGUGGA GGGACUGGCC CGAUGAAACC CGGCAACCAG AAAUGGUGCC AAUUCCUGCA GCGGAAACGU UGAAAGAUGA GCCA. The MFE secondary structure, as determined by `RNAfold -d2`, is shown to the upper left. This structure has free energy −42.3 kcal/mol, and the Boltzmann probability $p^0 = Z^0/Z$ of the MFE structure is 0.169854. The MFE^{24} structure computed by `RNAbor` for the same riboswitch sequence is shown in the middle. This structure has free energy of −38.8 kcal/mol, and the Boltzmann probability $p^{24} = Z^{24}/Z = 0.11$. The secondary structure S-adenosylmethionine riboswitch mRNA regulatory element determined from X-ray structure `2GIS` (Montange and Batey, 2006) is shown in the lower right.

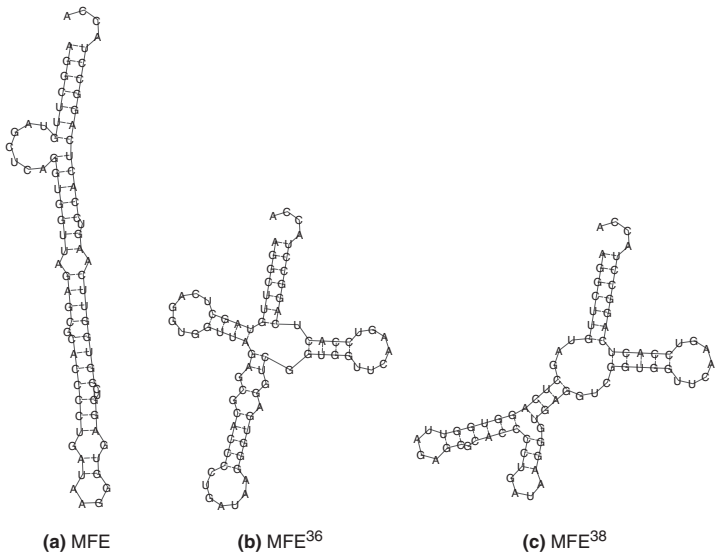


Fig. 7. (**a**) Minimum free energy secondary structure, as determined by `RNAfold -d2`, for Ile-tRNA from *E.coli* with anti-codon GAU and accession number DI1660 from Sprinzl's tRNA database (Sprinzl *et al.*, 1998). The MFE structure has free energy of −28.61 and Boltzmann probability $p^0 = 0.136856$. (**b**) MFE^{36} structure, as determined by `RNAbor`. This structure has free energy of −26.1 kcal/mol and $p^{36} = 0.033946$. (**c**) MFE^{38} structure, as determined by `RNAbor`. This structure has free energy of −27.5 kcal/mol, and $p^{38} = 0.074037$.

constraints, which allow the user to require all investigated structures to contain certain specified base pairs and for certain specified nucleotides to remain unpaired. This will allow RNAbor to be used together with chemical probing experiments to determine biologically active conformers.

ACKNOWLEDGEMENTS

Research of P.C. was partially supported by National Science Foundation DBI-0543506, which additionally supported some travel of E.F. All three authors would like to thank Elena Rivas, Eric Westhof and funding agencies for organizing the meeting RNA-2006 in Benasque, Spain, in July 2006, where some of this work was carried out.

Conflict of Interest: none declared.

REFERENCES

Adams,M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Bekaert,M. *et al.* (2003) Towards a computational model for -1 eukaryotic frameshifting sites. *Bioinformatics*, **19**, 327–335.

Brown,C. *et al.* (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**, 527–542.

Cannone,J. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2. (Erratum in: BMC Bioinformatics (2002), 3, 15).

Commans,S. and Böck,A. (1999) Selenocysteine inserting tRNAs: an overview. *FEMS Microbiol. Rev.*, **23**, 333–351.

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

Ding,Y. *et al.* (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.*, **32**, W135–W141.

Doudna,J. and Cech,T. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.

Flamm,C. *et al.* (2001) Design of multi-stable RNA molecules. *RNA*, **7**, 254–265.

Franch,T. *et al.* (1997) 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.*, **273**, 38–51.

Giegerich,R. *et al.* (1999) Prediction and visualization of structural switches in RNA. *Pac. Symp. Biocomput.*, **4**, 126–137.

Giegerich,R. *et al.* (2004) Abstract shapes of RNA. *Nucleic Acids Res.*, **32**, 4843–4851.

Griffiths-Jones,S. *et al.* (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.

Harborth,J. *et al.* (2003) Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. *Antisense Nucleic Acid Drug Dev.*, **13**, 83–105.

Hofacker,I. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, **125**, 167–188.

Ke,A. *et al.* (2004) A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature*, **429**, 201–205.

Lim,L. *et al.* (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.

Mathews,D. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mathews,D.H. and Turner,D. H. (2002) Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, **41**, 869–880.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, **29**, 1105–1119.

Montange,R. and Batey,R. (2006) Structure of the S-adenosylmethionine riboswitch mRNA regulatory element. *Nature*, **441**, 1172–1175.

Moon,S. *et al.* (2004) Predicting genes expressed via −1 and +1 frameshifts. *Nucleic Acids Res.*, **32**, 4884–4892.

Moulton,V. *et al.* (2000) Metrics on RNA secondary structures. *J. Comput. Biol.*, **7**, 277–292.

Nissen,P. *et al.* (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–923.

Nussinov,R. and Jacobson,A. B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. USA*, **77**, 6903–6913.

Penchovsky,R. and Breaker,R. (2005) Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nat. Biotechnol.*, **23**, 1424–1431.

Schlax,P. *et al.* (2001) Translational repression of the *Escherichia coli* alpha operon mRNA: importance of an mRNA conformational switch and a ternary entrapment complex. *J. Biol. Chem.*, **276**, 38494–38501.

Sprinzl,M. *et al.* (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **26**, 148–153.

Steffen,P. *et al.* (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.

Tuschl,T. (2003) Functional genomics: RNA sets the standard. *Nature*, **421**, 220–221.

Vicens,Q. and Cech,T. (2006) Atomic level architecture of group I introns revealed. *Trends Biochem. Sci.*, **31**, 41–51.

Voss,B. *et al.* (2004) Evaluating the predictability of conformational switching in RNA. *Bioinformatics*, **20**, 1573–1582.

Voss,B. *et al.* (2006) Complete probabilistic analysis of RNA shapes. *BMC Biol.*, **4**, 5.

Walter,P. and Blobel,G. (1982) Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, **299**, 691–698.

Weinger,J.S. *et al.* (2004) Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nat. Struct. Mol. Biol.*, **11**, 1101–1106.

Winkler,W.C. *et al.* (2002) An mRNA structure that controls gene expression by binding FMN. *Proc. Natl Acad. Sci. USA.*, **99**, 15908–15913.

Wuchty,S. *et al.* (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.

Xia,T.J. *et al.* (1999) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.

Yang,H. *et al.* (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3560.

Zuker,M. (1994) Prediction of RNA secondary structure by energy minimization. *Methods Mol. Biol.*, **25**, 267–294.

Zuker,M. and Sankoff,D. (1984) RNA secondary structures and their prediction. *Bull. of Math. Biol.*, **46**, 591–621.