Modeling Riboswitches by Spotting Out Switching Sequences

Marco Barsacchi [1], Eva Maria Novoa [2,3], Manolis Kellis [2,3], and Alessio Bechini[1,*]

[1]Univ. of Pisa, Dept. of Information Engineering, largo L. Lazzarino, 56122 Pisa, IT
[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA
[3]The Broad Institute of Massachusetts Institute of Technology and Harvard, 415 Main Street, Cambridge, MA 02142, USA

ABSTRACT

Motivation: Riboswitches are cis-regulatory elements in mRNA, mostly found in Bacteria, which exhibit two main secondary structure conformations. While one of them prevents the gene from being expressed, the other conformation allows its expression, and this switching process is typically driven by the presence of a specific ligand. Although there are a handful of known riboswitches, our knowledge in this field has been greatly limited due to our inability to identify them based on its sequence. Indeed, current methods are not able to predict the presence of the two functionally distinct conformations just from the knowledge of the plain RNA nucleotide sequence. Whether this would be possible, for which cases, and what prediction accuracy can be achieved, are currently open questions.

Results: Here we show that the two alternate secondary structures of riboswitches can be accurately predicted once the "switching sequence" of the riboswitch has been properly identified. For this aim, an algorithm to locate the switching sequence inside the complete riboswitch sequence has been developed, making use of a generated ensemble of configurations. The proposed approach is able to model the switching behavior of riboswitches whose generated ensemble covers both alternate configurations. Beyond structural predictions, the approach can also be paired to homology-based riboswitch searches.

1 INTRODUCTION

Gene regulation is essential to achieve organism versatility, giving cells control over structure and functions and being the basis for cellular differentiation, morphogenesis, and adaptability. Since the discovery of ribozymes, it has become more and more evident that RNA molecules are also actively involved in regulatory mechanisms, including the regulation of gene expression (Waters and Storz, 2009).

Riboswitches are RNA elements, mainly found in bacteria, that are embedded in 5'-untranslated regions (UTRs) of mRNA (Garst *et al.*, 2011). They are able to sense cellular metabolites with no involvement of protein factors, and consequently modulate either mRNA transcription or translation by adopting one out of two possible structures (Serganov and Nudler, 2013), known as "ON" and "OFF" conformations. Riboswitches are usually built around an *aptamer* domain, which binds to the ligand, and an *expression platform* domain, which undergoes a structural rearrangement upon the ligand-aptamer binding (Garst *et al.*, 2011). Furthermore, a central role in regulation has been recognized to an overlapping region between these two domains, referred to as *switching sequence* (or "SwSeq" for short hereafter). From an evolutionary standpoint, aptamers are typically highly conserved, as a consequence of its recognition specificity towards the ligand. In contrast, expression platforms are usually far less conserved (Breaker, 2012).

Previous works have shown that riboswitches can regulate gene expression via three main mechanisms: i) *transcription termination*, though the formation of a terminator hairpin or a competitive anti-terminator structure, ii) *translation inhibition*, by sequestering or releasing the Shine-Dalgarno (SD) sequence, and iii) *alternative splicing regulation*, via sequestration/release of alternative splicing sites (Peselis and Serganov, 2014). Understanding the mechanisms of riboswitch structural switching is of broad interest and applicable to a wide range of scientific fields, such as systems biology or drug design, e.g., to design novel engineered genetic circuits (Wittmann and Suess, 2012), or develop riboswitch-targeting drugs (Lünse *et al.*, 2014).

Previous computational efforts targeting riboswitches can be categorized in *riboswitch gene finders*, and *conformational switch predictors* (Clote, 2015). The former category includes several tools, such as Infernal, the founding component of the Rfam database (Nawrocki *et al.*, 2009), and more specific tools such as RibEx (Abreu-Goodger and Merino, 2005) or RiboSW (Chang *et al.*, 2009). Such methods are primarily used for genome-wide analyses, and are based on machine learning approaches. The second category encompasses various methods based on a structural classification of alternative structures, such as paRNAss (Voss *et al.*, 2004), RNAshapes (Janssen and Giegerich, 2014), and RNAbor (Freyhult *et al.*, 2007). Family-specific approaches for ON/OFF structure prediction have been employed as well (Clote *et al.*, 2012).

Here we present a novel computational method that can predict the two functional conformations of riboswitches, using merely as input the plain RNA nucleotide sequence. Importantly, this developed procedure can also be coupled with a classifier to identify putative riboswitch sequences, i.e. to uncover riboswitches based on its potential to generate two alternate configurations.

*to whom correspondence should be addressed

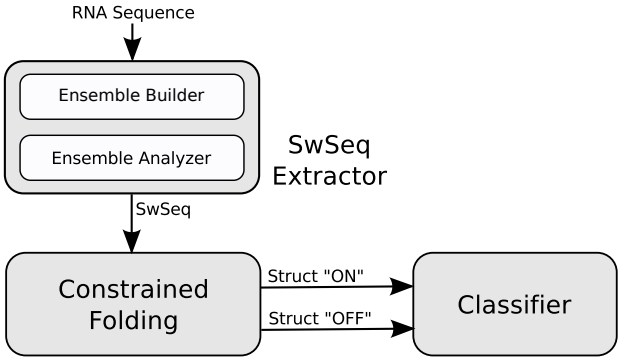

Fig. 1: Schematic view of the SwiSpot approach. Starting from an RNA sequence, a putative switching sequence is identified by the *SwSeq Extractor*. Such result is exploited by the *Constrained Folding* module to derive the two alternate structures. Finally, a *classifier* decides if a switching behavior can be associated to them.

2 SYSTEM AND METHODS

The presented work stems from investigations initially carried out to answer a simple question: *Is it possible to computationally predict the two functional conformations of a riboswitch once the SwSeq has been identified?* By imposing loose constraints on the pairings of the SwSeq bases, the RNAfold tool in the ViennaRNA package (Lorenz *et al.*, 2011) was able to closely approximate the two reference alternative structures in ten out of twelve case studies, while only minor errors were present in the outcomes for the other two (Table S1; further details in Section 3.2).

Such results suggest that spotting out the switching sequence may represent the basic step for the prediction of the two functional RNA conformations.

Throughout this work we used the ViennaRNA package, ver. 2.1.9, (Lorenz *et al.*, 2011), which is based on a validated set of thermodynamic parameters (Mathews *et al.*, 2004) and produces precise energy models.

2.1 Overall Approach

Although it has been found that tertiary structural arrangements affect the riboswitch functionality (Greenleaf *et al.*, 2008), their influence can hardly be captured by a simple, handy computational model. Instead, secondary structures can be investigated and analyzed in a much more efficient way. For these reasons, we shall focus on RNA *secondary structures* and their representations.

Let $r = r_1 r_2 \dots r_{l_r}$ be the RNA sequence of length l_r over the ordinary RNA alphabet; a secondary structure can be indicated by the integer row vector $p(r)$ (a "pairing") whose j-th element $p(r)[j]$ holds either the index of the element the j-th base is paired with, or j in case of no pairing. An ensemble of structures for r is a multiset of N pairing vectors $p(r)_i\,, i = 1 \dots N$, not necessarily distinct, stacked in the $N \times l_r$ matrix E_r. Thus, $E_r[i]$ indicates the i-th conformation in E_r, whose j-th element is $E_r[i, j]$. In this context, SwSeqs in the two alternate conformations show alternate pairing patterns (towards either upstream or downstream bases), thus suggesting a pattern-based method to identify and locate SwSeqs whenever they are not known.

The processing flow that leads from the input RNA sequence down to the possible alternate conformations and the evaluation of a riboswitch-like behavior can be structured as shown in Figure 1. It comprises three main functional modules indicated as *SwSeq Extractor*, *Constrained Folding*, and *Classifier*, respectively aimed at identifying the switching sequence, at deriving the alternate configurations, and at evaluating their potential to support a typical riboswitch behavior. Their implementations can be separately modified to improve their accuracy and/or efficiency.

The proposed method has been assessed against a set of well-known riboswitches, whose functional structures have been carefully described. Hence, a reference dataset has been defined for this purpose.

2.2 Dataset

Currently, riboswitches have been classified in more than twenty distinct classes (Breaker, 2012), according to significant structural and/or sequence similarity. At present, hundreds or thousands of representatives per class exist in the current RNA databases such as Rfam (Nawrocki *et al.*, 2015). However, a proper evaluation of the proposed approach requires a complete knowledge of both the ON and OFF conformations, and this is available only for a very small number of cases.

For the sake of our investigation, both the aptamer and the expression platform domain have to be precisely located in the riboswitch sequences. Unfortunately, in literature and in databases such as Rfam often only the expression platform is given. Thus, the complete sequence and the missing details must be obtained through a systematic review of studies on single riboswitches, whenever available. The riboswitch records used as our reference dataset and reported in Table 1 have been defined carrying out this reviewing work and, to the best of our knowledge, they can be regarded as one of the most comprehensive riboswitch dataset to date, at least in terms of the complete sequence set (aptamer + expression platform). It covers the main riboswitch families and the major regulatory mechanisms.

Our complete reference dataset contains forty records, divided in three groups. The first two groups include twenty riboswitches from sixteen families (Table 1). The first group encompasses twelve sequences (some already used in previous studies (Quarta *et al.*, 2012)) with reliable SwSeq information (first part of Table 1). For each record in the group, whenever not explicitly reported in the literature, the switching sequences have been identified by comparing the two known alternative secondary structures. The second group contains eight riboswitches whose SwSeq has not been identified yet (second part of Table 1); despite this lack of information, they can be used to validate the overall method, but not the SwSeq Extractor. The twenty elements in the last group (shown in Table S2) have been used as negative controls; they are not riboswitches, and consist of ncRNAs, in the same length range of the previous groups, selected

Table 1. Riboswitches of the reference dataset

ID[a]	Riboswitch Class	Gene - Organism	Switching Mechanism	Boltzmann Cov.[b]
thiM_TPP	Thiamine Pyrophosphate	*thiM - Escherichia coli*	Translation	YES
add_Adenine	Adenine	*add - Vibrio vulnificus*	Translation	YES
folT_THF	Tetrahydrofolate	*folT - Alkaliphilus metalliredigens*	Translation	N.A.
xpt_Guanine	Guanine	*xpt - Bacillus subtilis*	Transcription	NO
pbuE_Adenine	Adenine	*pbuE (ydhL) - Bacillus subtilis*	Transcription	NO
mtgE_Mg	Magnesium	*mgtE - Bacillus subtilis*	Transcription	YES
yitJ_SAM	S-adenosylmethionine	*yitJ - Bacillus subtilis*	Transcription	YES
lysC_Lysine	Lysine	*lysC - Bacillus subtilis*	Transcription	YES
tenA_TPP	Thiamine Pyrophosphate	*tenA - Bacillus subtilis*	Transcription	YES
metH_SAH	S-adenosylhomocisteine	*metH - Dechloromonas aromatica*	Transcription	N.A.
VEGFA	Het. nuclear ribonucleoprotein L	*VEGFA - Homo sapiens*	Alt. Splicing	YES
thiC_TPP	Thiamine Pyrophosphate	*thiC - Arabidopsis thaliana*	Alt. Splicing	N.A.
moaA_Moco	Molybdenum Cofactor	*moaA - Escherichia coli*	Translation	
btuB_Cobalamin	Cobalamin	*btuB -Escherichia coli*	Translation	
alx_PH	PH	*alx - Escherichia coli*	Translation	
GEMM_CDA	Cyclic-di-Guanosine Monoph.	*Daud_1768 - Desulforudis audaxviator*	Transcription	
crcB_Flouride	Flouride	*crcB - Bacillus cereus*	Transcription	
PreQ1	Pre-Queosine	*glyS - Fusobacterium nucleatum*	Transcription	
ydaO_ATP	ATP	*ydaO - Bacillus subtilis*	Transcription	
metA_SAH	S-adenosylmethionine	*metA - Agrobacterium tumefaciens*	Transcription	

(a) Riboswitches have been subdivided into those for which the SwSeq is known (upper panel) and not known (lower panel). (b) Boltzmann coverage has been evaluated with the extended sampling approach described in text. List of ncRNA used as negative controls are shown in Table S2. Full description of the reference dataset, including the annotated switching sequences, can be found in Table S1.

from the Rfam database (version 12.0, see `http://rfam.xfam.org`) to be representative of heterogeneous secondary structures. Among them, we have taken care to include RNAs that exhibit other types of switching behaviors.

3 ALGORITHMS

3.1 SwSeq Extractor

A central question in the proposed approach is how to spot the switching sequence out of an RNA sequence r, using no additional information. This task is assigned to the SwSeq Extractor module.

Scarceness of accurate structural data on riboswitches makes it impossible to follow a typical machine-learning approach. Moreover, recent studies on entropy contents of riboswitch sequences (Manzourolajdad and Arnold, 2015) do not focus on SwSeqs. Thus, we propose to use an ensemble (a multiset) E_r of conformations, which will likely contain conformations close to the two alternative structures. Once E_r has been defined, base pairings across different conformations in E_r must be analyzed to identify the SwSeq. These two tasks, named "Ensemble Builder" and "Ensemble Analyzer", are integral parts of the SwSeq Extractor module (Figure 1).

The actual content of E_r determines the effectiveness of the whole SwSeq extraction procedure. An implicit choice for E_r can be taken by referring to the McCaskill algorithm (McCaskill, 1990) to compute the frequency of *all* possible base pairings at equilibrium at a given temperature, subsumed in the matrix P. In this case, the ideal reference E_r accounts for all the conformations and their specific stability. We shall derive results according to this choice first, and then we shall compare them with results from a stochastic approach in building E_r.

The SwSeq is identified by the "Ensemble Analyzer", which searches for the subsequence in r that most frequently shows *alternative base-pairing* (upstream or downstream) across structures in E_r. We propose to capture this behavior by means of a proper scoring method; the predicted SwSeq would then be the one with the maximum score. The corresponding pseudo-code is shown in Algorithm S1.

In defining the scoring method, we make use of the basic promoting score $s_p > 0$, and the basic penalty $s_n < 0$. Let us consider the subsequence $r[j...j+l]$ of length l starting at index j; its *mean frequency of upstream/downstream base-pairing*, indicated respectively by $F_{up}(j,l)$ and $F_{do}(j,l)$, can be calculated through the McCaskill pairing probability matrix P (the probability for base pair (x,y) is indicted here as $P[x,y]$):

$$F_{up}(j,l) = \frac{1}{l} \sum_{h=j}^{j+l-1} \sum_{i=0}^{h-1} P[i,h] \tag{1}$$

$$F_{do}(j,l) = \frac{1}{l} \sum_{h=j}^{j+l-1} \sum_{i=h+1}^{N} P[i,h] \tag{2}$$

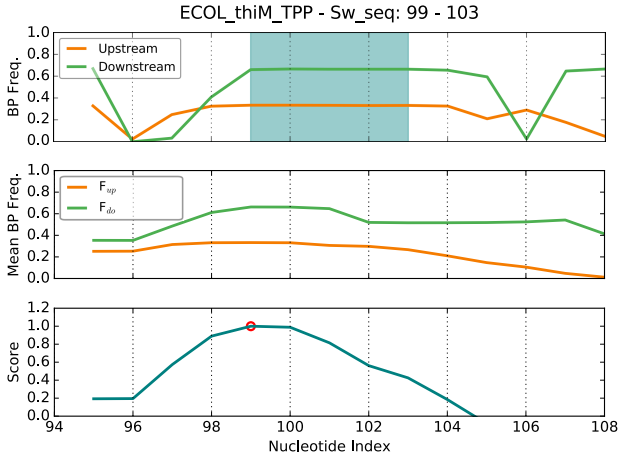

Fig. 2: In the upper two panels, upstream/downstream base pairing frequencies (respectively per single base, and over subsequences of length $l = 5$), in the neighborhood of the actual SwSeq (indicated by the shaded band) for *E. coli* thiM_TPP. In the lower panel, the corresponding score (Eq. 4) with the indicated maximum point (in red).

A synthetic way to express the tendencies of $r[j...j+l]$ to pair upstream and downstream is

$$\begin{array}{lcl} pairUp(j,l) & = & s_p \cdot F_{up}(j,l) + s_n \cdot F_{do}(j,l) \\ pairDown(j,l) & = & s_p \cdot F_{do}(j,l) + s_n \cdot F_{up}(j,l) \end{array} \tag{3}$$

and the propensity for $r[j...j+l]$ to show both of them can be indicated by

$$score(j,l) = pairUp(j,l) \cdot pairDown(j,l) \tag{4}$$

According to the proposed scoring model, the pair of indices (j^*, l^*) yielding the maximum for $score(j,l)$ indicates the putative switching sequence.

The actual values chosen for for s_p and s_n *may* affect the overall classification performance. Thus, according to tests performed on the reference dataset and discussed in Supplementary Material, they have been set to $s_p = 1$ and $s_n = -0.9$, yielding the most satisfying classification performance (Figure S1).

Figure 2 shows how base pairing frequencies can be exploited to build the described score. The upper chart reports the plain *per single base* upstream/downstream pairing frequencies. The middle chart reports the mean frequencies *per target subsequence*, referred to its starting position. Finally, the lower chart shows the overall score, as per Eq. 4. The starting position of the putative SwSeq corresponds to the base index yielding the maximum score (indicated by a red dot).

It might be argued that working with the overall pairing probability matrix P would account also for very unlikely conformations, loosing the focus on the most representative ones for our goal. Thus another possibility is to make use of an actual E_r, populating it via specific sampling methods. In this regard, the baseline option in our work was a Boltzmann-based sampling through a well-known algorithm (Ding and Lawrence, 2003).

For E_r to be representative of the real energy landscape, a very accurate energy model is required. It must be underlined that the Turner energy model is inherently unable to account for the ligand contributions in one of the structures. Moreover, the influence of kinetic phenomena in the formation of some native structures cannot be easily captured. As a consequence, as shown in Figure 3, by using a plain Boltzmann sampling not in all cases E_r covers the neighborhoods of the two alternate functional conformations: for thiM_TPP (a), both neighborhoods are encompassed, while for xpt_Guanine (b) one conformation is completely out of the E_r landscape. In practice, whenever conformations similar to one of the two alternatives are not adequately represented in E_r, it is unlikely to identify the SwiSeq by analyzing pairing frequencies.

In an attempt to partially address this limitation, we have chosen an ensemble size of 1200 structures and, to increase the representativeness of samples in E_r, we have applied an artifice proposed also in (Quarta *et al.*, 2012), by performing the sampling at different temperatures. By using the RNAsubopt tool, we sampled 300 structures at the default value of $37°C$ and we added 150 structures per each temperature value at six decile intervals towards the melting temperature of the RNA strand.

Once the reference ensemble has been built, we can refer to the same scoring method previously described. Thus, we can express F_{up} and F_{do} on the basis of actual structures in E_r. Using the step function $H(x)$ (defined as $H(x) = 0$ for $x \leq 0$, $H(x) = 1$ otherwise), these

values can be calculated as:

$$F_{up}(j,l) = \frac{1}{l \cdot N} \sum_{h=j}^{j+l-1} \sum_{i=1}^{N} H(h - E_r[i,h]) \tag{5}$$

$$F_{do}(j,l) = \frac{1}{l \cdot N} \sum_{h=j}^{j+l-1} \sum_{i=1}^{N} H(E_r[i,h] - h) \tag{6}$$

This possible approach is intrinsically stochastic, thus not reproducible in principle. Although a small variance exists across predicted SwSeqs from different runs, its effect on the two predicted conformations is minimal, with a negligible variance of the normalised bp-distance between conformations predicted in different runs (values are reported in Table S3).

3.2 Constrained Folding

The Constrained Folding module predicts the two alternate configurations, given a putative riboswitch sequence and its corresponding SwSeq. The founding idea is that SwSeq bases should pair with upstream bases in one case, and with downstream bases in the other: This can be forced as a (soft) constraint within the RNA folding algorithm. We have chosen RNAfold from the ViennaRNA package, because of its ability to accommodate the required constraint, specifying via a mask string, where a base may be paired upstream ('<'), downstream ('>'), or not constrained at all ('.').

The implementation of this module was tested on the first group of our reference dataset, i.e. containing riboswitches with experimentally validated SwSeqs. We have been able to closely predict the two alternate configurations in $\sim 80\%$ of the cases, and reconstruct the remaining ones with very minor errors (Table S1). Notably, this was the result that pushed us to shape the whole approach.

In a nutshell, we can conclude that the plain knowledge of the switching sequence is generally sufficient to predict the two functional structures of the riboswitch (Figure S3 for further consideration on switching sequences).

3.3 Classifier

Our method derives two alternative structures for a target RNA sequence and it is up to the Classifier module to determine whether they may be associated to a typical riboswitch behavior. The proposed solution relies on the identification of secondary motifs known to be related to specific gene regulation mechanisms in riboswitches. The extent to which such motifs are present in the input conformations is quantified by means of specific indices, namely i_{SD} for *SD sequestering*, and i_{TT} for *transcription termination*. Because of its complexities and subtleties, *alternative splicing* has not been directly addressed. In its current form, the Classifier can be sensibly applied only to prokaryotic sequences.

To quantify translation inhibition, we first locate the SD pattern in the sequence, and then count the fraction of unpaired bases n_{ub} in the found SD sequence interval (Chen *et al.*, 1994) in each of the two alternate structures. The greater n_{ub}, the lower the sequestering of the SD site. Our search consisted on the following consensus sequence: AGGAGG, followed by a start codon (AUG), separated by 5-10 bases from the SD site. If a given sequence does not possess a terminator (see equation 7 below), it is searched against one of the patterns in the database of small SD, located within a small distance from the end of the sequence. Once found the SD, we measure the difference between the unpaired fraction of the found SD in the two predicted structures. The index i_{SD}, $0 \leq i_{SD} \leq 1$ can be defined as the absolute difference of n_{ub} in the alternative conformations, over the SD length.

Transcription termination is typically determined by a "terminator hairpin" (Wilson and von Hippel, 1995), which can be detected as a stem-loop-like motif, followed by a run of 'U's. In practice, this can be performed by substituting dots in the dot-bracket notation with the symbols of the corresponding bases, and then looking for matches of one of the following patterns: the first, of the form

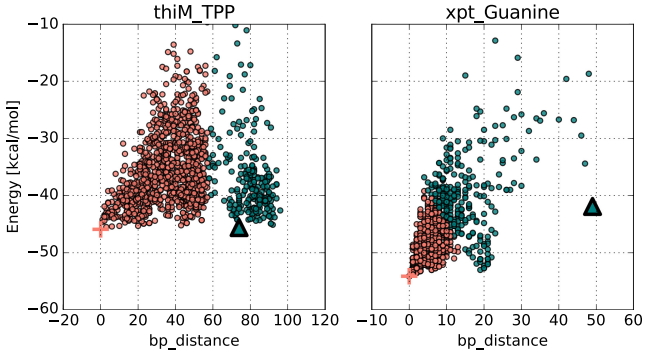

Fig. 3: Sampled and clustered structural ensembles for two different riboswitches. The two colours represent the clusters. The placement of ON/OFF conformations is indicated by the $+/\Delta$ symbols. In the right panel, the ensemble does not cover both of them (for sampled structural ensembles for the full reference dataset, see Figure S2).

"({3,}[ACGU]{3,8}){3,}U{3,}.{,20}$" [1] to identify long hairpins, and the other, "){8,}U{3,}.{,20}$", to catch terminators with more complex topology[2]. A termination-based riboswitch is correctly predicted if at least one of the patterns is matched *only in one* of the alternative structures. In this case, we set $i_{TT} = 1$; $i_{TT} = 0$ otherwise. It has been proposed that an RNA switch could be predicted by analyzing its energy landscape (Clote, 2015). According to this hypothesis, an RNA switch ideally shows two distinct clouds of points in the conformational landscape. This feature, although not exclusively held by riboswitches, could indeed be used to strengthen the classification procedure. For this reason, to quantify how well two separate clusters can be spotted in E_r, we perform a 2-medoids clustering on the ensemble, based on pairwise bp-distances (other possible metrics would derive similar results (Barsacchi *et al.*, 2016)). We then characterize the result with the silhouette index i_{sil}, i.e. by computing the mean of the silhouette coefficients for all the elements in the set (Rousseeuw, 1987). This is one of the most popular indexes to quantify clustering quality, and it approaches 1 for increasing quality.

The overall classification outcome can be based on a global index that comprises the previous ones:

$$I_r = i_{SD}(1 - i_{TT}) + i_{TT} + i_{sil} \tag{7}$$

A riboswitch behavior will be foreseen in all cases if I_r is beyond a given threshold value I_t.

3.4 Implementation

Algorithms have been initially implemented in Python, and the software have been built upon a Python wrapper from ViennaRNA library, compiled from sources provided by TBI (`http://www.tbi.univie.ac.at/RNA/`). A version in C language is available as well, suitable for faster computations. Being based on `RNAsubopt` and `RNAfold`, the current algorithm implementation does not support pseudoknotted structures.

The runtime of the Ensemble Analyzer increases as $O(l_r \cdot wlen_{max}) \approx O(l_r)$, where l_r and $wlen_{max}$ indicate the sequence length and the maximum length of the SwSeq, respectively. The complexity of the McCaskill algorithm is $O({l_r}^3)$, while the complexity of the suboptimal Boltzmann sampling is $O(n \cdot {l_r}^3 + N^2)$ (Ding and Lawrence, 2003), with N as the ensemble cardinality. In practice, in the ordinary ranges for such parameters, the runtime shows a linear dependence on input sequence length (Figure S4), and it takes a few seconds per sequence, thus allowing this method to be applied on large datasets in a genome-wide fashion.

4 RESULTS

The method presented here is capable of predicting the presence of a riboswitch — including its two alternate conformations — using as input only its nucleotide sequence. The procedure consists in three major steps, including: i) prediction of the switching Sequence (SwSeq), ii) prediction of the alternate secondary structure conformations, and iii) scoring, and consequent classification, of the sequence as a putative riboswitch (Figure 1). The proposed approach has been tested on the reference dataset, obtaining the results reported in Table S4; moreover, it has been used to investigate a complete set of annotated putative riboswitches from the Rfam database (Rfam 12.0). Our method is fully applicable to any prokaryotic organism at a genome-wide level, targeting single sequences up to 300nt each.

4.1 Application to the Reference Dataset

The ability to spot out the SwSeq depends both on the analysis procedure applied to the ensemble, and on the ensemble representativeness of the two functional conformations. Indeed, considering the computational steps in the SwSeq Extractor and the formulation of Eq. 4, the sampling procedure itself, whenever used, affects the reliability of the outcomes in this module (Figure 3).

We first tested the SwSeq Extractor module on the first group of our reference dataset, obtaining 10 exact SwSeq predictions out of 12 cases. We then predicted SwSeqs also for sequences in the second group, obtaining the putative alternative structures for all sequences in both groups; Figure 4 shows an example of such predicted conformations for tenA_TPP, highlighting the SwSeq location. It should be noted that, even in those cases in which an exact SwSeq prediction could not be made, the located subsequence is similar enough to the true SwSeq, thus allowing the Constrained Folding module to correctly identify the alternate conformations (as it happens for xpt_Guanine). Therefore, although SwSeq predictions may not exactly match SwSeq annotations from the literature, they are accurate enough to be effective in revealing the switching behavior.

Considering the complete reference dataset, the classifier module performance is summarized as a ROC curve, where sensitivity and specificity are respectively formulated as $TPR = TP/(TP + FN)$ and $SPC = TN/(TN + FP)$; true positives *TP* are the correctly predicted riboswitches, true negatives *TN* the correctly predicted non-riboswitches. Figure 5 reports the curves obtained using the McCaskill method and the ensemble-based analysis.

We have observed that a good trade-off between accuracy and *TPR* is obtained choosing a threshold value $I_t = 0.45$. Hence, predictions using the McCaskill method exhibit an accuracy of 0.6 and a true positive rate (*TPR*) of 0.83, while the stochastic method on average yields an Accuracy of 0.6 and $TPR = 0.7$. All the outcomes of the Classifier module on the reference dataset are reported in Table S4.

[1] Typically in regexes "(" and ")" are used as group delimiters; thus, parentheses symbols should be indicated by "\(" and "\)" instead.

[2] Although unlikely, this pattern might introduce false positives in the case of a stem loop within a multiloop which is followed by a run of 'U's and is near the end of the sequence.

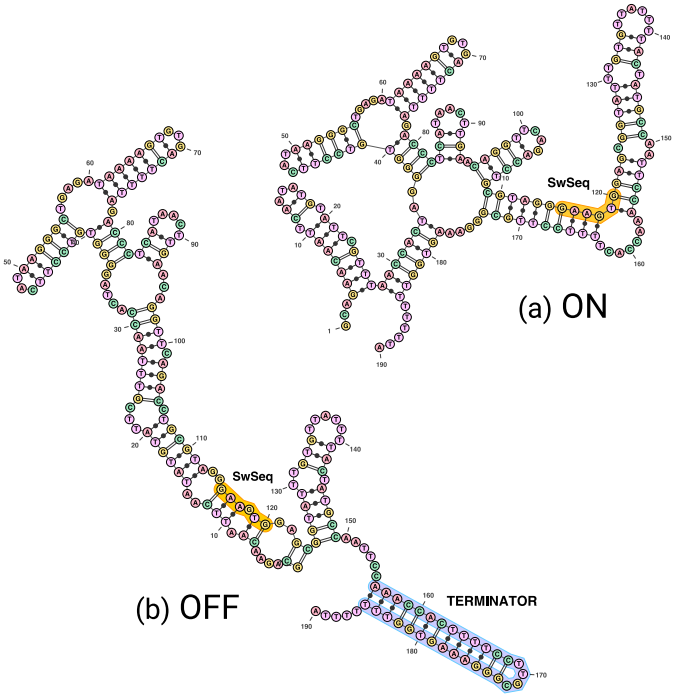

Fig. 4: Example of predicted alternate conformations ON (a) and OFF (b) for the tenA_TPP riboswitch. The switching sequence is highlighted in orange, and the terminator hairpin is highlighted in blue.

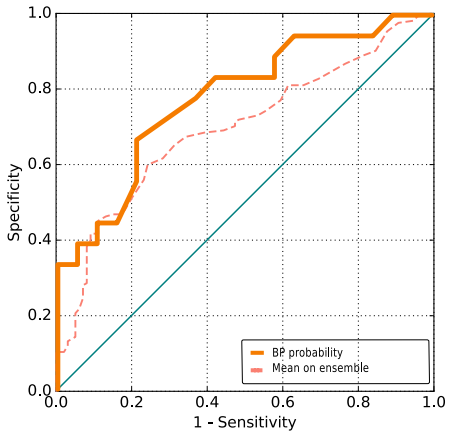

Fig. 5: ROC curve of performance on the complete reference dataset. The upper ROC curve has been obtained using base-pairing probability matrix, while the dashed curve represents the mean of 10 runs per ensemble.

We then proceeded to quantify the influence of the ensemble coverage on the classification outcomes. For this aim, we compared the classification outcome of all the riboswitches in the reference dataset, with the outcome of selected riboswitches that have E_r coverage. Figure 6 clearly shows that the proposed approach is more effective whenever E_r coverage is present.

4.2 Application to Rfam Putative Riboswitches

Currently, several motif search methods exist to uncover riboswitches in genomes. However, they typically identify riboswitches on the basis of the aptamer region only (Clote, 2015). Other methods search for conformational switches, avoiding specific aptamer considerations.

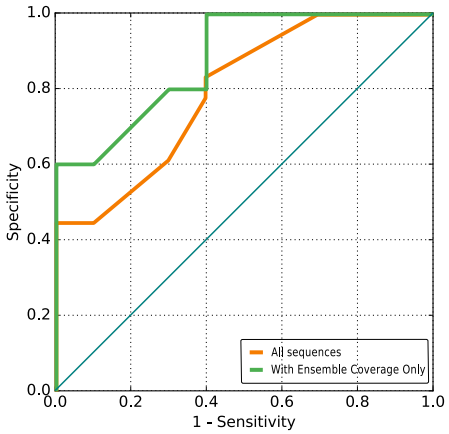

Fig. 6: Influence of ensemble coverage on effectiveness, evaluated in the case of ensemble sampling. The green ROC curve refers to the set of only riboswitches with E_r coverage, while the orange ROC curve considers all riboswitches.

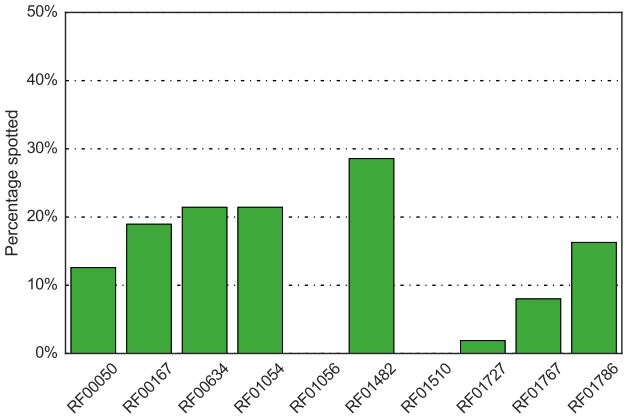

Fig. 7: Percentage of per-family prokaryote riboswitches identified in putative Rfam families containing riboswitches.

Our methodology can complement both of them, providing further structural insights. We applied our method upon a set of putative Rfam-annotated riboswitches, to illustrate how homology-based riboswitch searches can be coupled with our algorithm to obtain a higher confidence set of putative riboswitches. In practice, not only sequence similarity has to be taken into account, but also the potential of generating two alternative structures with a SwSeq.

For this aim, we first gathered the sequences included in the seed alignment for each of such families. We then used MAFFT (Katoh and Standley, 2013) to obtain a structure-based alignment of the seeds. Each alignment was assigned the relative "RNAz P-value" (Gruber *et al.*, 2012), accounting for both structural conservation and thermodynamic stability. The top-ranked families based on RNAz P-value were selected for further analysis; they included also some of the prokaryotic sequences described in our reference dataset. The target sequences were elongated downstream (to include the first occurrence of a start codon `ATG`) to assure the inclusion of both aptamers and switching sequences. We finally ran our algorithm on the these sequences, spotting the switching regulatory behaviour by checking either the SD sequestering/releasing or the presence of a terminator hairpin. A list of classified riboswitches has been obtained (downloadable from the SwiSpot website).

The histogram in Figure 7 summarizes per-family results: for each Rfam family, the percentage of sequences with a spotted switching regulatory behaviour is shown (see also Table S6). The low percentage of potential riboswitches actually spotted may be due to the presence of other types of switching mechanisms, an inaccurate sequence framing around the aptamer regions, or misannotations in Rfam , which may be in part due to ignoring the presence of expression platforms for their annotation.

4.3 Comparison with Other Approaches

Although other approaches for riboswitch prediction exist, none of them performs the two tasks that our method does, i.e. predicting if a sequence is a riboswitch, as well as identifying its switching sequence and its alternative conformations. Nevertheless, even if RNAbor does not produce an actual classification step, we have performed a general comparison with SwiSpot: by applying RNAbor to our reference dataset, peaks in the probability density profiles can be coarsely compared to SwiSpot results (see Figure S5 and Table S5). In general, multiple peaks in RNAbor profiles lead to interpretation problems, while Swispot provides clear outcomes. Considering the five RNAs with known ON/OFF structures in our dataset, that show also 2 distinct, well-defined RNAbor peaks, the structural predictions by SwiSpot are more accurate than those provided by RNAbor (average normalized bp-distance from reference structures: 0.319975 vs. 0.467507).

In Figure 8 we consider the particular case of thiM_TPP, previously used in the RNAbor validation (Freyhult *et al.*, 2007). The SD sequence is highlighted in red, while SwSeq pairings are highlighted in orange. The density plot for the complete sequence displays only one larger peak corresponding to the OFF structure, and a smaller peak for a partially formed ON structure; both conformations do not exactly match the reference ones. SwiSpot provides better results, with relative distances from the two annotated conformations -based on literature- of $(0.00, 0.00)$, while RNAbor peaks structure show less similarity with the reference: $(0.28, 0.09)$. This suggests that SwiSpot is capable of predicting existing alternative conformations regardless of the presence of distinct peaks in the probability density profile.

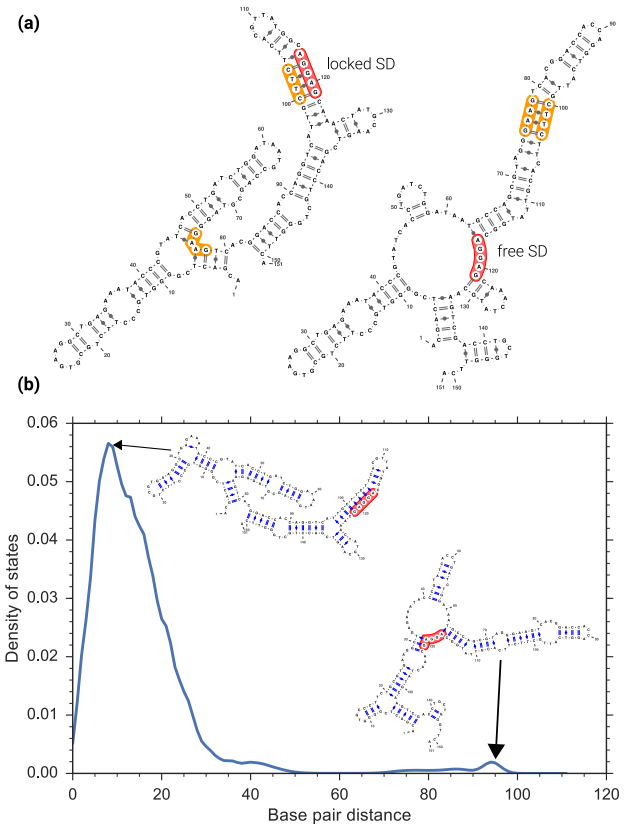

Fig. 8: A) Predicted alternate riboswitch structures according to the SwiSpot approach for thiM_TPP. The SD sequence is highlighted in red, while the SwSeq is highlighted in orange. B) RNAbor results, run on the same sequence, show only one peak for the OFF structure. A smaller peak with partially free SD is present.

5 DISCUSSION

Several methods for riboswitch prediction have been proposed, which are based on either predicting the presence of conformational switches, or on predicting riboswitches using homology-based searches (Clote, 2015). However, none of the current methods is focused on predicting the two riboswitch conformational structures. Such a capability would be extremely useful in driving riboswitch engineering and for synthetic biology. For this reason, we propose SwiSpot, a novel computational method to directly address conformational prediction of riboswitches.

The scarceness of precise data on riboswitch conformations makes it impossible to build a riboswitch predictor that merely uses a statistical or machine learning approach. Despite the mentioned data shortage, a reference dataset has been defined to properly assess the proposed approach. The proposed modeling effort stems from the first significant result presented in this work: we find a strong computational evidence of the role of the *switching sequence* in the characterisation of the whole riboswitch system. Moreover we find that the switching sequence incorporates most of the information required to drive the switching mechanisms.

The novelty of our approach relies on the ability to locate switching sequences, which are then used to predict the alternative conformations, and to finally classify the target RNA sequence as a riboswitch or not. Notably, previous genome-wide methods to search for riboswitches only looked for aptamer motifs (Clote, 2015), thus completely ignoring structural predictions. On the other hand, we show that SwiSpot can be paired to these methods, leading to more reliable predictions of putative riboswitches. Importantly, from this experience we observe that the quality of the results depends on the capability of selecting the proper subsequence to analyze, around the aptamer, down to the expression platform, and not beyond.

Lessons learned in using the proposed algorithms indicate possible future improvements, although apparently difficult to apply. Neglecting the energy contribution of the ligand and kinetic/co-transcriptional aspects of the folding process can likely lead to biased samplings of the energy landscapes, where conformations similar to both the alternative structures are not adequately represented. This issue could be addressed by introducing additional, diverse forms of sampling. Unfortunately, the lack of real structural data prevents any solid assessment of this kind of improvement. Furthermore, the limitation of excluding pseudoknots could be removed, devising new algorithms for the SwSeq Extractor and the Constrained Folder (Figure 1).

Future work will investigate possible improvements within each single module, as well as the provision of new functionalities in additional modules.

FUNDING

The work was partially supported by individual institutional funds from the University of Pisa. E.M.N. is supported by an HFSP Postdoctoral Fellowship (LT000307/2013-L).

Conflict of Interest: none declared.

REFERENCES

Abreu-Goodger, C. and Merino, E. (2005). RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res*, **33**(suppl 2), W690–W692.

Barsacchi, M., Baù, A., and Bechini, A. (2016). Extensive assessment of metrics on RNA secondary structures and relative ensembles. In *Proc. of 31st ACM Symp. on Applied Computing*, pages 44–47. ACM Press.

Breaker, R. R. (2012). Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol*, **4**(2).

Chang, T.-H., Huang, H.-D., Wu, L.-C., Yeh, C.-T., Liu, B.-J., and Horng, J.-T. (2009). Computational identification of riboswitches based on RNA conserved functional sequences and conformations. *RNA*, **15**(7), 1426–1430.

Chen, H., Bjerknes, M., Kumar, R., and Jay, E. (1994). Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res*, **22**(23), 4953–4957.

Clote, P. (2015). *Computational Methods for Understanding Riboswitches*, volume 553 of *Methods in Enzymology*. Elsevier.

Clote, P., Lou, F., and Lorenz, W. A. (2012). Maximum expected accuracy structural neighbors of an RNA secondary structure. *BMC Bioinformatics*, **13**(5), 1–18.

Ding, Y. and Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*, **31**(24), 7280–7301.

Freyhult, E., Moulton, V., and Clote, P. (2007). Boltzmann probability of rna structural neighbors and riboswitch detection. *Bioinformatics*, **23**(16), 2054–2062.

Garst, A. D., Edwards, A. L., and Batey, R. T. (2011). Riboswitches: structures and mechanisms. *Cold Spring Harb Perspect Biol*, **3**(6).

Greenleaf, W. J., Frieda, K. L., Foster, D. A. N., Woodside, M. T., and Block, S. M. (2008). Direct observation of hierarchical folding in single riboswitch aptamers. *Science*, **319**(5863), 630–3.

Gruber, A. R., FindeiB, S., Washietl, S., Hofacker, I. L., and Stadler, P. F. (2012). *RNAZ 2.0:*, chapter 9, pages 69–79. World Scientific.

Janssen, S. and Giegerich, R. (2014). The RNA shapes studio. *Bioinformatics*, **31**(3), 423–425.

Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Bio Evol*, **30**(4), 772–780.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithm Mol Biol*, **6**(1), 26.

Lünse, C. E., Schüller, A., and Mayer, G. (2014). The promise of riboswitches as potential antibacterial drug targets. *IJMM*, **304**(1), 79–92.

Manzourolajdad, A. and Arnold, J. (2015). Secondary structural entropy in RNA switch (riboswitch) identification. *BMC Bioinformatics*, **16**(1), 133.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *PNAS*, **101**(19), 7287–7292.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, **29**(6-7), 1105–1119.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**(10), 1335–1337.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2015). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*, **43**(D1), D130–D137.

Peselis, A. and Serganov, A. (2014). Themes and variations in riboswitch structure and function. *Biochim Biophys Acta*, **1839**(10), 908–918.

Quarta, G., Sin, K., and Schlick, T. (2012). Dynamic energy landscapes of riboswitches help interpret conformational rearrangements and function. *PloS Comp Biol*, **8**(2), e1002368.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*, **20**, 53–65.

Serganov, A. and Nudler, E. (2013). A decade of riboswitches. *Cell*, **152**(1-2), 17–24.

Voss, B., Meyer, C., and Giegerich, R. (2004). Evaluating the predictability of conformational switching in RNA. *Bioinformatics*, **20**(10), 1573–1582.

Waters, L. S. and Storz, G. (2009). Regulatory RNAs in bacteria. *Cell*, **136**(4), 615–28.

Wilson, K. S. and von Hippel, P. H. (1995). Transcription termination at intrinsic terminators: the role of the RNA hairpin. *PNAS*, **92**, 8793–8797.

Wittmann, A. and Suess, B. (2012). Engineered riboswitches: Expanding researchers' toolbox with synthetic RNA regulators. *FEBS Letters*, **586**, 2076–2083.