

Finding stable local optimal RNA secondary structures

Yuan Li and Shaojie Zhang*

Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Many RNAs, such as riboswitches, can fold into multiple alternate structures and perform different biological functions. These biologically functional structures usually have low free energies in their local energy landscapes and are very stable such that they cannot easily jump out of the current states and fold into other stable conformations. The conformational space of feasible RNA secondary structures is prohibitively large, and accurate prediction of functional structure conformations is challenging. Because the stability of an RNA secondary structure is determined predominantly by energetically favorable helical regions (stacks), we propose to use configurations of putative stacks to represent RNA secondary structures. By considering a reduced conformational space of local optimal stack configurations instead of all feasible RNA structures, we first present an algorithm for enumerating all possible local optimal stack configurations. In addition, we present a fast heuristic algorithm for approximating energy barriers encountered during folding pathways between each pair of local optimal stack configurations and finding all the stable local optimal structures.

Results: Benchmark tests have been conducted on several RNA riboswitches, whose alternate secondary structures have been experimentally verified. The benchmark results show that our method can successfully predict the native 'on' and 'off' secondary structures, and better rank them compared with other state-of-art approaches.

Availability: The software is freely available and can be downloaded at <http://genome.ucf.edu/RNASLOpt>.

Contact: shzhang@eecs.ucf.edu

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

Received on May 30, 2011; revised on August 14, 2011; accepted on September 4, 2011

1 INTRODUCTION

RNA molecules play important roles in many biological activities. The structures of RNAs are key to understanding their functionalities. The native secondary structures of RNA molecules are usually the minimum free energy (MFE) structures. However, many regulatory RNAs may form multiple alternate metastable structures, each of which, performs a different function (Baumstark *et al.*, 1997; Schultes and Bartel, 2000). For example, the *cis*-regulatory RNAs (called as riboswitches) are widely present at the non-coding regions of mRNAs in bacteria, and regulate both transcriptional and translational expression of their cognate

metabolic genes and thus comprise a metabolite-sensing system. Riboswitches can selectively bind with small metabolites and form alternative structure conformations, in order to exert control over translation initiation or formation of a transcription terminator (or an anti-terminator) helix and thus turn 'off' (or 'on') the gene transcription (Mandal and Breaker, 2004; Wakeman *et al.*, 2007).

Alternate structure conformations of RNA riboswitches can be experimentally verified using in-line probing (Mandal and Breaker, 2004), X-ray crystallography (Batey *et al.*, 2004) or Nuclear Magnetic Resonance (NMR) spectroscopy (Noeske *et al.*, 2005). However, these methods are usually time consuming and expensive. Computational approaches for accurately predicting alternate conformations are in need. The alternate functional structures are usually also energetically favored and are stable in their local energy landscapes. The conformational transitions between any pair of alternate structures may involve high-energy barriers, such that RNAs can easily become kinetically trapped by these structures. Accurate predictions of alternate structures of an RNA molecule should be conducted by exploiting the energy landscape and the folding dynamics of the RNA, in combination with the binding of the target metabolites. The ideal approach is to construct an exact energy landscape on all possible suboptimal secondary structures, then analyze every possible local optimal structures as well as all possible folding pathways in the landscape, and finally determine the most significant structures. In the following, we will briefly review existing methods for enumerating suboptimal structures and predicting alternate structures for RNA molecules.

Zuker (1989) devised the first algorithm, *mfold*, for predicting the MFE structure and multiple suboptimal structures. For a given sequence, it generates, for each admissible base pair, the energetically best structure containing that base pair. For a sequence of length n , *mfold* produces at most $n(n-1)/2$ suboptimal structures, which are a very small fraction of all the candidate suboptimal structures, and may miss some of the functional structures. In addition, *mfold* uses a filter based on the base pair metric to remove structures that are similar to one another. The filter is based on base pair difference, while it might be better to infer stability of structures in the context of energy landscape and remove unstable structures.

Wuchty *et al.* (1999) proposed the first exact solution, *RNAsubopt*, for predicting all possible suboptimal structures between the MFE and an arbitrary upper limit using a mathematical model proposed by Waterman and Byers (1985) based on the Turner energy model (Freier *et al.*, 1986; He *et al.*, 1991; Jaeger *et al.*, 1989; Turner *et al.*, 1988). Parisien and Major (2008) devised MC-Fold, a similar solution to the same problem that takes into account both non-canonical base pairings and pseudoknotted structures. In addition, Flamm *et al.* (2002) presented BARRIERS, an algorithm for constructing the exact energy landscape on all possible

*To whom correspondence should be addressed.

suboptimal structures produced by RNAsubopt. BARRIERS is able to distinguish all the local optimal structures and can build a barrier tree representing the energy landscape. However, the number of feasible structures grows quickly with the length of the RNA sequence and the energy range, and RNAsubopt enumerates enormous solutions for even a short sequence with a small energy range. For example, the free energies of the native 'off' and 'on' structures of the 110 nt long adenine riboswitch of *ydhL* from *B. subtilis* are -32.3 and -14.8 (kcal/mol), respectively. The number of feasible structures grows quickly as free energy increases (which is shown in Supplementary Figure S1), and the number of structures with free energies between the two native structures exceeds 10^9 . Therefore, it is very difficult and time consuming to find a few alternate structures from an enormous collection of candidates. Applications of these algorithms are generally limited to very short RNA sequences with a small energy range.

The conformational space of feasible structures not only is prohibitively large, but also renders redundant information, because many structures in the space are similar to one other. Thus, researchers have also proposed alternative approaches, which investigate reduced conformational spaces instead of the space of feasible suboptimal structures. Pipas and McMahon (1975) presented an algorithm that can construct the best k structures composed of compatible stacks (i.e. sharing no base in common and forming no pseudoknot). Nakaya *et al.* (1996) used a search tree for generating suboptimal structures by selecting a subset of stacking regions that can coexist, from the set of all possible stacking regions. The search tree is composed of m level of nodes, where m is the number of possible stacks and nodes at depth i determine whether the i -th stacking region is selected. Evers and Giegerich (2001) provided an algorithm that can enumerate all possible saturated structures such that no unpaired base can be paired without affecting the validity of the structures (Zuker, 1984). They employed a dynamic programming similar to that of Wuchty *et al.* (1999) and incorporated a saturation check to ensure that structures are saturated. Giegerich and his cooperators also presented RNAShapes (Giegerich *et al.*, 2004; Steffen *et al.*, 2006), an approach that first extracts RNA abstract shapes based on juxtaposition and embedding of stacks, and then clusters structures with the same shape together, and finally represents all the structures in a cluster by the 'shrep' of the cluster (i.e. the secondary structure with the lowest free energy in the cluster). One shortcoming of the stack-based approaches is that they may exclude incompatible stacks that overlap by only one or a few bases. If we consider shortening one of the stacks by cutting off the overlapping bases, it will result in a pair of compatible stacks. Another drawback of these approaches is that it is hard to infer the stability of RNA secondary structures in the context of energy landscape and thus is hard to accurately predict native structures. Recently, Lorenz and Clote (2011) proposed an approach, RNALocopt, that can sample a user-defined number of structures from the space of locally optimal structures. A locally optimal structure has the lowest free energy compared with its neighboring structures (obtained by adding or deleting a single base pair). One shortcoming is that when the sample size is small, RNALocopt may fail to predict the native structures, and when the sample size is large, it would be difficult to identify the significant structures from a large number of candidates.

We are interested in finding stable local optimal (SLOpt) structures that conform to the following criteria. First, a SLOpt

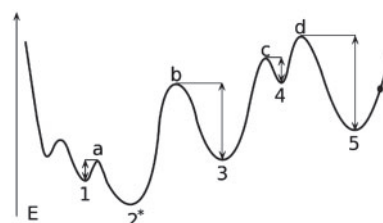


Fig. 1. A schematic representation of an energy landscape. Numbers 1, 3, 4, 5 represent local optima and 2* represents the global optimum. The dot adjacent to a local optimum 5 represents a none local optimal structure, which can transit to 5 along a gradient walk. Lower-case characters *a*, *b*, *c* and *d* are saddle points (i.e. structures with the highest free energies) of folding pathways between local optima 1&2, 2&3, 3&4 and 4&5, respectively. Bars represent the minimal additional energy required for the RNA molecule to 'jump' out of the corresponding energy basins.

structure should be local optimal (LOpt) in that it resides at the bottom of a basin in the energy landscape (i.e. has the lowest free energy compared with all its neighbors). None local optimal structures are unlikely to be biologically functional, because they can continuously transit to their lower energy neighboring structures, like climbing down a hill until a local optimum (the bottom of a basin in the energy landscape) is reached. Second, a SLOpt structure should be stable in that the minimal energy barrier between this structure and any other SLOpt structures should be high. This criterion is proposed because secondary structures with lower free energies are not guaranteed to be more stable than those with higher energies. This criterion ensures that the RNA molecule can be 'trapped' by the energy basin where the SLOpt structure resides, without being able to get out of the basin easily. Figure 1 illustrates a schematic representation of the energy landscape of an RNA molecule. Each LOpt structure (e.g. the local optimum, number 5) can represent a set of none LOpt structures in its associated energy basin (e.g. the dot). In addition, although both 1 and 3 are local optima and 1 has even lower energy than 3, 1 is still less stable. This is because the conformational transition from 1 to 2 involves a lower energy barrier, while the transitions from 3 to any lower free energy LOpt structures yield higher energy barriers.

We formalize the problem as follows: given an RNA sequence A , an energy range ΔE and an energy barrier cutoff ΔB , find all the stable and local optimal structures, of which (i) the free energies are within ΔE of the MFE and (ii) the minimal energy barrier between any pair of SLOpt structures is $\geq \Delta B$. We will describe our approach (RNASLOpt) for addressing the problem in Section 2. In Section 3, we will compare RNASLOpt against the state-of-art methods and show benchmark tests on known riboswitches. In Section 4, we will discuss possible applications of our approach and conclude the article.

2 METHODS

First, we introduce configurations of stacks to represent scaffolds of RNA secondary structures. RNA secondary structures involve both contributing base pairs and isolated base pairs, where stacking base pairs contribute significantly to the stabilization of RNA secondary structures (Yakovchuk *et al.*, 2006). Structures with isolated base pairs are usually unrealistic and the removal of these structures from the search space may yield more significant structures (Zuker, 1989). Since LOpt structures reside at bottoms of basins

in the energy landscape, and each can represent a set of similar secondary structures, we introduce LOpt stack configurations to approximate LOpt structures. LOpt stack configurations are configurations that have a maximal number of putative stacks such that no stacks can be added rendering lower energy structures. We then present algorithms for finding all possible LOpt stack configurations based on both the Nussinov model (Nussinov *et al.*, 1978) and the Turner energy model (Freier *et al.*, 1986; He *et al.*, 1991; Jaeger *et al.*, 1989; Mathews *et al.*, 1999; Turner *et al.*, 1988), using the mathematical scheme advocated by Wuchty *et al.* (1999). Next, we describe a fast heuristic algorithm for computing pairwise energy barriers among LOpt stack configurations. The energy barrier between a pair of LOpt stack configurations indicates the amount of additional energy required for the RNA molecule to fold from one structure to the other, and can be used to filter out unstable LOpt structures. Finally, we employ a simple neighbor joining algorithm to cluster unstable LOpt structures, obtain stable local optimal structures and assign rank accordingly.

2.1 RNA secondary structures and stack configurations

Consider an RNA sequence as a string $A = a_1 \dots a_n$ of n letters over alphabet $\Sigma = \{A, U, G, C\}$. A pair of nucleotides a_i and a_j ($i < j$) can interact with each other and form a base pair (denoted by (i, j)), if they are complementary to each other. In this article, we only consider the canonical base pairings (G–C and A–U) and the wobble base pairing (G–U). A *secondary structure* of an RNA can be represented by an ensemble of pairing bases. A secondary structure is *pseudoknotted* if it contains two base pairs (i, j) and (i', j') such that $i < i' < j < j'$. In this article, we only consider pseudoknot-free secondary structures.

The stability of an RNA secondary structure is determined predominantly by energetically favorable helical regions, where both base pair stacking and hydrogen bonding provide stabilizing energy contributions (Yakovchuk *et al.*, 2006). We denote a helical region by a *stack*. A stack $p = (p_b, p_e, p_l)$ has p_l consecutive base pairs, where (p_b, p_e) is the outmost base pair and $(p_b + p_l - 1, p_e - p_l + 1)$ is the innermost base pair. Without loss of generality, p_l can be 0. We define two arbitrary stacks as *compatible* with each other if they are parallel or one stack encloses the other. We define partial orders $<_p$ and $<_l$ between compatible stacks as follows. If a stack p is parallel to a stack q , and p resides to the 5' of q (i.e. $p_e < q_b$), then $p <_p q$. If p encloses q [i.e. $(p_b + p_l) \leq q_b$ and $q_e \leq (p_e - p_l)$], then $q <_l p$. We denote the ensemble of all possible putative stacks of an RNA sequence by \mathcal{P} . We can compute \mathcal{P} using the algorithm of Bafna *et al.* (2006) in $O(n^2)$ time. Following their work, we score hydrogen bonds between pairing bases G–C, A–U and G–U by 3, 2 and 1, respectively, and set the minimum length of putative stacks (ℓ_{\min}) as 4 and the minimum score of hydrogen bonds (h_{\min}) as 8, because statistics show that the fraction of true stacks missed is <10% with the cutoff (Bafna *et al.*, 2006). The number of putative stacks predicted is usually much less than the number of feasible pairing bases. This yields a faster algorithm for enumerating suboptimal structures, which recursively branches when a putative stack (instead of a feasible base pair) is encountered. In addition, the typical lengths of riboswitches are around 100–200, and the number of putative stacks predicted for an RNA of similar length may even be smaller than the sequence lengths. For example, we predicted 62 putative stacks for the 110 nt long adenine riboswitch of *ydhL* gene from *B. subtilis*.

In order to elucidate the basic idea, we define a notion of *stack configuration*. A stack configuration of an RNA sequence is composed of a set of putative stacks in \mathcal{P} that are pairwise compatible. Figure 2 shows a schematic representation of a stack configuration. A stack configuration φ is *local optimal* if there does not exist any stack p in \mathcal{P} that p can be added to φ without affecting the validity of φ (i.e. forming a pseudoknot or pairing a base with more than one partner). Next, let p and q be putative stacks and q is enclosed with p , we also define the following terms:

$|p|$: the length of the subsequence covered by p (i.e. $p_e - p_b + 1$).

$\mathcal{N}(p)$: the set of all possible putative stacks on a subsequence covered by p (i.e. $a_{p_b} \dots a_{p_e}$).

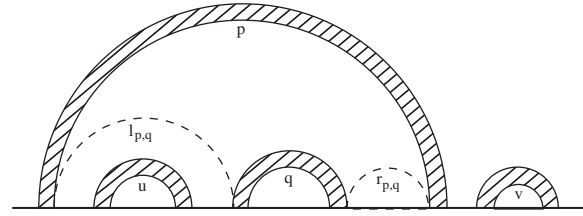


Fig. 2. A schematic representation of a stack configuration. Filled arcs represent putative stacks p, q, u and v . The relationships between these putative stacks are as follows: $p <_p v$, $u <_p q$, $u <_l p$, $q <_l p$ and $q \in \mathcal{F}_l(p)$. Dashed arcs represent $l_{p,q}$ and $r_{p,q}$, respectively.

$\mathcal{N}(p)$: all possible LOpt stack configurations composed of putative stacks in $\mathcal{P}(p)$.

$\mathcal{F}_l(p)$: a subset of putative stacks in $\mathcal{P}(p)$, where $\forall q \in \mathcal{F}_l(p)$, $\nexists q'$ such that $q' <_l p$ and $((q <_p q') \text{ or } (q <_l q'))$.

$l_{p,q}$: a stack $(p_b + p_l, q_b - 1, 0)$ that is enclosed by p and juxtaposes to the 5' end of q , provided that $q <_l p$.

$r_{p,q}$: a stack $(q_e + 1, p_e - p_l, 0)$ that is enclosed by p and juxtaposes to the 3' end of q , provided that $q <_l p$.

In the next two subsections, we will describe algorithms for generating all possible LOpt stack configurations based on the Nussinov model and the Turner model, respectively.

2.2 Stack-based folding using Nussinov model

2.2.1 Computing the maximum number of base pairs The RNA folding problem was formulated as a loop matching problem by Nussinov *et al.* (1978) and solved using dynamic programming. In the Nussinov model, the energy contribution of each base pair is 1, while base pair stacking and loop entropies have no energetic contributions. Given an arbitrary stack p , we define $N(p)$ as the maximal number of base pairs of all the stack configurations in $\mathcal{N}(p)$. The recursive formula for computing $N(p)$ is shown in Equation (1). If $\mathcal{F}_l(p)$ is an empty set, then no putative stack is enclosed with p and $N(p) = p_l$ (the number of base pairs in p). Otherwise, we can divide the sequence covered by p into three parts: (i) the stacking base pairs in p , (ii) an arbitrary stack q in $\mathcal{F}_l(p)$ and (iii) a stack $l_{p,q}$ which is enclosed with p and to the 5' of q . In this case, $N(p)$ is the sum of base pairs in the three parts. The time complexity for computing $N(p)$ is $O(|\mathcal{P}(p)|^2)$. In addition, we denote the entire RNA sequence by a stack $p^* = (1, n, 0)$ and can obtain the maximum number of base pairs over all possible stack configurations on the sequence by computing $N(p^*)$.

$$N(p) = p_l + \max_{q \in \mathcal{F}_l(p)} \{N(q) + N(l_{p,q})\} \quad (1)$$

2.2.2 Generating all possible LOpt stack configurations We present in Figure 3 an exact algorithm for enumerating all possible LOpt stack configurations with at least n_0 base pairs. We keep an array of *partial stack configurations* in R . Each partial stack configuration φ in R comprises an ordered list of stacks, which are labeled with either *finished* or *unfinished*. The label *finished* indicates that we have finished processing p and p should appear on all the stack configurations φ represents. The label *unfinished* means that the structures on the subsequence covered by p is not determined yet and p needs to be dealt with in the future. Each partial stack configuration φ can represent a set of LOpt stack configurations that contain all the *finished* stacks in φ . And, a partial stack configuration $(p^*, \text{unfinished})$ can represent all possible LOpt stack configurations on the entire RNA sequence. Besides, when all the stacks in φ are labeled with *finished*, φ only represents exactly one stack configuration.

The algorithm is as follows. First, we push $(p^*, \text{unfinished})$ to R . Then, we repeatedly pop up the last partial stack configuration φ from R and process φ according to the following procedures until R is empty. Given φ , we pop


```

procedure enumerate( $A, n_\theta$ )
 $p^* = (1, n, 0)$ ,  $\varphi = \{(p^*, \text{unfinished})\}$ ,  $R = \{(\varphi, N(p^*))\}$ 
while ( $R \neq \emptyset$ ) do
     $(\varphi, x) \leftarrow R$ ,  $(p, \text{label}) \leftarrow \varphi$ 
    if ( $\text{label}$  is unfinished) then
        for all stacks  $q$  in  $\mathcal{F}_I(p)$  do
             $(\varphi', x') = (\varphi, x - N(p))$ 
            if ( $p_l \neq 0$ ) then  $(p, \text{finished}) \Rightarrow \varphi'$  end if
             $(l_{p,q}, \text{unfinished}) \Rightarrow \varphi'$ ,  $(q, \text{unfinished}) \Rightarrow \varphi'$ 
             $x' = x' + p_l + N(q) + N(l_{p,q})$ 
            if ( $x' \geq n_\theta$ ) then  $(\varphi', x') \Rightarrow R$  end if
        end for
        if ( $\mathcal{F}_I(p)$  is  $\emptyset$  and  $x \geq n_\theta$ ) then  $(\varphi, x) \Rightarrow R$  end if
    else ( $\text{label}$  is finished) */
        if ( $x \geq n_\theta$ ) then output  $\varphi$  end if
    end if
end while
    
```

Fig. 3. An algorithm *enumerate*(A, n_θ) for enumerating all possible local optimal stack configurations on an RNA sequence A with at least n_θ base pairs. The meaning of \Rightarrow , \Leftarrow and \Leftarrow are pushing back an element to the end of an array, inserting an element to the beginning of an array, and popping up the last element from an array, respectively.

the last element (a stack p) from the array of φ and check its associated label. If the label of p is *finished*, then all the stacks in φ should have been processed. (Because we always insert stacks labeled with *finished* to the front of the array of φ and push stacks labeled with *unfinished* to the end.) In this case, we output the only stack configuration that φ represents. Otherwise, we decompose the unfinished stack p into three disjoint components: (i) the stacking base pairs of p , (ii) a stack $q \in \mathcal{F}_I(p)$ and (iii) a stack $l_{p,q}$. We can construct a stack configuration on the subsequence covered by p by combining (i) the stack p , (ii) a stack configuration taken from $\mathcal{N}(q)$ and (iii) a stack configuration taken from $\mathcal{N}(l_{p,q})$. If q is determined, we can construct $|\mathcal{N}(q)| \times |\mathcal{N}(l_{p,q})|$ possible new stack configurations. And, for each stack q in $\mathcal{F}_I(p)$, we construct a new stack configuration φ' by pushing $(p, \text{finished})$ to the end of φ and inserting $(l_{p,q}, \text{unfinished})$ and $(q, \text{unfinished})$ to the beginning of φ . We can compute the size of $\mathcal{N}(p)$ using Equation (2).

$$|\mathcal{N}(p)| = \sum_{q \in \mathcal{F}_I(p)} |\mathcal{N}(q)| \times |\mathcal{N}(l_{p,q})| \quad (2)$$

Next, we push all the new partial stack configurations that have at least n_θ base pairs to the end of R . We denote the maximal number of base pairs of a partial stack configuration φ by $N(\varphi)$. As shown in Equation (3), $N(\varphi)$ is the sum of $N(p)$ over all stacks p in φ . Each stack labeled with *finished* contributes exactly p_l base pairs, and each stack labeled with *unfinished* contributes at most $N(p)$ base pairs [where $N(p)$ can be computed using Equation (1)].

$$N(\varphi) = \sum_{p \in \varphi} \begin{cases} p_l & \text{the label of } p \text{ is } \textit{finished} \\ N(p) & \text{the label of } p \text{ is } \textit{unfinished} \end{cases} \quad (3)$$

2.3 Stack-based RNA folding using the Turner model

According to the Turner model, the free energy of a stack configuration is the additive sum of energy contributions of all the stacking base pairs, hairpin loops, bulges, interior loops, multiloops and dangling bases (Mathews *et al.*, 1999). We describe the energy parameters and terminal symbols used in the following:

- M_c : offset penalty for opening a multibranching loop.
- M_b : free base penalty for each unpaired base in a multibranching loop.
- M_i : helix penalty for each helix in a multibranching loop.
- $H(p)$: destabilizing energy of the hairpin loop enclosed with a stack p .

$I(p, q)$: destabilizing energy of the interior loop or bulge between stacks p and q .

$S(p)$: stabilizing energies of all the stacking base pairs in a stack p .

M_c , M_b and M_i are constant energy parameters. $H(p)$ and $I(p, q)$ can be obtained from the tabulated energy parameters, and $S(p)$ can be computed as the sum of tabulated stacking energies of adjacent stacking base pairs in p . All the free energy parameters are taken from the work of Mathews *et al.* (1999). We also define the following non-terminal symbols as follows:

$F(p)$: the MFE of all stack configurations in $\mathcal{N}(p)$, provided that $p_b = 1$ and $p_l = 0$.

$C(p)$: the MFE of all stack configurations in $\mathcal{N}(p)$, provided that $p_l \neq 0$ and p closes the structure on $a_{p_b} \dots a_{p_e}$.

$FM1(p)$: the MFE of all stack configurations in $\mathcal{N}(p)$, provided that p is within a multibranching loop, and there exists at least a stack q such that $q_l \neq 0$ and $q <_I p$.

$FM(p)$: the MFE of all stack configurations in $\mathcal{N}(p)$, provided that p is within a multibranching loop.

2.3.1 Computing the minimum free energy The recursive formula for computing the minimum free energy is shown in Equation (4), with a time complexity of $O(|\mathcal{P}(p)|^3)$ [which is $O(n^6)$ with a small factor]. For the sake of simplicity, we do not discuss dangling energy contributions in the recursive formula, but take them into account in the implementation.

$$\begin{aligned}
 F(p) &= \min_{q \in \mathcal{F}_I(p)} \{C(q) + F(l_{p,q})\} \\
 C(p) &= S(p) + \min \begin{cases} H(p), \\ \min_{q <_I p} \{C(q) + I(p, q)\}, \\ \min_{q \in \mathcal{F}_I(p)} \left\{ \begin{aligned} &C(q) + FM1(l_{p,q}) + M_c \\ &+ 2 * M_i + |r_{p,q}| * M_b \end{aligned} \right\} \end{cases} \quad (4) \\
 FM1(p) &= \min_{q \in \mathcal{F}_I(p)} \{C(q) + FM(l_{p,q}) + M_i + |r_{p,q}| * M_b\} \\
 FM(p) &= \min \begin{cases} |p| * M_b, \\ \min_{q \in \mathcal{F}_I(p)} \{C(q) + FM(l_{p,q}) + M_i + |r_{p,q}| * M_b\} \end{cases}
 \end{aligned}$$

2.3.2 Generating all possible LOpt stack configurations In this section, we describe an algorithm for numerating all possible local optimal stack configurations of an RNA sequence A within ΔE of the MFE. The detailed procedures are described in Supplementary Figures S4–S8. We denote the free energy upper limit for stack configurations by e_θ , where e_θ is equivalent to the MFE of all possible stack configurations plus ΔE . We keep an array of paired objects $R = \{(\varphi, E(\varphi)), (\varphi', E(\varphi')), \dots\}$. Each paired object of R comprises of a partial stack configuration φ and its associated minimum free energy $E(\varphi)$. Each partial stack configuration φ comprises an ordered list of stacks, each with a label (i.e. $\varphi = \{(p, \text{label}), (p', \text{label}'), \dots\}$). There are five types of labels, including *finished*, F , C , $FM1$ and FM . The label *finished* indicates that we have finished processing stack p , and p will appear on all the stack configurations that φ represents. The remaining labels correspond to the following cases: $F(p)$, $C(p)$, $FM1(p)$, and $FM(p)$, respectively.

The algorithm starts with a partial stack configuration $\varphi_0 = (p^* = (1, n, 0), F)$ and its associated minimum free energy $E(\varphi_0)$. φ_0 represents all possible stack configurations on A , and $E(\varphi_0)$ is the minimum free energy of φ_0 (i.e. $E(\varphi_0) = F(p^*)$). We push $(\varphi_0, E(\varphi_0))$ to R and repetitively process the last element of R according to the following procedure until R is empty. Let $(\varphi, E(\varphi))$ be the last partial stack configuration and its associated energy in R , and let (p, label) be the last stack and its associated label in φ . First, we check the label of p . Similar to the algorithm based on the Nussinov model, we also ensure that stacks labeled with *finished* are inserted to the front of the array of φ and other stacks are pushed back to the end of the array. If the label of p is *finished*, then all the stacks should have been processed. In this case, we output φ if E_φ is less than e_θ . Otherwise, we will construct a set of new partial stack configurations according to the label. Each new

partial stack configuration φ' is constructed by combining all the remaining stacks other than p in φ (denoted by φ^- , where $\varphi^- = \varphi - \{(p, \text{label})\}$) with stacks enclosed with p . Next, we compute $E(\varphi')$ for each new partial stack configuration φ' , and push them to the end of R if $E(\varphi')$ is less than or equal to e_θ , as described in the following:

Case F: p ($p_b=1$ and $p_l=0$) is a stack. For each stack q in $\mathcal{F}_l(p)$, we construct a new partial stack configuration φ' by pushing (q, C) and $(l_{p,q}, F)$ to the end of φ^- . $E(\varphi')$ is given by Equation (5).

$$E(\varphi') = E(\varphi) - F(p) + C(q) + F(l_{p,q}) \quad (5)$$

Case C: p ($p_l \neq 0$) should appear on all the stack configurations that φ represents. We construct a set of new partial stack configurations according to Cases C.1, C.2 and C.3.

C.1: p closes a hairpin loop. We construct a new partial stack configuration φ' by inserting $(p, \text{finished})$ to the front of φ^- . $E(\varphi')$ is given by Equation (6).

$$E(\varphi') = E(\varphi) - C(p) + \underline{S}(p) + \underline{H}(p) \quad (6)$$

C.2: p closes a stack q and forms an interior loop (or a bulge) with q . For each stack $q <_l p$, we construct a partial stack configuration φ' by inserting $(p, \text{finished})$ to the front of φ^- and then pushing (q, C) to the end. $E(\varphi')$ is given by Equation (7).

$$E(\varphi') = E(\varphi) - C(p) + \underline{S}(p) + \underline{I}(p, q) + C(q) \quad (7)$$

C.3: p closes a multibranched loop. For each stack $q \in \mathcal{F}_l(p)$, we construct a new partial stack configuration φ' by inserting $(p, \text{finished})$ to the front of φ^- , and then pushing (q, C) and $(l_{p,q}, FM1)$ to the end. $E(\varphi')$ is given by Equation (8).

$$E(\varphi') = E(\varphi) - C(p) + \underline{S}(p) + C(q) + FM1(l_{p,q}) + \underline{M}_c + 2 * \underline{M}_i + |r_{p,q}| * \underline{M}_b \quad (8)$$

Case FM1: p ($p_l=0$) is directly enclosed with a multibranched loop, and there exists at least a stack q such that $q_l \neq 0$ and $q <_l p$. For each stack q in $\mathcal{F}_l(p)$, we construct a new partial stack configurations φ' by pushing (q, C) and $(l_{p,q}, FM)$ to the end of φ^- . $E(\varphi')$ is given by Equation (9).

$$E(\varphi') = E(\varphi) - FM1(p) + C(q) + FM(l_{p,q}) + \underline{M}_i + |r_{p,q}| * \underline{M}_b \quad (9)$$

Case FM: p ($p_l=0$) is directly enclosed with a multibranched loop. We construct a set of new partial stack configurations according to cases FM.1 and FM.2.

FM.1: all the bases covered by p are unpaired. We construct a partial stack configuration $\varphi' = \varphi^-$. $E(\varphi')$ is computed as Equation (10).

$$E(\varphi') = E(\varphi) - FM(p) + |p| * \underline{M}_b \quad (10)$$

FM.2: there exists a stack q ($q_l \neq 0$) enclosed with p . For each stack $q <_l p$, we construct a partial stack configuration φ' by pushing (q, C) and $(l_{p,q}, FM)$ to the end of φ^- . $E(\varphi')$ is given by Equation (11).

$$E(\varphi') = E(\varphi) - FM(p) + C(q) + FM(l_{p,q}) + \underline{M}_i + |r_{p,q}| * \underline{M}_b \quad (11)$$

2.3.3 Redefining partial orders $<_l$ and $<_p$ Stack configurations produced by our approach consist of pairwise compatible stacks, therefore incompatible stacks that overlap one another by only a few bases cannot coexist in a structure. To solve this problem, we use looser definitions of partial orders $<_l$ and $<_p$, which allow compatible stacks to share a small portion of bases in common. RNASLOpt is able to produce stack configurations containing incompatible stacks overlapping by a few (by default, no $>20\%$) bases.

2.4 Obtaining SLOpt structures

Using the algorithm described above, we can produce a set of all possible LOpt stack configurations on an RNA sequence, and denote it by R . However, although the conformational space of LOpt stack configurations is dramatically reduced compared with the space of feasible secondary

structures, the number of structures considered may still be enormous. In literature, many distance metrics, such as base pair metrics (Zuker, 1989; Zuker *et al.*, 1991), tree metrics (Shapiro and Zhang, 1990), mountain metrics (Moulton *et al.*, 2000), metrics based on base pairing probability matrices (Hofacker *et al.*, 2004) and metrics using the Lempel–Ziv algorithm (Liu and Wang, 2006; Zhang and Wang, 2010) have been proposed for filtering out similar structures and reducing the number of structures considered. In contrast, we are only interested in stable local optimal (SLOpt) structures. And, we will filter out unstable structures from the space instead of removing similar structures that share base pairs, shapes or pairing probabilities in common. The SLOpt structures should be difficult for an RNA molecule to escape, and the associated energy barrier between any pair of SLOpt structures should be greater than or equal to a certain threshold ΔB . Using pairwise energy barriers among LOpt stack configurations as a distance matrix, we can evaluate the stability of RNA secondary structures in the context of energy landscape.

The problem of determining the minimal energy barrier between two conformational structures has been well studied, and it is usually solved in conjunction with finding the optimal folding pathways with the minimal energy barrier. Many approaches have been proposed to address the problem. These approaches can either be based on the Nussinov model [e.g. an exact algorithm proposed by Thachuk *et al.* (2010) and a greedy algorithm by Morgan and Higgs (1998)] or the Turner model [e.g. an exact solution devised by Flamm *et al.* (2002) and heuristic algorithms developed by Morgan and Higgs (1998), Flamm *et al.* (2001), Voss *et al.* (2004), Geis *et al.* (2008) and Dotu *et al.* (2010)]. In this article, we focus on using energy barriers to find SLOpt stack configurations (instead of determining the optimal folding pathways). Therefore, here, we propose a fast heuristic for computing pairwise energy barriers among LOpt stack configurations. Upon these pairwise energy barriers, we cluster unstable LOpt stack configurations using a simple neighbor joining algorithm, and obtain all the SLOpt stack configurations with the minimal pairwise energy barrier no less than ΔB . Finally, we rank these SLOpt structures either according to their free energies or their minimal associated energy barriers.

2.4.1 Approximating barrier energy Consider secondary structures S and S' , the folding pathway between S and S' involves a series of intermediate structures, among which, the saddle point structure S^* is the one with the highest free energy (e.g. in Fig. 1, a is the saddle point for the folding pathway from local optima 1 to 2). We denote the energy barrier from S to S' by $\mathcal{B}(S \rightarrow S')$ and denote the energy barrier between S and S' by $\mathcal{B}(S \rightleftharpoons S')$. $\mathcal{B}(S \rightarrow S')$ is equivalent to the absolute difference in the free energies of S and S^* [i.e. $|E(S) - E(S^*)|$] and $\mathcal{B}(S \rightleftharpoons S')$ can be computed using Equation (12).

$$\mathcal{B}(S \rightleftharpoons S') = \min\{\mathcal{B}(S' \rightarrow S), \mathcal{B}(S \rightarrow S')\} \quad (12)$$

We list our assumptions for approximating barrier energy $\mathcal{B}(S \rightarrow S')$ in the following. The saddle point S^* between S and S' can be achieved when all the base pairs in S are opened or shifted such that S' can be formed without opening more base pairs. The amount of additional energy required for opening an entire stack p is roughly $\underline{S}(p)$, and the amount for opening a base pair in p is about $\frac{1}{p_l} * \underline{S}(p)$, while the amount for sliding one endpoint of a base pair in p is $\frac{\alpha}{p_l} * \underline{S}(p)$, ($0 \leq \alpha \leq 1$, by default, α is 0.5). Given a base pair (i, j) in S and an arbitrary stack p' in S' , we determine the necessary operation to apply to (i, j) (i.e. operations that can make the formation of p' possible) according to the positional relationship between (i, j) and p' . Let $w((i, j), p')$ denote the additional energy associated with the operation. We describe the four types of positional relationships and the corresponding $w((i, j), p')$ in Table 1. Case 1 (i, j) is *compatible* to p' (i.e. either be nested or juxtapose to each other). In this case, we cannot infer the operation to apply to the base pair, because the stack can be formed anyway. Case 2, (i, j) is *consistent* with p' [(i, j) is in p']. We do not apply any operation to the base pair so as to keep it intact during the folding. Case 3 (i, j) *partially conflicts* to p' [i.e. there exist two base pairs (i, i') and (j', j) in p']. In this case, we may slide either

Table 1. Positional relationships between a base pair (i, j) and a stack p'

Cases	Relationships	Descriptions	$w((i, j), p')$
1	Compatible	(i, j) and p' either are nested or juxtapose to each other	— (not applicable)
2	Consistent	(i, j) is in p'	0
3	Partially-Conflict	there exist base pairs (i, i') and (j', j) in p'	$\frac{\alpha}{p_i}$
4	Conflict	Otherwise	$\frac{1}{p_i}$

endpoint i or j to its new partner (i' or j') to form p' . Case 4 (i, j) conflicts to p' . In this case, we have to open (i, j) in order to make the formation of p' possible. Since S' usually contains more than one stack, we use the smallest $w((i, j), p')$ over all the stacks p' in S' , to represent the least amount of additional energy required so as to form S' . If (i, j) is compatible with all the stacks in S' , we have to delete (i, j) , which requires $\frac{1}{p_i} * \underline{S}(p)$ additional energy. We present the approximated algorithm for computing $\mathcal{B}(S \rightarrow S')$ in Equation (13).

$$\mathcal{B}(S \rightarrow S') = \sum_{p \in S} \sum_{(i, j) \in p} \min_{p' \in S'} \{w((i, j), p') * \underline{S}(p)\} \quad (13)$$

2.4.2 Pairwise energy barrier-based clustering A LOpt stack configuration φ is considered as stable if the minimal energy barrier between φ and any other stable structures is no less than $\Delta\mathcal{B}$. φ can be seen as a representative of all the unstable structures in the energy basin it resides. Let R^* denote the set of SLOpt stack configurations. We describe the procedure for constructing R^* from the set of LOpt stack configurations R in Supplementary Figure S9. First, we sort LOpt stack configurations in R by their free energies (i.e. the lower the free energy is, the higher the stack configuration ranks). Then, we push the MFE LOpt stack configuration (i.e. $R[0]$) to R^* . Next, we define a lower triangular matrix M^* for saving pairwise energy barriers of SLOpt stack configurations in R^* , where $M^*[k, l]$ represents the energy barrier between $R^*[k]$ and $R^*[l]$ [i.e. $\mathcal{B}(R^*[k] \rightarrow R^*[l])$]. We analyze each LOpt stack configuration φ in R . If the energy barrier between φ and any SLOpt stack configuration in R^* is less than $\Delta\mathcal{B}$, we consider φ as unstable, and discard it. Otherwise, we push φ to R^* as a SLOpt stack configuration and update M^* accordingly. When M^* is constructed completely, we stepwisely neighbor join SLOpt stack configurations in R^* which have the lowest pairwise energy barrier in M^* , and obtain a cluster tree. Finally, we rank SLOpt structures in R^* either by their free energies or by their associated minimal energy barriers.

3 RESULTS

3.1 Reducing the conformational space

The number of feasible secondary structures within a certain energy range of the MFE can be enormous. Therefore, instead of investigating the vast conformational space of feasible secondary structures, we want to reduce the size of the conformational space to consider. First, we only enumerate LOpt stack configurations instead of feasible structures, the number of which is greatly reduced compared with that of feasible structures. In addition, we can further reduce the number of candidates to consider by filtering out unstable structures and only investigate SLOpt stack configurations. Note that the reduced space still grows exponentially with the RNA length and the energy range. Comparisons of sizes of different conformational spaces are shown in Supplementary Figures S2 and S3.

Table 2. The numbers of structures produced by mfold, RNASHapes and RNASLOpt

Riboswitch	Len	SubOpt	mfold	RNASHapes	RNASLOpt	
					LOpt	SLOpt
		(%)				
Adenine-BS	110	55	43	25	19	5
Adenine-VV	113	20	20	9	14	4
Guanine	148	55	38	759	1216	70
SAM	134	20	18	53	410	31
c-di-GMP	124	20	25	81	259	38
Lysine	233	20	20	>1000	4798	346
TPP	185	20	33	247	1384	91

Len, lengths of riboswitches. SubOpt, the percentage suboptimality.

3.2 Predicting alternative structures for riboswitches

We will show that although the conformational space of SLOpt stack configurations is greatly reduced compared with the space of feasible structures, it does not miss native structures for all the benchmark tests. Therefore, we can predict alternate structures for riboswitches by exploring the space of SLOpt stack configurations. We performed benchmark tests on seven riboswitches, including the adenine riboswitch of the *ydhL* gene from *B.subtilis* (Mandal and Breaker, 2004) (denoted by adenine-BS), the adenine riboswitch of *add* gene from *Vibrio vulnificus* (Lemay *et al.*, 2011) (denoted by adenine-VV), the guanine riboswitch of *xpt-pbuX* operon from *B.subtilis* (Mandal *et al.*, 2003), the *S*-adenosylmethionine (SAM) riboswitch of *metE* from *Thermoanaerobacter tencongensis* (Epshtein *et al.*, 2003), the c-di-GMP riboswitch of *tfoX* from *Candidatus desulfurudis* (Smith *et al.*, 2009), the lysine riboswitch of *lysC* from *B.subtilis* (Blouin *et al.*, 2011) and the thiamin pyrophosphate (TPP) riboswitch of *thiamin* from *B.subtilis* (Mironov *et al.*, 2002; Rentmeister *et al.*, 2007). We describe the parameters used in the tests as follows. By default, the minimum length of putative stacks (i.e. ℓ_{\min}) is 4, and the minimum score for hydrogen bonds (i.e. h_{\min}) is 8. However, ℓ_{\min} is 3 for the SAM riboswitch and c-di-GMP riboswitch, because a large proportion of stacks in the native structures of both cases are of lengths ≤ 3 . Percentage suboptimality is a parameter that determines the free energy upper limit for the predicted structures. If percentage suboptimality is $x\%$, then only structures that have free energies $\leq (1 - x\%)$ of the MFE will be computed. The default value is 20%, since usually the native structures are within a lower energy range from the MFE. However, for the adenine-BS riboswitch and the guanine riboswitch, suboptimality is assigned a greater value (i.e. 55%), because the free energies of the 'on' structures for both riboswitches are $>20\%$ of the MFE. The default energy barrier cutoff $\Delta\mathcal{B}$ is 12 (kcal/mol), which is empirically chosen to reflect the stability of alternative structures, and it can be changed by users.

First, we compare the number of structures produced by mfold (v3.5), the number of 'shreps' by RNASHapes (v2.1.6) and the numbers of LOpt and SLOpt stack configurations by RNASLOpt in Table 2. The number of SLOpt produced by RNASLOpt is less than that of RNASHapes in all the cases. It reveals that although the number of candidates considered by both methods are exponential, the space of RNASLOpt is reduced compared with the space of RNASHapes. Interestingly, the number of candidates produced by RNASLOpt is even less than that of mfold [which generates $O(n^2)$]

Table 3. The ranks of the best structures corresponding to the native ‘off’ and ‘on’ structures produced by mfold, RNAShapes, RNALocopt and RNASLOpt

Riboswitch	SubOpt (%)	mfold	RNAShapes	RNALocopt			RNASLOpt	
				$n = 10$	$n = 100$	$n = 1000$	RankE	RankB
Adenine-BS	55	(1, 18)	(1, –)	(3, –)	(3, –)	(3, –)	(1, 4)	(1, 2)
Adenine-VV	20	(3, 1)	(4, 1)	(7, –)	(28, –)	(42, 25)	(2, 1)	(4, 1)
Guanine	55	(1, 25)	(1, 66)	(1, –)	(1, –)	(1, –)	(1, 15)	(1, 3)
SAM	20	(6, 11)	(8, 14)	(–, –)	(66, 60)	(180, 98)	(1, 5)	(1, 13)
c-di-GMP	20	(10, 12)	(22, 3)	(–, 1)	(38, 1)	(68, 1)	(6, 14)	(10, 4)
Lysine	20	(4, 5)	(22, 35)	(1, –)	(2, 92)	(658, 806)	(24, 31)	(18, 22)
TPP	20	(1, 17)	(1, 24)	(1, –)	(2, –)	(190, 410)	(1, 5)	(1, 3)

For each pair of numbers in the table, the two numbers denote ranks of the best structures corresponding to the native ‘off’ and ‘on’ structures, respectively. SubOpt represents percentage suboptimality used by mfold, RNAShapes and RNASLOpt. RNAShapes were run using the most abstract shape type. RNALocopt were run with sample size $n = 10$ (the default value), 100 and 1000 (instead of using suboptimality). RankE and RankB represent that secondary structures are ranked by their free energies and minimal associated energy barriers, respectively. Bold faced numbers indicate the best pair of ranks produced among all the approaches. ‘–’ represents no secondary structure similar to the specified native structure found.

structures at most], when the RNA sequence is short (e.g. the adenine riboswitch). The running time for all the test cases on a 32 bit PC with 2.4 GHz Quad-processor, 3.2 GB memory (running Fedora 11) are 0.04, 0.04, 6, 2.4, 1.5, 227.6 and 15.8 s, respectively (see Supplementary Table S1). Usually, RNASLOpt can be applied on RNAs of around 200 nt long and finish the computation within a few minutes.

Next, we compare the ranks of the best structures corresponding to the native structures produced by mfold, RNAShapes, RNALocopt and RNASLOpt in Table 3. The best structures should share the most backbone structures in common with the native structures. RNASLOpt can rank predicted structures both according to their free energies and minimal associated energy barriers. In all the cases, RNASLOpt ranks the best structures corresponding to the native ‘on’ and ‘off’ structure conformations among the top. And, in six out of seven cases, RNASLOpt provides better ranks than the others.

For example, Figure 4 show both the native ‘on’ and ‘off’ structures of adenine riboswitch from the *ydhL* gene of *B.subtilis* (Breaker, 2004) and the best stack configurations produced by RNASLOpt. RNASubopt produces $>10^9$ feasible secondary structures, mfold selects 43 representative structures and RNAShapes predicts 25 shreps (with the most abstract option). In contrast, RNASLOpt enumerates 19 LOpt stack configurations within 55% of the MFE, filters out 14 unstable stack configurations, and obtains 5 SLOpt stack configurations. Two SLOpt stack configurations among the five have the similar backbone structures to the native conformations and are ranked among the top according to both free energies (i.e. ranked 1 and 4, respectively) and the minimal associated energy barriers (i.e. ranked 1 and 2, respectively). Since the ‘on’ and ‘off’ structures predicted by RNASLOpt are LOpt stack configurations, an extra stack was predicted for each configuration (Fig. 4C and D) without affecting the backbone structure. We also list the native ‘on’ or ‘off’ conformations of the seven riboswitches, together with the best structures produced by mfold, RNAShapes, RNALocopt and RNASLOpt in Supplementary Figures S10–S16.

4 CONCLUSION

In this article, we designed an approach RNASLOpt for predicting stable local optimal stack configurations of an RNA molecule.

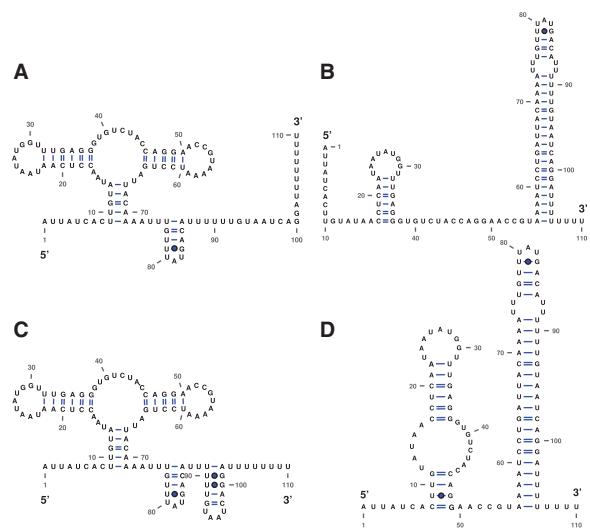


Fig. 4. The native and predicted ‘on’ and ‘off’ structure conformations of the adenine riboswitch from the *ydhL* gene of *B.subtilis*. (A) and (B) show the native ‘on’ and ‘off’ structure conformations; (C) and (D) plot the best corresponding stack configurations predicted by RNASLOpt.

We first predict all possible local optimal stack configurations that are significantly different from one another. With each stack configuration representing a set of similar RNA secondary structures, we are able to greatly reduce the size of the conformational space considered, and make applications on longer sequences with a higher energy range possible. In addition, we also employ a fast heuristic to compute pairwise energy barriers among LOpt stack configurations. Finally, we filter out unstable structures based on their pairwise energy barriers, obtain stable structures and rank them either according to their free energies or their minimal associated energy barriers.

RNASLOpt can produce interesting stable local optimal stack configurations for an RNA, and can be used to predict alternate structure conformations for riboswitches. We benchmarked RNASLOpt on seven riboswitches and the results show that RNASLOpt is able to successfully predict both native ‘on’ and ‘off’

structures for all the riboswitches. In addition, RNASLOpt is able to assign better ranks than the other methods in most cases. Since RNASLOpt is still an exponential algorithm, it is best applied to RNAs of up to 250 nt long. In conclusion, RNASLOpt can possibly be used in the following applications: (i) detecting alternate structure conformations for novel riboswitches; (ii) producing stable local optimal structures for an RNA molecule, which comprise a compact representation of the energy landscape; and (iii) unraveling the folding kinetics of an RNA molecule by exploiting the compact energy landscape and then determining its functionality. Stack configurations of an RNA molecule contain only the backbone helices and represent scaffolds of secondary structures. When finer grained structures conforming to the backbone structure are in need, we can take stack configurations as constraints and obtain the MFE structures that conform to the constraints using constraint folding (e.g. RNAfold with -C option, Hofacker *et al.*, 1994).

Conflict of Interest: none declared.

REFERENCES

- Bafna,V. *et al.* (2006) Consensus folding of unaligned RNA sequences revisited. *J. Comput. Biol.*, **13**, 283–295.
- Batey,R.T. *et al.* (2004) Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature*, **432**, 411–415.
- Baumstark,T. *et al.* (1997) Viroid processing: switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation. *EMBO J.*, **16**, 599–610.
- Blouin,S. *et al.* (2011) Folding of the lysine riboswitch: importance of peripheral elements for transcriptional regulation. *Nucleic Acids Res.*, **39**, 3373–3387.
- Breaker,R.R. (2004) Natural and engineered nucleic acids as tools to explore biology. *Nature*, **432**, 838–845.
- Dotu,I. *et al.* (2010) Computing folding pathways between RNA secondary structures. *Nucleic Acids Res.*, **38**, 1711–1722.
- Epshtein,V. *et al.* (2003) The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc. Natl Acad. Sci. USA*, **100**, 5052–5056.
- Evers,D. and Giegerich,R. (2001) Reducing the conformation space in RNA structure prediction. In *Proceedings of the German Conference on Bioinformatics*. German Research Center for Biotechnology, Braunschweig, Germany, pp. 118–124.
- Flamm,C. *et al.* (2001) Design of multistable RNA molecules. *RNA*, **7**, 254–265.
- Flamm,C. *et al.* (2002) Barrier trees of degenerate landscapes. *Z. Phys. Chem.*, **216**, 155–174.
- Freier,S.M. *et al.* (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl Acad. Sci. USA*, **83**, 9373–9377.
- Geis,M. *et al.* (2008) Folding kinetics of large RNAs. *J. Mol. Biol.*, **379**, 160–173.
- Giegerich,R. *et al.* (2004) Abstract shapes of RNA. *Nucleic Acids Res.*, **32**, 4843–4851.
- He,L. *et al.* (1991) Nearest-neighbor parameters for G.U mismatches. *Biochemistry*, **30**, 11124–11132.
- Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
- Hofacker,I.L. *et al.* (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.
- Jaeger,J.A. *et al.* (1989) Improved predictions of secondary structures for RNA. *Proc. Natl Acad. Sci. USA*, **86**, 7706–7710.
- Lemay,J.F. *et al.* (2011) Comparative study between transcriptionally- and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genet.*, **7**, e1001278.
- Liu,N. and Wang,T. (2006) A method for rapid similarity analysis of RNA secondary structures. *BMC Bioinformatics*, **7**, 493.
- Lorenz,W.A. and Clote,P. (2011) Computing the partition function for kinetically trapped RNA secondary structures. *PLoS One*, **6**, e16178.
- Mandal,M. and Breaker,R.R. (2004) Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol.*, **11**, 29–35.
- Mandal,M. *et al.* (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*, **113**, 577–586.
- Mathews,D.H. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
- Mironov,A.S. *et al.* (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, **111**, 747–756.
- Morgan,S. and Higgs,P. (1998) Barrier heights between ground states in a model of RNA secondary structure. *J. Phys. A: Math. Gen.*, **31**, 3153–3170.
- Moulton,V. *et al.* (2000) Metrics on RNA secondary structures. *J. Comput. Biol.*, **7**, 277–292.
- Nakaya,A. *et al.* (1996) Classification of RNA secondary structures using the techniques of cluster analysis. *J. Theor. Biol.*, **183**, 105–117.
- Noeske,J. *et al.* (2005) An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 1372–1377.
- Nussinov,R. *et al.* (1978) Algorithms for loop matchings. *SIAM J. App. Math.*, **35**, 68–82.
- Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
- Pipas,J.M. and McMahon,J.E. (1975) Method for predicting RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **72**, 2017–2021.
- Rentmeister,A. *et al.* (2007) Conformational changes in the expression domain of the *Escherichia coli* thiM riboswitch. *Nucleic Acids Res.*, **35**, 3713–3722.
- Schultes,E.A. and Bartel,D.P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.
- Shapiro,B.A. and Zhang,K.Z. (1990) Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci.*, **6**, 309–318.
- Smith,K.D. *et al.* (2009) Structural basis of ligand binding by a c-di-GMP riboswitch. *Nat. Struct. Mol. Biol.*, **16**, 1218–1223.
- Steffen,P. *et al.* (2006) RNAsHapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
- Thachuk,C. *et al.* (2010) An algorithm for the energy barrier problem without pseudoknots and temporary arcs. *Pac. Symp. Biocomput.*, **15**, 108–119.
- Turner,D.H. *et al.* (1988) RNA structure prediction. *Annu. Rev. Biophys. Chem.*, **17**, 167–192.
- Voss,B. *et al.* (2004) Evaluating the predictability of conformational switching in RNA. *Bioinformatics*, **20**, 1573–1582.
- Wakeman,C.A. *et al.* (2007) Structural features of metabolite-sensing riboswitches. *Trends Biochem. Sci.*, **32**, 415–424.
- Waterman,M.S. and Byers,T.H. (1985) A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Math. Biosci.*, **77**, 179–188.
- Wuchty,S. *et al.* (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
- Yakovchuk,P. *et al.* (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.*, **34**, 564–574.
- Zhang,S. and Wang,T. (2010) A complexity-based method to compare RNA secondary structures and its application. *J. Biomol. Struct. Dyn.*, **28**, 247–258.
- Zuker,M. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621.
- Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
- Zuker,M. *et al.* (1991) A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res.*, **19**, 2707–2714.