Published in final edited form as:
Int J Bioinform Res Appl. 2014 ; 10(4): 498–518. doi:10.1504/IJBRA.2014.062997.

Finding consensus stable local optimal structures for aligned RNA sequences and its application to discovering riboswitch elements

Yuan Li, **Cuncong Zhong**, and **Shaojie Zhang**
Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, Florida 32816, USA

Yuan Li: liy@eecs.ucf.edu; Cuncong Zhong: cczhong@eecs.ucf.edu; Shaojie Zhang: shzhang@eecs.ucf.edu

Abstract

Many non-coding RNAs (ncRNAs) can fold into alternate native structures and perform different biological functions. The computational prediction of an ncRNA's alternate native structures can be conducted by analysing the ncRNA's energy landscape. Previously, we have developed a computational approach, RNASLOpt, to predict alternate native structures for a single RNA. In this paper, in order to improve the accuracy of the prediction, we incorporate structural conservation information among a family of related ncRNA sequences to the prediction. We propose a comparative approach, RNAConSLOpt, to produce all possible consensus SLOpt stack configurations that are conserved on the consensus energy landscape of a family of related ncRNAs. Benchmarking tests show that RNAConSLOpt can reduce the number of candidate structures compared with RNASLOpt, and can predict ncRNAs' alternate native structures accurately. Moreover, an application of the proposed pipeline to bacteria in *Bacillus* genus has discovered several novel riboswitch candidates.

Keywords

bioinformatics; RNA consensus secondary structure; RNA energy landscape; RNA stable local optimal structure; riboswitch

1 Background

Non-coding RNAs (ncRNAs) play important roles in the biological regulatory system by folding into specific structures. Many ncRNAs, such as riboswitches, can transit among more than a single native structure in order to participate in different biological activities (Schultes and Bartel, 2000). For example, the adenine riboswitch of *ydhL* gene of *Bacillus subtilis* can selectively couple the adenine metabolites, causing a structural rearrangement that can turn 'off' the formation of a transcription terminator and preclude the gene transcription of its downstream genes (Mandal and Breaker, 2004). Determination of ncRNAs' alternate functional structures can provide deep insights into the regulatory

Correspondence to: Shaojie Zhang, shzhang@eecs.ucf.edu.

mechanisms of ncRNAs in cellular life. Furthermore, analysis of putative RNAs' potential structure conformations can lead to discovery of novel riboswitches.

1.1 Stable local optimal structures and energy landscape of a single RNA

The alternate functional structures of an ncRNA can be determined by analysing its energy landscape. The exact energy landscape of an RNA consists of all feasible suboptimal structures within a certain energy range, where each suboptimal structure is directly connected to its neighbouring structures (i.e. structures that differ from it by exactly one base pair). We can use approaches, such as RNAsubopt (Wuchty et al., 1999), to enumerate all possible suboptimal structures, and then use approaches, such as BARRIERS (Flamm et al., 2002), to construct the exact energy landscape. However, the conformational space of feasible suboptimal structures can be extremely large, rendering a lot of redundant information (many suboptimal structures are similar to one another). For example, for the adenine riboswitch, the number of suboptimal structures with free energies between the 'on' and 'off' state structural conformations exceeds 10^9.

Researchers have also developed approaches that only investigate a subset of suboptimal structures. Zuker (1989) has developed mfold, an approach that is able to generate, for each admissible base pair in an RNA, the minimum energy structure containing the base pair. The approaches of Pipas and McMahon (1975) and Nakaya et al. (1996) consider structures composed of co-existing stacks to reduce the number of candidates. Evers and Giegerich (2001) have implemented an approach for enumerating all saturated suboptimal structures. Giegerich et al. (2004) have also developed RNAShapes, which can cluster suboptimal structures according to their shapes. Lorenz and Clote (2011) have developed RNALocopt, which can sample a user-defined number of locally optimal structures. Also, Lou and Clote (2011) have contributed RNAborMEA, which can compute the structure with maximum expected accuracy over all *k*-neighbours for an RNA secondary structure *S* and a number *k*.

In our previous work (Li and Zhang, 2011), we have proposed a novel approach, RNASLOpt, for predicting functional structural conformations of a single RNA by finding stable local optimal (SLOpt) structures on the RNA's energy landscape. Usually, ncRNAs' functional structural conformations have some distinctive features. First, the functional structures are energetically favourable and optimal on their local energy landscapes, which we call local optimal (LOpt). They tend to reside at the bottom of energy basins to ensure being favoured over an ensemble of other structural conformations (Russell et al., 2002). This is because non-local optimal structures can progressively fold into their neighbouring structures with lower free energies easily, like rolling down a hill until reaching an energy basin (a LOpt structure). Second, the conformational transitions between any pair of alternate functional structures may involve high energy barriers, such that the ncRNA can become kinetically trapped on the energy landscape (i.e. if the energy barrier between two structures is low, then conformational transition between the two structures may occur easily).

Therefore, in order to predict ncRNAs' native structures, we have proposed to exploit ncRNAs' underlying energy landscapes and search for SLOpt structures, that are not only thermodynamically stable, but also involve high energy barriers during the folding pathways to any other SLOpt structures. That is, given an ncRNA sequence, how to enumerate all the SLOpt structures such that (a) their free energies are within a certain energy range ΔE from the Minimum Free Energy (MFE), (b) they are local optimal on the ncRNA's energy landscape and (c) they are kinetically stable such that the minimal energy barrier between any two SLOpt structures is no less than a certain threshold $\Delta \mathcal{B}$?

We have employed stack configurations (each of which contains a set of compatible stacks) to represent scaffolds of RNA secondary structures. And, we have used LOpt stack configurations to approximate LOpt structures, where each LOpt stack configuration consists of a maximal number of compatible stacks (i.e. no additional stack can be added without forming pseudoknots). We have enumerated all the LOpt stack configurations within an energy range ΔE from the MFE, and then used a fast heuristic to compute the approximated pairwise energy barriers among these LOpt stack configurations, and finally applied a clustering algorithm to obtain all the SLOpt stack configurations (among which all the pairwise energy barriers are greater than or equal to $\Delta \mathcal{B}$). Based on the generated SLOpt stack configurations, we can infer a compact representation of the RNA's energy landscape with a remarkably reduced conformational space. Moreover, from the reduced search space, we can distinguish the ncRNA's alternate native structural conformations more accurately.

1.2 Predicting the optimal consensus structure for a family of related RNAs

The biological functions of ncRNAs are usually determined by their structures. And, ncRNAs that carry out similar biological functions are likely to share similar structural conformations. Predicting secondary structures for a single RNA based on energy minimisation alone typically has limited accuracy. More accurate prediction can be obtained by using comparative approaches to compute consensus structures that are conserved among related ncRNAs. Comparative approaches for predicting consensus structures can either (a) conduct sequence alignment and thermodynamic-based folding simultaneously (e.g. the Sankoff algorithm (Sankoff, 1985), Foldalign (Gorodkin et al., 1997), Dynalign (Mathews and Turner, 2002)), or (b) rely on well-aligned sequence alignments and fold consensus structures (e.g. RNAalifold (Hofacker, 2007; Hofacker et al., 2002), Pfold (Knudsen and Hein, 2003), PETfold (Seemann et al., 2008), McCaskill-MEA (Kiryu et al., 2007), CentroidAlifold (Hamada et al., 2011)), or (c) first fold each individual RNA separately and then align all the predicted structures to obtain the consensus structure (e.g. RNACast (Reeder and Giegerich, 2005), RADAR (Khaladkar et al., 2007)). One of the most popular comparative approaches is RNAalifold, which takes into account thermodynamic stability, covariant mutations and inconsistent base pairing into consensus folding.

1.3 Consensus stable local optimal structures and energy landscapes for a family of related RNAs

Most of the comparative approaches can predict only the optimal consensus structure, while ignoring consensus suboptimal structures. These approaches are not appropriate for analysing ncRNAs with alternate functional structures. In order to predict ncRNAs' alternate functional structures more accurately and confidently, we want to study the consensus suboptimal structures that are conserved in evolution among related ncRNAs on their consensus energy landscapes. We assume that the consensus functional structures of ncRNAs should also be local optimal, residing at energy basins of the consensus energy landscape. In addition, the consensus folding pathways between any two consensus functional structures should involve high energy barriers such that the conformational transitions cannot occur easily.

We propose the following problem: given a family of related ncRNAs, how to enumerate all the consensus stable local optimal structures such that (a) they are conserved among the family of related ncRNAs, (b) their consensus free energies are within a certain energy range ΔE from the MFE, (c) they are local optimal on the consensus energy landscape, and (d) they are dynamically stable such that the pairwise energy barrier between any two of them is no less than $\Delta \mathcal{B}$? So far, to our knowledge, *no* specific method has been proposed to address this problem. In this paper, we describe our comparative approach, RNAConSLOpt, for finding consensus SLOpt (denoted by ConSLOpt) structures on the consensus energy landscape of a family of related ncRNAs.

1.4 Novel riboswitch elements discovery

An application of our proposed approach, RNAConSLOpt, is to search for novel riboswitch elements. Computational detection of novel riboswitches is a very challenging task. RNAConSLOpt is particularly suitable for addressing this problem, because riboswitches can switch between allosteric structure conformations that are mutually exclusive, while RNAConSLOpt can find evolutionarily conserved and thermodynamically stable structures.

Many researchers have developed a variety of methods for identifying new riboswitch elements in bacterial genomes. Barrick et al. (2004) have proposed an approach that integrates intergenic sequence search, pairwise sequence alignment, and structure-based motif search for novel riboswitches detection. They have discovered and experimentally verified several novel riboswitches in *B. subtilis* genome. Bengert and Dandekar (2004) have developed RiboswitchFinder, a method that searches an input sequence for specific riboswitch elements according to the sequence and structure patterns of the elements, and the energy-based folding of the input sequence. Abreu-Goodger and Merino (2005) have created RibEx, a web server that can search for known riboswitches and conserved regulatory elements in bacteria. In addition, Yao et al. (2006) have contributed CMfinder, an effective motif search tool that performs well in finding motifs that are present in a subset of unaligned sequences. CMfinder integrates energy-based secondary structure prediction and covariance models for characterising motifs. CMfinder can be applied to genome-wide homolog search and is shown to have identified many homologous instances of known ncRNA families. Moreover, Chang et al. (2009) have implemented RiboSW, a systematic method that searches putative riboswitch elements through considering secondary structures of known riboswitches, as well as sequence conservations of their functional regions. However, although these approaches perform well in identifying homologous instances of known riboswitch families, they cannot be used for *de novo* detecting novel riboswitches. In this paper, we propose to make use of RNAConSLOpt to develop a pipeline for *de novo* detecting riboswitch elements in bacteria 5′ Untranslated Regions (UTRs).

We arrange this paper as follows. In Section 2, we elucidate algorithms of RNAConSLOpt in detail. In Section 3, we show benchmarking tests of RNAConSLOpt on known riboswitches, and compare RNAConSLOpt against RNASLOpt. In addition, we present the pipeline utilising RNAConSLOpt to discover novel riboswitch elements within *Bacillus* bacteria genomes. In Section 4, we discuss further applications of RNAConSLOpt and finally conclude this paper in Section 4.

2 Methods

RNAConSLOpt incorporates not only free energies of structures, but also covariance and conservation signals into enumerating ConSLOpt structures. RNAConSLOpt consists of three algorithms: (a) the stack-based consensus folding algorithm, (b) the algorithm for generating all possible ConSLOpt stack configurations and (c) and the algorithm for filtering out unstable consensus LOpt stack configurations and obtaining ConSLOpt stack configurations. In the following, we first review the covariance and conservation score of aligned RNA sequences used in RNAalifold, and then define notations related to consensus stack configurations, and finally describe the three algorithms.

2.1 Covariant mutations and structural conservation

We represent an alignment of n ($n > 1$) related RNAs, each containing exactly L bases, by $\mathbb{A} = \{a_1, \dots, a_n\}$. By a_k^i, we denote the i-th base of the k-th RNA. The alphabet includes nucleotides $\{A, U, G, C\}$ and a gap '–'. Complementary nucleotides (including $\{A{\cdot}U, G{\cdot}C$ and $G{\cdot}U\}$ can form base pairs. Following the idea of RNAalifold (Hofacker et al., 2002), we consider the i-th and j-th columns of $\mathbb{A}$ to be complementary, if the covariance and conservation score between the two columns, γ_{ij}, is not less than a threshold value γ^* (with a default value −0.4). Recall that γ_{ij} is composed of a covariance score C_{ij} and an inconsistent score q_{ij}. Note that C_{ij} is the bonus to compensatory mutations that maintain the pairing pattern between i-th and j-th columns; while q_{ij} is the penalty to RNAs, of which the i-th and j-th columns cannot pair. The values of γ_{ij}, C_{ij} and q_{ij} are, respectively, computed using equations (1)–(3),

$$\gamma_{ij} = \frac{1}{n}\left(C_{ij} - \phi_1 q_{ij}\right) \quad (1)$$

where φ_1 is the relative weight of the inconsistent score and its default value is 1.0;

$$C_{ij} = \frac{2}{n-1} \sum_{1 \le k < l \le n} \begin{cases} d\left(a_k^i, a_l^i\right) + d\left(a_k^j, a_l^j\right) & if\left(a_k^i \cdot a_k^j\right) \wedge \left(a_l^i \cdot a_l^j\right) \\ 0 & otherwise \end{cases} \quad (2)$$

where $d(x, y)$ is the hamming distance between two nucleotides x and y (0, if $x = y$ and 1, if $x \neq y$);

$$q_{ij} = \sum_{1 \le k \le n} \begin{cases} 0 & if\ a_k^i \cdot a_k^j \\ 0.25 & if\ both\ a_k^i\ and\ a_k^j\ are\ gaps \\ 1 & otherwise \end{cases} \quad (3)$$

2.2 Notations of consensus stacks and structures

By computing γ_{ij} for all possible i and j, where $1 \le i < j \le L$, we can determine the consensus base-pairing pattern in $\mathbb{A}$. Following the convention of RNASLOpt (Li and Zhang, 2011), we define the following notations. Let (i, j) represent a consensus base pair between the i-th and j-th columns of $\mathbb{A}$. A consensus stack of $\mathbb{A}$ is a helical region consisting of a set of *consecutive* consensus base pairs, which cannot extend on both ends. We use $p = (p_b, p_e, p_l)$ to represent a consensus stack containing the following p_l consecutive consensus base pairs, $\{(p_b, p_e), (p_b + 1, p_e - 1), \ldots, (p_b + p_l - 1, p_e - p_l + 1)\}$. p_b and p_e are the 5′ and 3′ ends of the outmost base pair in p. $|p|$ is the sequence length covered by the stack p and is equal to $p_e - p_b + 1$. We use $\gamma(p)$ to denote the covariance and conservation score of p. $\gamma(p)$ can be computed by adding up the γ scores of all the consensus base pairs in p.

We use $\mathcal{P}(\mathbb{A})$ to denote a set of all possible consensus stacks of $\mathbb{A}$, which contains at least a user-defined number of base pairs (the default value is 4). For any two stacks p and q in $\mathcal{P}(\mathbb{A})$, if p is parallel to the 5′ of q (i.e. $p_e < q_b$), then $p <_P q$; if p is enclosed by q (i.e. $q_b + q_l \le p_b$ and $p_e \le q_e - q_l$), then $p <_I q$; otherwise, p and q are incompatible. (The partial orders $p <_P q$ and $p <_I q$ can be loosely defined, allowing p and q to overlap by a few columns.) In case that p is enclosed by q, we use a stack $l_{p,q} = (q_b + q_l, p_b - 1, 0\}$ or $(r_{p,q} = (p_e + 1, q_e - q_l, 0))$ to represent the region that is enclosed by q and appears to the 5′ (or 3′) end of p. We define $\mathcal{P}(p)$ to be the set of all possible consensus stacks within p, and $\mathcal{F}_l(p)$ to be a subset of $\mathcal{P}(p)$. A stack $q \in \mathcal{P}(p)$ belongs to $\mathcal{F}_l(p)$, if and only if there is no stack q' in $\mathcal{P}(p)$, such that either $q <_P q'$ (i.e. q' appears to the 3′ of q), or $q <_I q'$ (i.e. q is embedded in q').

We use configurations of consensus stacks (containing a set of compatible consensus stacks allowing no pseudoknots) to represent scaffolds of consensus structures. We also employ consensus LOpt stack configurations (each of which contains a maximal number of compatible consensus stacks) to approximate consensus LOpt structures. We use consensus free energy for evaluating each generated consensus structure. The consensus free energy contains both the covariance and conservation score, and the average free energy over all single RNAs in the alignment, and is computed in a similar manner to RNAalifold.

We define the following terminal symbols. By $\underline{S}(p)$, we denote the normalised stabilising consensus energy of all the stacking base pairs in a consensus stack p. $\underline{H}(p)$ is the normalised destabilising consensus energy of hairpin loops enclosed by p, and $\underline{I}(p, q)$ is the normalised consensus energy of an interior loop or a bulge between stacks p and q. In case that an RNA in the alignment cannot form a base pair (or a loop or a bulge) which exists in the consensus structure, the energy contribution of the particular base pair in the RNA will not be counted. $\underline{M_c}$ is a constant offset penalty for closing a multi-loop. $\underline{M_b}$ and $\underline{M_i}$ are constant penalties for each unpaired base and each helix in a multi-loop. We also define non-terminal symbols: $F(p)$, $C(p)$, $FM1(p)$ and $FM(p)$, each represents the minimum consensus energy over all stack configurations within p conforming to the following constraints:

- $F(p)$: $p_b = 1$ and $p_l = 0$;
- $C(p)$: $p_l \ne 0$ and p closes some structures within itself;

- *FM*1(*p*): *p* is within a multi-loop, and there exists at least a consensus stack *q* such that $q_l \neq 0$ and then $q <_I p$;
- *FM(p): p* is within a multi-loop.

2.3 Stack-based consensus folding algorithm

In the work of RNASLOpt (Li and Zhang, 2011), we have described a recursive formula for computing the MFE for all possible LOpt stack configurations of a single RNA. Here, we modify the formula in order to compute the minimum consensus energy for aligned sequences of related ncRNAs, as in equation (4):

$$\begin{aligned}
F(p) &= \min_{q \in \mathscr{F}_I(p)} \{C(q) + F(l_{p,q})\} \\
C(p) &= \underline{S}(p) + \phi_2 \gamma(p) + \min \begin{cases} \underline{H}(p), \\ \min\limits_{q <_I p} \{C(q) + \underline{I}(p,q)\}, \\ \min\limits_{\substack{q \in \mathscr{F}_I(p) \\ \mathscr{F}_I(l_{p,q}) \neq \varnothing}} \left\{ \begin{array}{l} C(q) + FM1(l_{p,q}) + \underline{M_c} \\ +2 * \underline{M_i} + |r_{p,q}| * \underline{M_b} \end{array} \right\} \end{cases} \\
FM1(p) &= \min_{q \in \mathscr{F}_I(p)} \{C(q) + FM(l_{p,q}) + \underline{M_i} + |r_{p,q}| * \underline{M_b}\} \\
FM(p) &= \min \begin{cases} |p| * \underline{M_b}, \\ \min\limits_{q \in \mathscr{F}_I(p)} \left\{ \begin{array}{l} C(q) + FM(l_{p,q}) \\ +\underline{M_i} + |r_{p,q}| * \underline{M_b} \end{array} \right\} \end{cases}
\end{aligned} \tag{4}$$

where φ_2 is the weight of the covariance and conservation score and its default value is 0.5. The major differences are that (a) we consider the consensus structures shared among related ncRNAs, instead of structures of a single ncRNA and (b) we integrate the covariance and conservation score in evaluating the generated structures.

2.4 Generating all possible consensus local optimal stack configurations

Next, we enumerate all possible consensus LOpt stack configurations of $\mathbb{A}$ within an energy range of Δ*E* from the minimum consensus free energy. In the work of RNASLOpt (Li and Zhang, 2011), we have proposed an approach for enumerating all possible LOpt stack configurations for a single RNA. We modify it for aligned RNA sequences as follows.

We use *p** (where *p** = (1, *L*, 0)) to denote the stack that covers the overall alignment of $\mathbb{A}$. The minimum consensus free energy of $\mathbb{A}$ is *F*(*p**), and the energy upper bound is Δ*E* + $F(p^*)$. We use a partial stack configuration ϕ_0 (where $\phi_0 = \{(p^*, F)\}$) to represent all possible consensus LOpt stack configurations on $\mathbb{A}$. A partial stack configuration ϕ is composed of a set of compatible consensus stacks, where each consensus stack *p* is associated with one of the five labels: *finished*, *F*, *C*, *FM*1 and *FM*. For each consensus stack *p* in ϕ, we decompose the region covered by *p* into several separated sub-regions according to the label of *p*, and then construct a set of new partial stack configurations accordingly. The decomposition and construction are conducted through back tracking the recursive formula of equation (4) (similar to the procedures described in Li and Zhang, 2011). We repeatedly process each partial stack configuration ϕ, until either the consensus free energy of ϕ is greater than the energy upper bound, or all the consensus stacks in ϕ are labelled as *finished.*

During the back tracking phase, at each step, we determine whether to include a consensus stack. This procedure differs from those of RNASLOpt and RNAsubopt in that: at each step, RNASLOpt decides whether to include a stack of a single RNA; and RNAsubopt chooses whether to form a feasible base pair. RNASLOpt can greatly reduce the search space compared with RNAsubopt, because it encounters far fewer branching points (as the number of stacks is less than the number of feasible base pairs) (Li and Zhang, 2011). Similarly, RNAConSLOpt is expected to explore a further reduced, yet evolutionarily conserved, conformational space of consensus structures compared with RNASLOpt (as the number of consensus stacks of aligned RNAs is usually less than the number of stacks in a single RNA). Note that, although RNAConSLOpt still considers a search space that grows exponentially with sequence length, it can further reduce the number of candidate structures, and thus can be applied to longer sequences with a greater energy range.

2.5 Clustering consensus stable local optimal stack configurations

Finally, we select consensus stable local optimal structures from the consensus LOpt stack configurations based on pairwise consensus energy barriers. To achieve this goal, we need to compute the pairwise consensus energy barriers among LOpt structures. The problem of determining the minimal energy barrier between two secondary structures, even for a single RNA, is hard (Manuch et al., 2009). Although both exact solutions (Flamm et al., 2002; Thachuk et al., 2010) and heuristic approaches (Morgan and Higgs, 1998; Dotu et al., 2010; Flamm et al., 2001; Geis et al., 2008; Morgan and Higgs, 1998; Voss et al., 2004) have been proposed to address this problem for single RNAs, they are not tailored for computing consensus energy barriers for aligned RNAs and are not fast enough to apply to thousands of pairs of conformational structures. Therefore, we use the fast heuristic described in our previous work (Li and Zhang, 2011) to compute consensus energy barriers. Finally, we obtain a set of ConSLOpt structures (among which all the pairwise consensus energy barriers are greater than or equal to $\Delta \mathscr{B}$) using neighbour joining clustering (Li and Zhang, 2011).

3 Results and discussion

3.1 Benchmarking tests on known riboswitches

In order to test whether RNAConSLOpt is able to predict alternate functional structures for riboswitches, we conducted benchmarking tests on the adenine riboswitch, the thiamine pyrophosphate (TPP) riboswitch, the lysine riboswitch and the flavin mononucleotide (FMN) riboswitch. First, we obtained primary sequences and native structural conformations of the following riboswitches as the reference: adenine – *ydhL* gene of *B. subtilis* (Mandal and Breaker, 2004), TPP – *thiamine* of *B. subtilis* (Mironov et al., 2002; Rentmeister et al., 2007), lysine – *lysC* of *B. subtilis* (Blouin et al., 2011) and FMN – *ribD* of *B. subtilis* (Winkler et al., 2002). Next, for each riboswitch, we constructed an alignment of homologous sequences. We downloaded the seed alignment of each riboswitch from the Rfam database (Griffiths-Jones et al., 2003). Note that we could not use the seed alignment directly, because it is an alignment of partial sequences that are too short when compared to the full reference sequence. For each partial sequence in the seed alignment, we inferred the genomic location of the full sequence accordingly. After extracting all the full sequences from the EMBL Nucleotide Sequence Database (Kanz et al., 2005), we selected the reference sequence and four other sequences which have lower than 90% sequence identity with the reference, and aligned them using ClustalW2 (Larkin et al., 2007). We applied RNAConSLOpt to the constructed riboswitch alignments in order to produce ConSLOpt stack configurations. Finally, we evaluated the generated ConSLOpt structures using the reference native structural conformations and compared RNAConSLOpt against RNASLOpt.

The native and predicted ‘on’ and ‘off’ structural conformations of the adenine riboswitch are shown in Figure 1. We found that covariant mutations exist in both ‘on’ and ‘off’ structures and are informative for the prediction. In Table 1, we also compared ranks of the best predicted structures corresponding to the native ‘on’ and ‘off’ structures produced by RNAConSLOpt against the ranks by RNASLOpt. We can see that ranks of ‘on’ and ‘off’ structures predicted by RNAConSLOpt are better than those of RNASLOpt. This is due to the power of comparative analysis in ncRNA structure prediction. RNAConSLOpt only investigates consensus stable local optimal structures residing at energy basins of the consensus energy landscape. It can further reduce the search space compared with RNASLOpt, retaining the ability to predict both alternate native structures for riboswitches. The running time for the four benchmarking tests (on a 32 bit, 2.4 GHz Quad-processor, 3.2 GB memory PC) are 1 second, 3 seconds, 8 seconds, and 14 seconds, respectively. This indicates that RNAConSLOpt can be applied to alignments of length around 250 with great efficiency.

In addition, we also compared the number of ConSLOpt structures of aligned riboswitches (produced by RNAConSLOpt) against the number of SLOpt structures of the reference sequence (produced by RNASLOpt). In general, the number of ConSLOpt structures of aligned riboswitches is a small fraction of the number of SLOpt structures of the reference sequence, as shown in Figure 2. The source code and benchmarking tests for RNAConSLOpt (version 1.1) are available at http://genome.ucf.edu/RNAConSLOpt

3.2 A pipeline for de novo detection of riboswitch elements in bacterial genomes

We present a pipeline that utilises RNAConSLOpt in detecting novel riboswitch elements. RNAConSLOpt can predict consensus stable local optimal structures for aligned orthologous sequences, while putative riboswitches are likely to have allosteric structure conformations. Therefore, by analysing covariant mutation patterns of the predicted ConSLOpt structures, we can obtain additional information and then discover putative riboswitch elements with more confidence. We have applied this riboswitch detection pipeline to a set of bacteria in *Bacillus* genus, and carried out the following procedures.

First, we downloaded 82 complete genomes of 37 *Bacillus* bacteria, as well as their gene annotations from the National Institutes of Health (NIH) National Center for Biotechnology Information (NCBI) ftp server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). We selected *B. subtilis 168* (with GenBank accession number NC_000964) as the reference genome. *B.*

subtilis is a well studied and annotated organism commonly used as a model in bacteria research. *B. subtilis* has 4155 non-redundant genes annotated. For each gene, we collected the upstream sequences of all orthologous genes from the 82 *Bacillus* bacteria genomes, aiming at constructing an orthologous sequence alignment. Each sequence consists of up to 500 nucleotides in 5′-UTR of the specific gene and the starting 50 nucleotides of the gene’s protein coding region. We kept the starting 50 nucleotides of the protein coding region so that we can use them as an anchor to construct high-quality alignments. We also discarded short orthologous sequences which have less than 100 nucleotides in 5′-UTR. After collecting all the orthologous sequences for a specific gene, we then employed ClustalW2 (Larkin et al., 2007) to construct an alignment.

With the constructed orthologous sequence alignments, we then divided them into many small overlapping windows. The window size can be 100, 120, 140 and 160 and the step size is 20. We refined each alignment window using rnazSelectSeqs.pl in RNAz (Gruber et al., 2010) package (version 2.1 with default parameters). Note that the refined alignments produced by RNAz are usually shorter in length than the original alignments. We only chose windows with lengths between 90 and 120. We also filtered out windows which contain less than four sequences, as they cannot provide enough covariant mutation information. Further, for each remaining alignment window, we used RNAz (with – no-shuffle option) to predict whether the alignment is likely to be a real RNA. We removed windows which have less than 50% probability of being classified as an RNA by RNAz, and finally obtained 10,577 high-quality alignment windows.

After selecting 10,577 alignment windows, we applied RNAConSLOpt to each of them with the default parameters (ΔE = 15 kcal/mol, $\Delta\ \mathcal{B}$ =12 kcal/mol). RNAConSLOpt produced ConSLOpt structures for each window and ranked these structures by their associated minimal energy barriers. We denoted the rank 1st and rank 2nd ConSLOpt structures by R_1 and R_2, respectively. $E(R_1)$ and $E(R_2)$ represent consensus energies with covariant scores for R_1 and R_2, respectively. Among all the selected windows, 4037 of them were predicted with putative allosteric consensus structures.

Since many of the remaining 4037 windows may overlap with one another, for each group of overlapping windows, we selected the one with the lowest $E(R_2)$ as the representative. After trimming redundant information from the results, we obtained 630 non-overlapping windows. To make the prediction more conservative, we only analysed 506 windows for which the average distance to the starting codons of their downstream genes is less than 100. With $E(R_2)$ less than −10 (kcal/mol) and −20 (kcal/mol), we obtained 161 and 38 putative riboswitch candidates, respectively.

In order to check whether the putative riboswitches have already been studied or not, we searched their orthologous sequences in the alignments against known riboswitch families. First, we used BLAST (Altschul et al., 1990) (with option megablast) to compare each orthologous sequence against the full sequence alignments of RNA families in the Rfam database (Griffiths-Jones et al., 2003). We considered a riboswitch candidate belonging to a known RNA family if one of its orthologous sequences ‘hit’ an Rfam RNA family with an e-value less than 10^{-5}. The Rfam RNA family would be denoted as the best matching RNA family for the putative riboswitch. In addition, we also conducted homolog search against covariance models of known ncRNAs in Rfam using Infernal's cmsearch (Nawrocki et al., 2009) with a significant e-value cut-off ($E < 10^{-10}$).

Finally, we sorted all the windows based on their $E(R_2)$ values (i.e. the consensus energy with covariance for the rank 2nd ConSLOpt structure R_2). Table 2 shows all the predictions with $E(R_2)$ less than −20 (kcal/mol). Supplementary information of the list of genomes used in this pipeline, the riboswitch candidates with $E(R2)$ value less than −10 (kcal/mol), and the predicted ConSLOpt structures for all the riboswitch candidates is available at http://genome.ucf.edu/RNAConSLOpt.

3.3 Discovery of novel riboswitch elements in Bacillus bacteria genomes

Genome-wide discovery of riboswitch elements in *Bacillus* bacteria genomes using the pipeline results in 38 hits with $E(R_2)$ less than −20 (kcal/mol). These 38 potential riboswitch elements are sorted based on $E(R_2)$ and are listed in Table 2. Among the 38 genes whose 5′-UTR contain potential riboswitch elements, 28 of them are recognised by the KEGG pathway analysis (Kanehisa et al., 2004). Of these recognised genes, 60.7% (17/28) of them are involved in metabolic pathways. The major pathways consist of aminoacyl-tRNA biosynthesis, biosynthesis of secondary metabolites, microbial metabolism in diverse environments, thiamine metabolism, pyrimidine metabolism, purine metabolism, methane metabolism and histidine metabolism.

BLAST (Altschul et al., 1990) search of the 38 regions against Rfam database reveals that 34.2% (13/38) of them are annotated riboswitches or mRNA leader elements (see Table 2). In addition, we further use Infernal's cmsearch to annotate the other 25 regions that are not registered in Rfam. The cmsearch results indicate another 7 potential riboswitch elements with significant e-values. An example of this category resides in the 5′-UTR of *cysE*, which codes serine acetyltransferase. This enzyme, together with acetyl-coA, catalyses the reaction of producing O-acetylserine from serine. O-acetylserine participates in the sulphur metabolic pathway, which synthesises organic sulphur metabolites such as cysteine, methionine and S-adenosyl-methionine (Andre et al., 2008). Although experimental evidence suggests that many steps of this pathway are regulated by T-box and S-box riboswitches, whether *cysE* is also regulated by riboswitch is still unclear (Andre et al., 2008). The discovery of an allosteric structure of this element, and its sequence and structural resemblance to T-box riboswitch, supports the hypothesis that these genes are regulated by T-box riboswitch.

The other 18 genes whose 5′-UTR do not contain known riboswitch elements are likely to be regulated by novel riboswitch elements. We selected two elements as examples for detailed discussion. The first gene *greA* codes for the transcription elongation factor GreA. It has been recently experimentally verified that this gene is regulated by the *greA* attenuator (Potrykus et al., 2010) in *E. coli*. The presence of such an attenuator indicates that this gene is under certain transcriptional regulation by its 5′-UTR. However, the mechanism of this regulation is still unclear (Naville and Gautheret, 2010). Our results indicate that the attenuator may act like a riboswitch, which regulates the transcription of the gene by alternating its structure. Interestingly, homolog search (using cmsearch) of the *greA* attenuator profile against *B. subtilis* does not return any significant hits. This implies that the *greA* attenuator adopts its own structures in *B. subtilis*, which in turn suggests that the gene may participate in different biological pathways and under the different regulation in *B. subtilis*. The predicted allosteric structures R_1 and R_2 of *greA* are shown in Figure 3.

The second gene *nadD* codes nicotinate mononucleotide adenylyl transferase (NMNAT), which catalyses the adenylation of nicotinate mononucleotide to nicotinate adenine dinucleotide (NAD). The biochemical function of the enzyme NMNAT resembles that of FMN adenylyl transferase (FMNAT), which also catalyses adenylation as an enzyme, but produces flavin adenine dinucleotide (FAD) from flavin mononucleotide (FMN). The interaction catalysed by FMNAT is a critical step of FMN biosynthesis pathway, and the expression of FMNAT is considered to be regulated by the FMN riboswitch (Nudler and Mironov, 2004; Gusarov et al., 1997; Mack et al., 1998). As a result, it is highly possible that the enzyme NMNAT, which is coded by *nadD* gene, is also regulated by riboswitch elements in the 5′-UTR. Using RNAConSLOpt, we are able to identify a potential allosteric RNA element in the 5′-UTR (see Figure 4), which further implies the existence of such a riboswitch element. Homolog search with cmsearch against this region does not result in any significant matches with known riboswitch families, suggesting that the riboswitch element that regulates *nadD* is novel. The sequences of this region are relatively diverse (79.8% mean pairwise identity), yet most of the mutations are covariant. More importantly, we identified a covariant mutation that is compatible for both structures that the riboswitch-like element can adopt. Therefore, *nadD* is likely to be regulated by a riboswitch-like element, and its predicted allosteric structures R_1 and R_2 are shown in Figure 4.

4 Conclusion

We have proposed the first comparative approach, RNAConSLOpt, for producing all possible ConSLOpt (i.e. consensus stable local optimal) stack configurations given an alignment of related ncRNAs. Based on these ConSLOpt structures, we can distinguish alternate functional structures for ncRNA families more accurately and confidently. Moreover, we can construct a compact representation of the consensus energy landscape of an ncRNA family. The benchmarking tests on four riboswitch families show that RNAConSLOpt outperforms RNASLOpt in reducing the number of candidate structures and improving the ranks of both predicted alternate functional structures.

In addition, we have built a pipeline making use of RNAConSLOpt to discover novel riboswitch elements genome wide. The advantage of this pipeline is that it requires no preliminary knowledge about sequences and structures of known riboswitches. Therefore, it can be used not only for identifying homologous instances of known riboswitches, but also for *de novo* riboswitch detection. An application of this pipeline to a set of bacteria in *Bacillus* genus results in the recovering of many known riboswitches and the detection of many novel riboswitch candidates. The KEGG pathway analysis and biological function annotation of proteins associated with several riboswitch candidates, together with studies of their putative allosteric structures, provide strong evidence that they are likely to be real riboswitches. Our future work involves applying the proposed pipeline to systematically detect riboswitch elements in more bacterial genomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM102515).

References

Abreu-Goodger C, Merino E. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Research. 2005; 33:W690–W692. [PubMed: 15980564]

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215(3):403–410. [PubMed: 2231712]

Andre G, Even S, Putzer H, Burguiere P, Croux C, Danchin A, Martin-Verstraete I, Soutourina O. S-box and T-box riboswitches and antisense RNA control a sulfur metabolic operon of Clostridium acetobutylicum. Nucleic Acids Research. 2008; 36(18):5955–5969. [PubMed: 18812398]

Bengert P, Dandekar T. Riboswitch finder – a tool for identification of riboswitch RNAs. Nucleic Acids Research. 2004; 32:W154–159. [PubMed: 15215370]

Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser JK, Breaker RR. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(17):6421–6426. [PubMed: 15096624]

Blouin S, Chinnappan R, Lafontaine DA. Folding of the lysine riboswitch: importance of peripheral elements for transcriptional regulation. Nucleic Acids Research. 2011; 39:3373–3387. [PubMed: 21169337]

Chang TH, Huang HD, Wu LC, Yeh CT, Liu BJ, Horng JT. Computational identification of riboswitches based on RNA conserved functional sequences and conformations. RNA. 2009; 15(7): 1426–1430. [PubMed: 19460868]

Dotu I, Lorenz WA, Van Hentenryck P, Clote P. Computing folding pathways between RNA secondary structures. Nucleic Acids Research. 2010; 38:1711–1722. [PubMed: 20044352]

Evers, D.; Giegerich, R. Reducing the conformation space in RNA structure prediction. Proceedings of the German Conference on Bioinformatics; 2001. p. 118-124.

Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M. Design of multistable RNA molecules. RNA. 2001; 7:254–265. [PubMed: 11233982]

Flamm C, Hofacker IL, Stadler PF, Wolfinger MT. Barrier trees of degenerate landscapes. Zeitschrift für Physikalische Chemie. 2002; 216:155–174.

Geis M, Flamm C, Wolfinger MT, Tanzer A, Hofacker IL, Middendorf M, Mandl C, Stadler PF, Thurner C. Folding kinetics of large RNAs. Journal of Molecular Biology. 2008; 379:160–173. [PubMed: 18440024]

Giegerich R, Voss B, Rehmsmeier M. Abstract shapes of RNA. Nucleic Acids Research. 2004; 32:4843–4851. [PubMed: 15371549]

Gorodkin J, Heyer LJ, Stormo GD. Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Research. 1997; 25:3724–3732. [PubMed: 9278497]

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Research. 2003; 31:439–441. [PubMed: 12520045]

Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. RNAZ 2.0: improved noncoding RNA detection. Pacific Symposium on Biocomputing. 2010; 15:69–79. [PubMed: 19908359]

Gusarov II, Kreneva RA, Rybak KV, Podcherniaev DA, Iomantas IV, Kolibaba LG, Polanuer BM, Kozlov II, Perumov DA. Primary structure and functional activity of the Bacillus subtilis ribC gene. Molecular Biology. 1997; 31(5):820–825.

Hamada M, Sato K, Asai K. Improving the accuracy of predicting secondary structure for aligned RNA sequences. Nucleic Acids Research. 2011; 39:393–402. [PubMed: 20843778]

Hofacker IL. RNA consensus structure prediction with RNAalifold. Methods in Molecular Biology. 2007; 395:527–544. [PubMed: 17993696]

Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA sequences. Journal of Molecular Biology. 2002; 319:1059–1066. [PubMed: 12079347]

Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Diez FG, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R. The EMBL nucleotide sequence database. Nucleic Acids Research. 2005; 33:29–33.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Research. 2004; 32:D277–280. [PubMed: 14681412]

Khaladkar M, Bellofatto V, Wang JT, Tian B, Shapiro BA. RADAR: a web server for RNA data analysis and research. Nucleic Acids Research. 2007; 35:W300–304. [PubMed: 17517784]

Kiryu H, Kin T, Asai K. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. Bioinformatics. 2007; 23:434–441. [PubMed: 17182698]

Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research. 2003; 31:3423–3428. [PubMed: 12824339]

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. ClustalW and ClustalX version 2. Bioinformatics. 2007; 23(21):2947–2948. [PubMed: 17846036]

Li Y, Zhang S. Finding stable local optimal RNA secondary structures. Bioinformatics. 2011; 27:2994–3001. [PubMed: 21903624]

Lorenz WA, Clote P. Computing the partition function for kinetically trapped RNA secondary structures. PLoS ONE. 2011; 6:e16178. [PubMed: 21297972]

Lou, F.; Clote, P. Maximum expected accurate structural neighbors of an RNA secondary structure. IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS’2011); 2011. p. 123-128.

Mack M, van Loon AP, Hohmann HP. Regulation of riboavin biosynthesis in Bacillus subtilis is affected by the activity of the avokinase/avin adenine dinucleotide synthetase encoded by ribC. Journal of Bacteriology. 1998; 180(4):950–955. [PubMed: 9473052]

Mandal M, Breaker RR. Adenine riboswitches and gene activation by disruption of a transcription terminator. Nature Structural & Molecular Biology. 2004; 11:29–35.

Manuch, J.; Thachuk, C.; Stacho, L.; Condon, A. In: Deaton, R.; Suyama, A., editors. NP-completeness of the direct energy barrier problem without pseudoknots; 15th International Conference DNA Computing and Molecular Programming; Berlin, Heidelberg: Springer-Verlag; 2009. p. 106-115.

Mathews DH, Turner DH. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. Journal of Molecular Biology. 2002; 317:191–203. [PubMed: 11902836]

Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA, Nudler E. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. Cell. 2002; 111:747–756. [PubMed: 12464185]

Morgan SR, Higgs PG. Barrier heights between ground states in a model of RNA secondary structure. Journal of Physics A: Mathematical and General. 1998; 31(14):3153–3170.

Nakaya A, Yonezawa A, Yamamoto K. Classification of RNA secondary structures using the techniques of cluster analysis. Journal of Theoretical Biology. 1996; 183:105–117. [PubMed: 8959113]

Naville M, Gautheret D. Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation. Genome Biology. 2010; 11(9):R97. [PubMed: 20920266]

Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009; 25(10):1335–1337. [PubMed: 19307242]

Nudler E, Mironov AS. The riboswitch control of bacterial metabolism. Trends in Biochemical Sciences. 2004; 29:11–17. [PubMed: 14729327]

Pipas JM, McMahon JE. Method for predicting RNA secondary structure. Proceedings of the National Academy of Sciences of the United States of America. 1975; 72:2017–2021. [PubMed: 1056009]

Potrykus K, Murphy H, Chen X, Epstein JA, Cashel M. Imprecise transcription termination within Escherichia coli greA leader gives rise to an array of short transcripts, GraL. Nucleic Acids Research. 2010; 38(5):1636–1651. [PubMed: 20008510]

Reeder J, Giegerich R. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. Bioinformatics. 2005; 21:3516–3523. [PubMed: 16020472]

Rentmeister A, Mayer G, Kuhn N, Famulok M. Conformational changes in the expression domain of the Escherichia coli thiM riboswitch. Nucleic Acids Research. 2007; 35:3713–3722. [PubMed: 17517779]

Russell R, Zhuang X, Babcock HP, Millett IS, Doniach S, Chu S, Herschlag D. Exploring the folding landscape of a structured RNA. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99:155–160. [PubMed: 11756689]

Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM Journal of Applied Mathematics. 1985; 455:810–825.

Schultes EA, Bartel DP. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. Science. 2000; 289:448–452. [PubMed: 10903205]

Seemann SE, Gorodkin J, Backofen R. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Research. 2008; 36:6355–6362. [PubMed: 18836192]

Thachuk C, Manuch J, Rafiey A, Mathieson LA, Stacho L, Condon A. An algorithm for the energy barrier problem without pseudoknots and temporary arcs. The Pacific Symposium on Biocomputing. 2010; 15:108–119.

Voss B, Meyer C, Giegerich R. Evaluating the predictability of conformational switching in RNA. Bioinformatics. 2004; 20:1573–1582. [PubMed: 14962925]

Winkler WC, Cohen-Chalamish S, Breaker RR. An mRNA structure that controls gene expression by binding FMN. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99:15908–15913. [PubMed: 12456892]

Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. 1999; 49:145–165. [PubMed: 10070264]

Yao Z, Weinberg Z, Ruzzo WL. CMfinder – a covariance model based RNA motif finding algorithm. Bioinformatics. 2006; 22(4):445–452. [PubMed: 16357030]

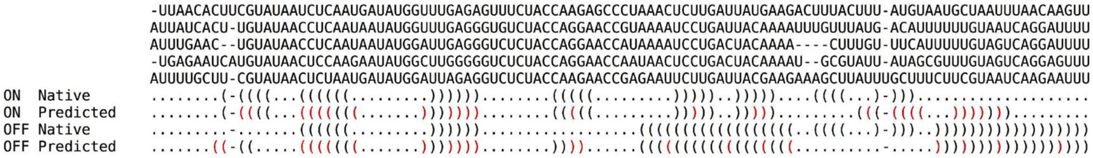
Zuker M. On finding all suboptimal foldings of an RNA molecule. Science. 1989; 244:48–52. [PubMed: 2468181]

Biographies

Yuan Li obtained her PhD in Computer Science from the University of Central Florida. She is currently a Senior Software Engineer of Algorithms in the R&D Department of Pacific Biosciences of California, Inc. Her research interests are in the area of fast alignment of high-throughput, error-prone, long reads in application of de novo genome assembly, non-coding RNA analysis and prediction of consensus isoforms for transcriptomic sequences.

Cuncong Zhong obtained his PhD and MS in Computer Science from the University of Central Florida. He also earned his BS in Computer Science and BS in Biotechnology from Huazhong University of Science and Technology, China. He is currently developing novel algorithms and software for metagenomic data annotation at the J. Craig Venter Institute as a post-doctoral fellow.

Shaojie Zhang received his PhD in Computer Science from the University of California, San Diego. He is currently an Associate Professor of Computer Science at the Department of Electrical Engineering and Computer Science at the University of Central Florida. His research is focused on bioinformatics, which includes ncRNA gene finding, RNA analysis, and computational genomics.



```
                 -UUAACACUUCGUAUAAUCUCAAUGAUAUGGUUUGAGAGUUUCUACCAAGAGCCCUAAACUCUUGAUUAUGAAGACUUUACUUU-AUGUAAUGCUAAUUUAACAAGUU
                 AUUAUCACU-UGUAUAACCUCAAUAAUAUGGUUUGAGGGUGUCUACCAGGAACCGUAAAAUCCUGAUUACAAAAUUUGUUUAUG-ACAUUUUUUGUAAUCAGGAUUUU
                 AUUUGAAC--UGUAUAACCUCAAUAAUAUGGAUUGAGGGUCUCUACCAGGAACCAUAAAAUCCUGACUACAAAA----CUUUGU-UUCAUUUUUGUAGUCAGGAUUUU
                 -UGAGAAUCAUGUAUAACUCCAAGAAUAUGGCUUGGGGGUCUCUACCAGGAACCAAUAACUCCUGACUACAAAAU--GCGUAUU-AUAGCGUUUGUAGUCAGGAGUUU
                 AUUUUGCUU-CGUAUAACUCUAAUGAUAUGGAUUAGAGGUCUCUACCAAGAACCGAGAAUUCUUGAUUACGAAGAAAGCUUAUUUGCUUUCUUCGUAAUCAAGAAUUU
ON  Native       ........(-((((...((((((.........))))))........(((((.........)))))..)))))....((((...)-)))....................
ON  Predicted    ........(-((((...(((((((.......)))))))........(((((.........)))))..)))))..........(((-((((...)))))))........
OFF Native       .........-.......((((((.........))))))..................((((((((((((((((((..((((...)-)))..))))))))))))))))))
OFF Predicted    .......((-((.....(((((((.......)))))))........))))......((((((((((((((((((..........-.....))))))))))))))))))
```

Figure 1.

Aligned sequences of the adenine riboswitches and the corresponding native and predicted consensus ‘on’ and ‘off’ conformational structures. Pairing columns with covariant mutations in the predicted consensus structures are coloured red

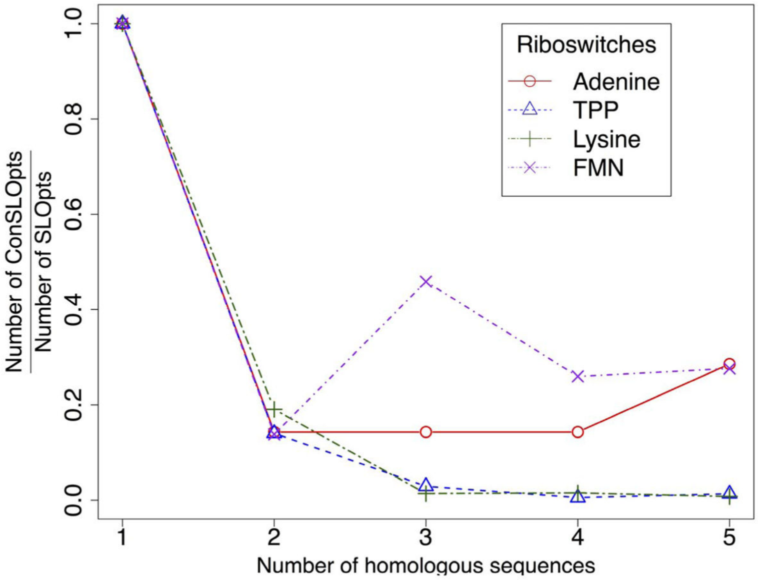


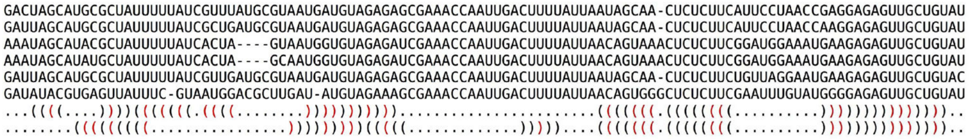
Figure 2.

Comparison of the number of ConSLOpt structures and that of SLOpt structures. ConSLOpts and SLOpts represent the consensus SLOpt stack configurations of aligned RNA sequences, and the SLOpt stack configuration of the reference RNA, respectively

```
GGGGUUGUAUGUGACAACUCCGCUAGUAC-AGGCGUGCUAGAAACCUCCGCUCUCUAUAAAGCGGAGGAGUUUCAUAUG-GAACUCCUCUUUUUUUCGGGGGAUUGGUAUAUAA
GGGGUUGUAUGUGACAACUCCACUAGUGCUACGUGUGCUAGAAACCUUCGCU----AUAAAGCGGAGGAGUUUCAUAUG-GAACUCCUCUUUUUUUCGGGGGAUUGGUAUAUAA
GGGGUUGUAUGUGACAACUCCGCUAGUAC-AGGCGUGCUAGAAACCUCCGCUUUACAUAAAGCGGGGGAGUUUCAUAUG-GAACUCCUCUUUUUUUCGGGGGAUUGGUAUAUAA
GGGGUUGUAUGUGACAACUCCGCUAGUGC-AAGGGUACUAGAAACCUCCGCUAACAAUGAAGCGGAGGAGUUUCAUAUG-GAACUCCUCUUUUUUUCAGGGGAUUGGUAUAUAA
GGGGUUGUAUGUGACAACUCCGCUAGUGC-AAGGGUACUAGAAACCUCCGCUAACAAUGAAGCGGAGGAGUUUCAUAUG-GAACUCCUCUUUUUU-CGGGGGAUUGGUAUAUAA
GGGGUUGUAUGUGACAACUCCGCUAGUGC-AUAUGUACUAGAAACCUCCGCUAU-UGGAAUGCGGAGGAGUUUCAUAUUUGAACUCCUCUUUUCU-CGGGGGAUUGGUAUAUAA
((((((.(((((((.((((((.(((((((-....))))))).....((((((.......))))))))))))).)))))))-.))))))((((((....))))))...........
((((((((.....))))))))).(((((((-....)))))))...((((((............(((((((((((.....)-))))))))))......))))))............
```

Figure 3.

The predicted rank 1st and 2nd ConSLOpt structures for a putative riboswitch element upstream of *greA.* An alignment of orthologous sequences located in 5′-UTR of *greA*, together with its rank 1st and 2nd ConSLOpt structures produced by RNAConSLOpt are shown. Pairing columns with covariant mutations are coloured red



```
GACUAGCAUGCGCUAUUUUUAUCGUUUAUGCGUAAUGAUGUAGAGAGCGAAACCAAUUGACUUUUAUUAAUAGCAA-CUCUCUUCAUUCCUAACCGAGGAGAGUUGCUGUAU
GAUUAGCAUGCGCUAUUUUUAUCGCUGAUGCGUAAUGAUGUAGAGAGCGAAACCAAUUGACUUUUAUUAAUAGCAA-CUCUCUUCAUUCCUAACCAAGGAGAGUUGCUGUAU
AAAUAGCAUACGCUAUUUUUAUCACUA----GUAAUGGUGUAGAGAUCGAAACCAAUUGACUUUUAUUAACAGUAAACUCUCUUCGGAUGGAAAUGAAGAGAGUUGCUGUAU
AAAUAGCAUAUGCUAUUUUUAUCACUA----GCAAUGGUGUAGAGAUCGAAACCAAUUGACUUUUAUUAACAGUAAACUCUCUUCGGAUGGAAAUGAAGAGAGUUGCUGUAU
GAUUAGCAUGCGCUAUUUUUAUCGUUGAUGCGUAAUGAUGUAGAGAGCGAAACCAAUUGACUUUUAUUAAUAGCAA-CUCUCUUCUGUUAGGAAUGAAGAGAGUUGCUGUAC
GAUAUACGUGAGUUAUUUC-GUAAUGGACGCUUGAU-AUGUAGAAAGCGAAACCAAUUGACUUUUAUUAACAGUGGGCUCUCUUCGAAUUUGUAUGGGGAGAGUUGCUGUAU
...((((....))))(((((((.((((........))))))))))).......................(((((((.((((((((..........)))))))))))))))..
........(((((((((................)))))))))(((((.............)))))....(((((((.((((((((..........)))))))))))))))..
```

Figure 4.

The predicted rank 1st and 2nd ConSLOpt structures for a riboswitch element upstream of *nadD*. An alignment of orthologous sequences located in 5′-UTR of *nadD*, together with its rank 1st and 2nd ConSLOpt structures produced by RNAConSLOpt are shown. Pairing columns with covariant mutations are coloured red

Table 1

Ranks of the best structures corresponding to the native ‘off’ and ‘on’ structures by RNASLOpt and RNAConSLOpt. RNAConSLOpt was run with the default parameters for all the riboswitches (minimum stack length: 4; ΔE:15 kcal/mol; and ΔB: 12 kcal/mol). For each (*a*, *b*) in the table, *a* and *b* denote ranks of the best consensus structures corresponding to the native ‘off’ and ‘on’ structures respectively. RankE is the rank of each predicted structure based on its free energy. RankB is the rank of each predicted structure based on its minimal associated energy barrier (Li and Zhang 2011). Len represents length of each alignment. PairId represents the mean pairwise identity of each alignment. For each riboswitch, the best pair of ranks produced by RNASLOpt and RNAConSLOpt are in bold face

Name	**ΔE (kcal/mol)**	**RNASLOpt**			**RNAConSLOpt**				
		Rank E	**RankB**	**# of SLOpt**	**Len**	**PairId**	**RankE**	**RankB**	**# of ConSLOpt**
Adenine	25	(1, 5)	(1, 3)	6	108	0.67	**(1, 2)**	**(1, 2)**	2
TPP	15	(1, 5)	(1, 4)	369	194	0.62	(1, 5)	**(1, 3)**	5
Lysine	15	(25, 32)	(76, 33)	673	237	0.62	**(1, 2)**	**(1, 2)**	5
FMN	15	(64, 49)	(7, 27)	234	247	0.60	(1, 23)	**(1, 20)**	50

Table 2

Predicted riboswitch elements. Thirty eight predicted riboswitch elements in *Bacillus* genus with $E(R_2)$ less than −20 (kcal/mol) are shown. Genes represent names of related downstream genes. $E(R_1)$ and $E(R_2)$: consensus energy with covariance and conservation score of R_1 and R_2, respectively, where R_1 and R_2 are the rank 1st and 2nd ConSLOpt structures according to associated energy barriers. COG represents the Clusters of Orthologous Groups of related proteins. Rfam shows the best matching RNA family in Rfam Database.

Genes	COG	Rfam	Riboswitch	$E(R_1)$	$E(R_2)$	$Cov(R_1)$	$Cov(R_2)$	$\mathcal{B}$ (R_1, R_2)	PairId
hisZ	COG3705	–	–	−49.83	−45.2	1.33	1.03	32.67	0.94
greA	COG0782	–	–	−41.95	−39.18	0.9	0.73	44.72	0.9
yjcI	COG0626	RF00162^{+*}	SAM	−37.74	−33.27	3.75	3.5	30.48	0.75
yxkD	COG1284	RF00442^{+*}	ykkC-yxkD	−42	−33.06	3.7	3.4	29.02	0.79
ileS	COG0060	RF00230^{+*}	T-box	−40.02	−32.83	1.57	0.58	16.43	0.89
glyQ	COG0752	RF00230*	T-box	−35.25	−30.27	−0.03	0.55	45.9	0.79
thiM	COG2145	RF00059*	TPP	−34.88	−29.9	0.43	0.6	18.72	0.97
yugI	COG1098	–	–	−31.22	−29.52	3.62	3.45	21.95	0.88
trpE	COG0147	RF00230^{+*}	T-box	−36.37	−29.23	0.78	0.35	17.59	0.96
cysE	COG1045	RF00230*	T-box	−32.57	−28.9	3.53	2.17	20.52	0.79
ylxS	COG0779	–	–	−30.13	−28.75	−0.45	0.25	14.61	0.86
hutH	COG2986	–	–	−41.95	−28.02	0.47	0.93	16.99	0.96
glyS	COG0751	RF00230*	T-box	−35.11	−27.45	−1.35	0.15	38.54	0.8
leuS	COG0495	RF00230^{+*}	T-box	−34.32	−26.35	2.98	1.67	13.28	0.68
yrhG	COG2116	–	–	−37.07	−25.65	0.35	−0.15	14.02	0.88
argH	COG0165	–	–	−28.44	−25.38	0.2	0	21.55	0.97
secG	–	–	–	−29.97	−25.25	0.18	0.35	12.24	0.9
pyrH	COG0528	–	–	−33.6	−24.92	0.63	0.35	12.17	0.9
secDF	COG0342	–	–	−25.02	−24.28	1.45	0.97	17.27	0.94
tenA	COG0819	RF00059^{+*}	TPP	−29.73	−24.27	1.78	1.2	16.96	0.81
narH	COG1140	–	–	−29.12	−24.18	0	0.25	22.62	0.97
infC	COG0290	RF00558^{+*}	L20-leader	−24.82	−23.9	0.1	−0.08	23.57	0.88
ilvB	COG0028	RF00230^{+*}	T-box	−32.95	−23.85	1.32	0.8	25.71	0.82

Genes	COG	Rfam	Riboswitch	$E(R_1)$	$E(R_2)$	$Cov(R_1)$	$Cov(R_2)$	$\mathcal{B}$ (R_1, R_2)	PairId
glmS	COG0449	RF00234^{+*}	glmS	−26.98	−23.77	3.3	4.23	16.53	0.6
proI	COG0345	RF00230^{+*}	T-box	−33.12	−23.57	1.15	1.57	17.13	0.85
ykkC	COG2076	RF00442^{+*}	ykkC-yxkD	−25.91	−22.86	1.79	2.29	18.34	0.81
cysH	COG0175	RF00162*	SAM	−25.32	−22.52	−1.87	−0.42	12.1	0.79
odhB	COG0508	–	–	−28	−22.42	−0.1	0.4	15.29	0.96
glyA	COG0112	–	–	−23.23	−22.4	0.52	0.37	14.72	0.85
glgA	COG0297	–	–	−33.15	−22.35	2.13	1.43	17.11	0.86
valS	COG0525	RF00230^{+*}	T-box	−32.28	−21.88	1.33	1.47	16.87	0.8
rtpA	COG0484	RF00230^{+*}	T-box	−32.49	−21.25	3	2.08	15.12	0.77
gabP	COG1113	–	–	−27.42	−21.12	2.12	1.58	23.42	0.76
ribD	COG1985	RF00050*	FMN	−29.25	−20.8	1.8	0.33	13.58	0.76
pyrG	COG0504	–	–	−23.55	−20.6	1.07	0.87	15.72	0.76
guaA	COG0519	RF00167*	Purine	−28.65	−20.45	0.68	0.92	16.7	0.93
atpD	COG0055	–	–	−21.12	−20.13	0.25	0.23	20.8	0.89
nadD	COG1057	–	–	−22.48	−20.12	2.55	1.6	12.86	0.8

+ *and* * indicate that the best matching RNA families were identified by BLAST and Infernal's cmsearch, respectively. $\mathcal{B}(R_1, R_2)$ denotes the approximated consensus energy barrier between R_1 and R_2. PairId is the mean pairwise identity among orthologous sequences