Thermodynamic and kinetic folding of Riboswitches

Stefan Badelt[a], Stefan Hammer[a,b], Christoph Flamm[a,*], Ivo L. Hofacker[a,b]

[a] *Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17/3, A-1090 Vienna, Austria*
[b] *Research Group Bioinformatics and Computational Biology, University of Vienna, Währingerstraße 29, A-1090 Vienna, Austria*

Abstract

Riboswitches are structured RNA regulatory elements located in the 5'-UTRs of mRNAs. Ligand-binding induces a structural rearrangement in these RNA elements, effecting events in downstream located coding sequences. Since they do not require proteins for their functions they are ideally suited for computational analysis using the toolbox of RNA structure prediction methods. By their very definition riboswitch function depends on structural change. Methods that consider only the thermodynamic equilibrium of an RNA are therefore of limited use. Instead, one needs to employ computationally more expensive methods that consider the energy landscape and the folding dynamics on that landscape. Moreover, for the important class of kinetic riboswitches, the mechanism of riboswitch function can only be understood in the context of co-transcriptional folding. We present a computational approach to simulate the dynamic behavior of riboswitches during co-transcriptional folding in the presence and absence of a ligand. Our investigations show that the abstraction level of RNA secondary structure in combination with a dynamic folding landscape approach is expressive enough to understand how riboswitches perform their function. We apply our approach to a experimentally validated theophylline binding riboswitch.

Keywords: riboswitch, kinetic folding, dynamic landscape, RNA secondary structure, co-transcriptional folding, theophylline aptamer

*Corresponding author
Email address: xtof@tbi.univie.ac.at (Christoph Flamm)

PACS: [2010] 87.15.A-, 87.14.gn, 87.15.db, 87.15.hm, 87.15.hp, 87.10.Mn

1. Introduction

The last decades witnessed a dramatic expansion of our knowledge on RNA as a regulatory molecule. A myriad of functional small RNAs influencing a diverse set of cellular processes have been described for bacteria and eukaryotes. Among them are riboswitches Serganov & Nudler (2013), structured RNA elements located in the 5'-UTR of mRNAs Nudler & Mironov (2004), that are capable of regulating gene expression. Regulation works either on the transcriptional or on the translational level. Translational riboswitches regulate the formation of the translation initiation complex which enables them to switch between an on- and off-state. In contrast, transcriptional riboswitches induce early termination of the whole transcription process and are therefore not reversible.

Usually, riboswitches are composed of two parts (i) a relatively conserved aptamer domain responsible for ligand-binding, and (ii) a variable sequence region termed expression platform for regulating the downstream located coding sequences. The ligand recognition sites vary greatly in size and complexity of their secondary and tertiary structures. Environmental stimuli like temperature changes or the binding of ligands such as ions, enzyme co-factors, RNA or DNA trigger switching due to changes in the expression platform which are then translated into a modulation of downstream events.

Riboswitches can furthermore be classified into thermodynamic and kinetic switches. Thermodynamic switches are found in energetic equilibrium between their on- and off-state. If switching is triggered, the equilibrium distribution shifts towards the new energetically best conformation. This implies that thermodynamic switches can reversibly and repeatedly toggle between on- and off-states. In contrast, kinetic switches are trapped in one state, depending on whether the trigger was present at the time of folding. The functional states correspond to local minima of the energy landscape that cannot be escaped during the lifetime of the molecule. Therefore, kinetic switches are not reversible

without addition of extrinsic energy, RNA turnover through degradation and synthesis is responsible for changing the state of a cell.

The high modularity in the structural architecture of riboswitches allows for a high degree of functional portability to other contexts making riboswitches an highly attractive design target to achieve context dependent gene regulation in synthetic biology Isaacs et al. (2004); Dawid et al. (2009); Qi et al. (2012); Rodrigo et al. (2013).

2. Characterization and prediction of riboswitches

Riboswitches implement a particularly direct mechanism of gene expression, since they effect the expression of an mRNA via the structure of the RNA molecule itself without requiring any protein cofactors. Therefore this mechanisms can be easily modeled $in-silico$ on the level of secondary structures using well-established and efficient methods exist. Nevertheless a number of caveats make the application of such programs to riboswitches less than straightforward. (i) The effect of ligand binding on the aptamer structures is not included in energy models for RNA secondary structures. (ii) Many aptamer structures form pseudo-knots or complex tertiary structures ignored by secondary structure prediction and (iii) the commonly used methods for describing RNA molecules in thermodynamic equilibrium are insufficient for modeling riboswitches whose mechanisms depend on RNA folding kinetics.

The computational effort to characterize riboswitches is therefore dependent on the type of the riboswitch. Temperature dependent riboswitches can be modeled with standard free energy parameters, while modeling riboswitches that bind ligands need empirical data on the binding free energy. Also, thermodynamic switches can be characterized by methods predicting equilibrium properties, while kinetic switches require the much harder computation of folding kinetics.

Figure 1 shows a designed and experimentally tested example of a temperature dependent, thermodynamic switch. Computing the specific heat us-

ing `RNAheat` Hofacker et al. (1994) readily identifies a structural transition at around 34°C.

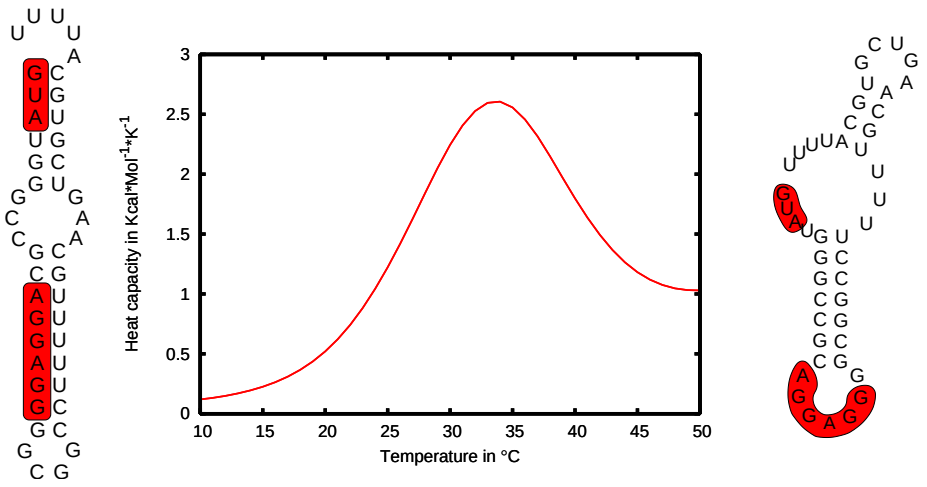

Figure 1: **RNA thermometer** Specific heat of a thermo-sensitive RNA switch. The peak at around 34°C marks the structural transition between the low temperature (left) and high temperature (right) structures. The Shine-Dalgarno sequence and the start codon (highlighted in red) are inaccessible at the low temperature structure, but accessible at high temperatures.

To model switches responsive to small RNA molecules it is sufficient to use methods able to predict RNA-RNA interactions, such as `RNAcofold` Bernhart et al. (2006) or `RNAup` Mückstein et al. (2008), given one assumes high concentrations of the small RNA. Ligand binding riboswitches are more difficult to analyze, as the mechanism for binding such small molecules is not captured by standard RNA energy models. However, experimentalists have measured binding free energies of ligands interacting with particular RNA motifs, e.g. Jenison et al. (1994); Jucker et al. (2003); Gouda et al. (2003) for the theophylline aptamer. These free energies can be included in energy landscape predictions as a energy correction on binding competent structures.

Section 3 describes fast approaches for thermodynamic riboswitches. For switches that are trapped in a kinetically favored structure, we describe methods based on RNA landscape computations in section 4. The most complex case probably are switches depend on co-transcriptional folding, as they have a

dynamic energy landscape that changes with every newly transcribed nucleotide. Section 5 shows how a co-transcriptional theophylline-riboswitch Wachsmuth et al. (2013) can be modeled by computing RNA folding kinetics on such dynamic energy landscapes.

3. Thermodynamic RNA folding

If we define the set of RNA secondary structures Ω as all structures that (i) are formed from nested, isosteric base-pairs (GC, CG, AU, UA, GU, UG), (ii) have hairpins with at least 3 unpaired nucleotides, and (iii) have interior-loops of at most 30 unpaired nucleotides, then this includes the vast majority of known pseudo-knot free secondary structures. There are experimentally determined energy parameters Mathews et al. (1999, 2004) that enable to compute a free energy for any RNA secondary structure $S \in \Omega$. In particular, these energy parameters assign energies to every loop (hairpin, interior, exterior and multi-loop). They mostly depend on the loop-type and size, with some sequence dependence. Most beneficial are stacking energies, i.e. base-pairs that close an interior loop with no unpaired bases in between, but there are also tabulated energy values, e.g. for common interior loops that are known for stable non-canonical interactions. This energy model is known as the Nearest Neighbor energy model Turner & Mathews (2009). The total energy of an RNA structure can be computed as the sum of all loops

$$E(S) = \sum_{L \in S} E(L) \tag{1}$$

3.1. RNA structure prediction

Based on the described energy model several methods exist to efficiently predict minimum free energy (MFE) structure as well as various equilibrium properties of the RNA. These methods solve the problem by dynamic programming and typically require $\mathcal{O}(n^2)$ space and $\mathcal{O}(n^3)$ time. They can thus be used routinely even for very long RNA molecules. In this contribution, we will focus on methods available in the `ViennaRNA package` Lorenz et al. (2011), which provides an

especially large selection of prediction methods. Other popular methods include e.g., `RNAstructure` Mathews (2014) and `mfold / UNAfold` Markham & Zuker (2008).

The most commonly mode of structure prediction, where programs, such as `RNAfold` will return a single structure corresponding to the lowest free energy state of the RNA. Since riboswitch function depends on the presence of at least two functional conformations, MFE folding is clearly insufficient.

A more complete picture of the thermodynamic folding can be gained by computing the partition function Z of an RNA molecule. From the partition function

$$Z = \sum_{S \in \Omega} e^{\frac{-E(S)}{RT}} \tag{2}$$

various equilibrium properties can be derived. In particular, we can compute the probability P of observing a structure S

$$P(S) = \frac{1}{Z} e^{\frac{-E(S)}{RT}} \tag{3}$$

and the ensemble free energy G

$$G = -RT \ln(Z) \tag{4}$$

The partition function can be computed with the same $\mathcal{O}(n^3)$ effort as computing the MFE structure. In addition, the algorithm allows to compute the equilibrium probability p_{ij} for every possible base-pair (i, j). Pair probabilities provide a compact representation of the complete Boltzmann ensemble of structures of an RNA molecule.

Most folding programs allow to specify constraints, such as base-pairs that have to be present or positions that are not allowed to pair. For a riboswitch with a known aptamer structure, this can be used to compute the partition function only over those structures which form the aptamer, i.e. binding competent structures. The ratio of the constrained and unconstrained partition function yields the equilibrium probability that the aptamer structure is formed

$$p(\text{aptamer}) = \frac{Z^{\text{constrained}}}{Z}. \tag{5}$$

If we have information on how strong the aptamer structure is stabilized by ligand-binding, e.g. from measurements of the dissociation constant K_d, and about ligand and RNA concentrations, we can even compute the fraction of ligand-bound RNAs as a function of the concentrations.

Another approach to gain a more complete picture than only a single MFE structure, is to compute suboptimal structures in addition to the MFE structure. At least three commonly used strategies for this exist. `mfold` Zuker (1989) first introduced an algorithm to compute all suboptimals detectable by picking one base-pair and asking for the optimal structure containing this pair. This approach yields a small, but generally incomplete list of alternative structures. `RNAsubopt` Wuchty et al. (1999) will produce all suboptimal structures in a defined energy range, resulting in a number of structures that grows exponentially with sequence length. Finally, it is possible to directly sample structures from the Boltzmann ensemble after computing the partition function Z.

Since riboswitches possess at least two functionally important conformations, it seems natural to use the prediction of suboptimal structures to search for novel riboswitches. One of the first methods to attempt this was `paRNAss` Giegerich et al. (1999); Voss et al. (2004). This program generates a sample of suboptimal structures, computes pairwise distances between those structures using two different distance measures and performs a clustering. RNAs which exhibit two well separated clusters of structures are classified as RNA switches. This procedure works well for a number of known switching RNA molecules, such as attenuator sequences, but is not too successful for ligand binding aptamers. The reason simply is that the aptamer binding conformation typically is only stable in the presence of the ligand. Since the structure predictions do not take ligand binding into account, they fail to recognize the aptamer conformation as a low energy state. In practice, computational efforts for riboswitch discovery have therefore focused on the detection of known aptamer structures using structural homology search.

3.2. RNA2Dfold

The `paRNAss` method, mentioned above, introduced a so called validation plot as visualization of the clustering result. Once the procedure has identified two clusters and their representative structure, it computes for every suboptimal structure the distances d_1, d_2 to these two reference structures. The resulting distance pairs plotted as points in a 2D coordinate system.

The idea of classifying each structure by its distance to two reference structures is pursued in a more principled way in `RNA2Dfold` Lorenz et al. (2009). Rather than working with a sample of suboptimal structures, `RNA2Dfold` considers all possible secondary structures and performs a classified dynamic programming. In short, we define a distance class (κ, λ) to comprise all structures with distance κ to the first reference structure and λ to the second. An extension of classical RNA folding algorithms then computes the MFE structure (or partition function) for every distance class. In effect, `RNA2Dfold` computes a projection of the high-dimensional conformation space into two dimensions spanned by the distance to the reference structures. The result is ideal for visualizing the folding landscape by plotting the folding energy as a function of κ and λ, see Figure 2.

The additional bookkeeping makes `RNA2Dfold` much more expensive than normal RNA folding, requiring $\mathcal{O}(n^7)$ time and $\mathcal{O}(n^4)$ space. Nevertheless, the approach is readily applicable to sequences of up to about 400nt, easily exceeding the length of typical riboswitches. A remaining problem is to best choose the two reference structures. In general we want to choose the MFE structure as the first and a meta-stable structure as the second reference. A common work flow is to first perform a run of `RNA2Dfold` with the MFE and the open chain conformation as references, and choose a suitable meta-stable structure from the results of this first run. The meta-stable structure is then used as reference in a subsequent second `RNA2Dfold` computation.

This procedure is illustrated in Figure 2 using the leader sequence of the *E. coli* tRNA^{phe} synthetase operon. The ground state structure of the sequence forms a terminator hairpin that switches the transcription of the downstream

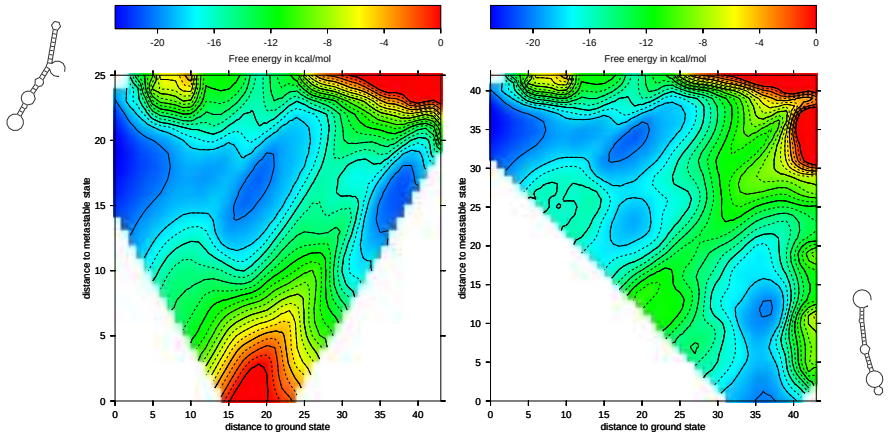

Figure 2: `RNA2Dfold` computed projection of the energy landscape for the leader sequence of the *E. coli* tRNA^{phe} synthetase operon. Left panel: Projection using the MFE structure (terminator hairpin, far left) and the open chain as references. Right panel: Projection using the MFE structure and the meta-stable structure found at position (36,17) of the first projection. The two structures are shown on the left and right of the landscapes.

operon off. The 2D landscape clearly indicates the presence of a meta-stable state with 17 base-pairs and a base-pair distance of 36 to the MFE structure. Using the meta-stable structure as the second reference shows the clearly separated conformational states of the leader sequence even better.

It is worth noting that 2D landscapes immediately provide a lower bound on the energy barrier between the two structures and thus an estimate how quickly the RNA can switch conformations. They can also be used as a starting point for more sophisticated path-finding heuristics. By computing a series of 2D landscapes for successively longer sequences, one can obtain a qualitative impression of co-transcriptional folding in order to study kinetic switches.

The caveat about the effect of ligand binding applies here as well, since we only obtain a landscapes for the unbound riboswitch. An upcoming version of the Vienna RNA package will allow to specify flexible soft constraints, such as energy bonuses for particular structural motifs. Given suitable experimental binding energies this should allow us to compute 2D landscapes for the

riboswitch in the presence of the ligand.

3.3. RNAsubopt, barriers

`RNA2Dfold` as described above will generally not find all local minima, i.e. meta-stable states, of an RNA. For a more complete characterization of riboswitches we need to consider the whole energy landscape $\mathcal{L}$ of an RNA molecule and identify all stable alternative conformations. In general, RNA molecules can adopt multiple conformations and also non-riboswitches might have alternative structures that are kinetically favored. Moreover, the lifespan of an RNA molecule can be simply too short to reach the MFE structure at all. Whether this is the case for a particular RNA can be determined by analyzing the energy landscape.

More formally, denote the energy landscape as $\mathcal{L} = (\Omega, \mathcal{M}, E)$, with Ω being the previously introduced set of RNA conformations, $\mathcal{M}$ being a move-set to define a neighborhood relation and E being an energy function to assign a fitness value to each conformation. For an ergodic move-set $\mathcal{M}$ we chose the most elementary modification of an RNA secondary structure, the formation or opening of a single base-pair.

`RNAsubopt` Wuchty et al. (1999) computes all conformations $S \in \Omega$ that are within a certain energy range above the MFE. Since RNA energy landscapes grow exponentially with sequence length, this results in a massive amount of secondary structures even for very short sequences. The program `barriers` Flamm et al. (2002) can then process such an energetically sorted list of suboptimal structures with a flooding algorithm to find all local minima and the according saddle points connecting them. In particular, every structure is either a local minimum, a saddle point connecting at least two local minima, or it belongs to the *basin* of one local minimum. This allows for computing the partition functions (see equation 2) for every basin. The level of coarse-graining can be adjusted to the inspected landscape by specifying the minimal depth of a local minimum or the total number of energetically best local minima. The results can be visualized in form of a barrier-tree.

For temperature sensitive RNA switches, as shown in Figure 1, one can compute the suboptimal structures for two temperatures and compare the energy landscapes. See Figure 3 for barrier trees depicting the landscapes at temperatures 30°C and 40°C.

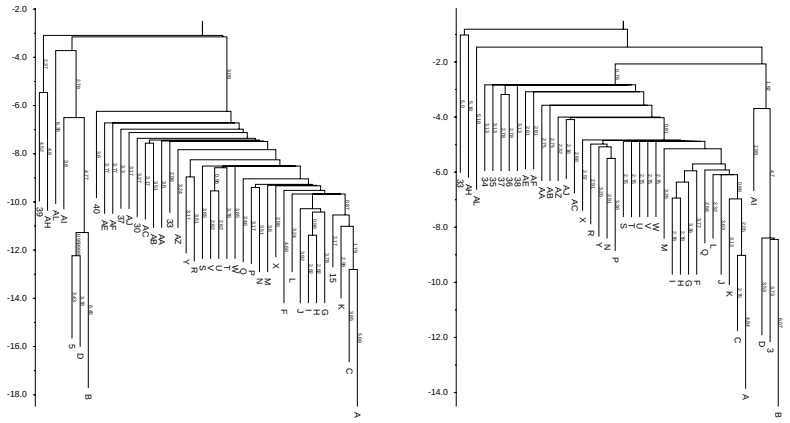

Figure 3: **Barrier tree representation of the folding landscape.** Barrier trees were computed for the RNA thermometer shown in Fig. 1) at 30°C (left) and 40°C (right). Local minima occurring in both landscapes are labeled by letters, minima labeled by numbers occur in only on of the two landscapes. The ordinate gives the energy of local minima and saddle points in `kcal/mol`. The two prominent structures A and B change place in response to the temperature change, while the overall shape of the tree stays qualitatively the same.

As mentioned previously, modeling ligand binding riboswitches requires to take into account the stabilizing effect of the bound ligand to the aptamer structure. For some aptamers binding affinities and thus binding free energies have been experimentally determined and in addition the structural requirements for ligand binding are often known. The energy landscape of the riboswitch in presence of the ligand can then be analyzed by adding the binding free energy to all conformations that are binding competent, i.e. contain an intact aptamer structure.

In the following we use an artificially designed theophylline-dependent riboswitch termed RS10 Wachsmuth et al. (2013). It is positioned at the 5'UTR of its target gene (*bgaB*) and leads to the formation of an early terminator

hairpin in the absence of theophylline. As soon as theophylline is present, a co-transcriptionally formed aptamer structure is stabilized, the terminator cannot be formed and the mRNA is transcribed in its full length. The binding energy of theophylline to the aptamer was estimated from the dissociation constant of $K_d = 0.32\mu M$ at 25°C Jenison et al. (1994) as $\Delta G = -\mathrm{RT} \ln K_d = -8.86$kcal/mol. Since RS10 regulates at the transcription level, it necessarily falls into the category of kinetic switches. The terminator hairpin can be effective only if it forms quickly enough, i.e. before the polymerase has continued into the coding region. Transcription speed and therefore the choice of nucleotides and the length of the spacer region play a crucial role for a proper functionality.

Figure 4 shows two barrier trees of the RS10 riboswitch, representing the energy landscape with and without the ligand. Structures A and B contain the terminator hairpin, while structure I does not and therefore represents the on-state. Note that even in the presence of theophylline, structure A remains the ground state. The terminator free structure I is only meta-stable, but separated by an energy barrier of $\approx$ 12kcal/mol from the ground state. The static landscape picture is, however, insufficient to decide whether the on-state structure I will indeed be reached by the co-transcriptional folding process.

The limitation of the `RNAsubopt/barriers` approach lies in the lengths of inspected molecules. The `barriers` program has to read and store all low energy structures in memory. This limits the approach to RNA molecules where the relevant low energy part of the landscape comprises less than, say, 10^8 conformations which is often reached by molecules of about 100nt. For longer RNAs it might be still possible to identify the most important local minima, but not the saddle points connecting them.

Recent work has aimed to overcome the length limits of `barriers` by using heuristics for sampling low lying local minima as well as for estimating barrier heights between local minima. The Basin hopping graph approach of Kuchark et al. (2014), for example, can handle RNAs of several 100nt at the expense of loosing exact barrier heights.

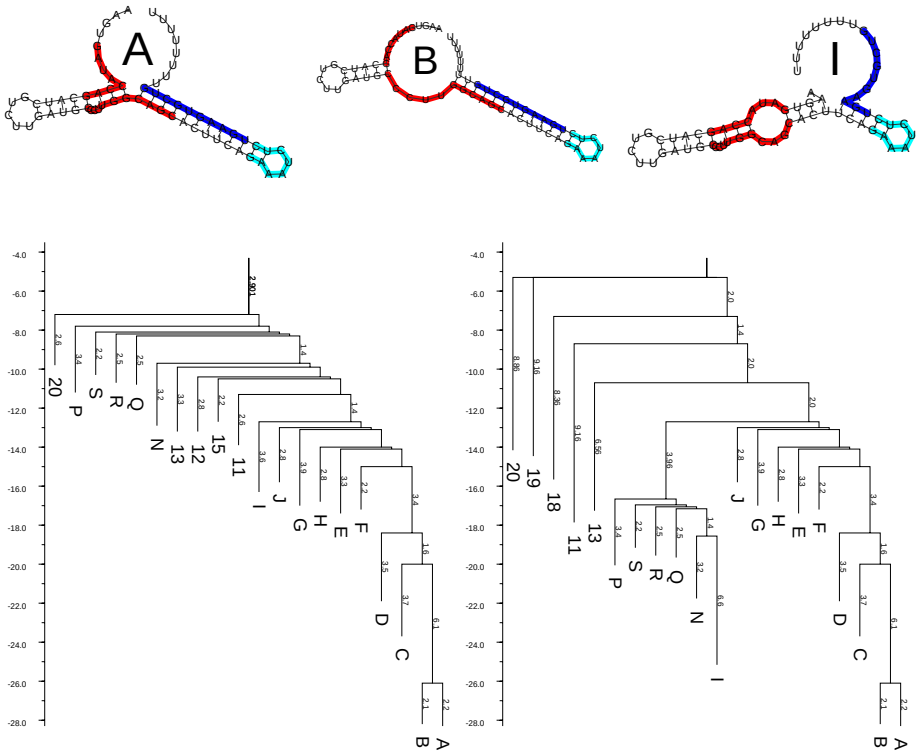

Figure 4: **Folding Landscape Picture:** (top): The three main structures of the RS10 riboswitch, using the color code as in Wachsmuth et al. (2013): **red** indicates the aptamer region, **blue** is the 3'-part of the terminator hairpin and **cyan** is a spacer region. Structures A and B contain the terminator hairpin and therefore correspond to the off-state of the switch. Structure I is the theophylline binding competent structure corresponding to the on-state of the switch. This structure is stabilized by -8.86 `kcal/mol` upon ligand binding. The barrier tree on the left and the right correspond to the undistorted (theophylline free) and distorted folding landscape, respectively. Note the gain in ruggedness for the distorted folding landscape and the emergence of a distinct sub-tree containing structure I.

4. RNA folding kinetics on static landscapes

The physical process of RNA folding is usually modeled as a stochastic process on an RNA energy landscape specifying (i) the state space, (ii) the neighborhood relation and (iii) the transition rate model. With these three key con-

cepts at hand the folding process can be described as a continuous time Markov process

$$\frac{dP_x(t)}{dt} = \sum_{y \neq x} [P_y(t) k_{xy} - P_x(t) k_{yx}]$$

where $P_x(t)$ gives the probability to observe the folding RNA chain in conformation x at time t, and k_{xy} is the transition rate from conformation y to conformation x. Clearly, $k_{xy} > 0$ only if conformation x is reachable from conformation y via the neighborhood relation. Most existing approaches for kinetic RNA folding are based on the master equation model above and mainly differ in the set of allowed states (e.g. with or without pseudo-knots), the neighborhood relation, as well as in the energy rules and the resulting rate model. However, the existing approaches can be partitioned into two major classes according to the method how the master equation is solved. The first class of approaches apply Gillespie-type simulation algorithms Gillespie (1977) to generate statistically correct trajectories as possible solutions. The second class of approaches solve the master equation directly.

4.1. Stochastic simulation of folding kinetics

The program `Kinfold` Flamm et al. (2000) implements a rejection-less Monte-Carlo method together with the most elementary neighborhood relation, the insertion or deletion of a single base-pair. While this combination allows for a very detailed simulation of folding pathways, the elementary step resolution leads necessarily to long simulation runs. Many approaches therefore choose to allow larger structural changes by using the formation or destruction of an entire helix as the basic step Mironov & Lebedev (1993); Isambert & Siggia (2000); Danilova et al. (2006); Huang & Voß (2014). Using helix insertion/deletion as basic transformation strongly restricts the space of allowed conformations. This reduction allows to explore the conformation space in a much smaller number of steps. Consequently simulation of larger RNAs become feasible. However, due to the larger structural changes during a simulation step, the quality of the rate model becomes extremely important. The extension of these approaches to

gain folding during transcription, or the incorporation of pseudo-knotted structures is straight forward. For a recent review on the advantages and problems of kinetic folding approaches see Flamm & Hofacker (2008).

On the downside, simulation approaches require a fairly large number of trajectories in order to give statistically robust results. In general, they also require sophisticated post processing in order to interpret the trajectories in a meaningful way.

4.2. Barriers/Treekin

Formally, the master equation 4 is solved by

$$P(t) = e^{t \cdot \mathbf{K}} \cdot P(0)$$

where $P(0)$ is the vector of initially populated conformations for $t = 0$, and $\mathbf{K} = (k_{xy})$ is the matrix of transition rates between individual conformations of the conformation space. Integrating the master equation thus involves computing matrix exponentials, usually by first diagonalizing the matrix $\mathbf{K}$. This limits the dimension of the number of $\mathbf{K}$ to a few thousand. Since the number of conformations grows exponentially with sequence length, the solution 4.2 is applicable only for short toy examples.

In order to treat RNAs of biological interest, we need a coarse-graining that reduces the number of conformations. The program `barriers` Flamm et al. (2002) performs such a coarse-graining of the conformation space into macro-states, by partitioning the folding landscape into gradient basins and their connecting saddle points. The resulting hierarchical structure, called barrier tree (see figure 3), offers a compact representation of the entire folding landscape, where leaf nodes of the tree correspond to local minima and internal tree nodes to the energetically lowest saddle points connecting two local minima. During the construction of the barrier tree, the program `barriers` identifies these "gradient basins" and calculates the partition function of each macro-state as well as effective transition rates between any two macro-states α, β as

$$k(\alpha \to \beta) \approx \sum_{x \in \alpha} \sum_{y \in \beta} k(x \to y) e^{-E(x)/RT} / Z_\alpha.$$

The approximation assumes a local equilibrium between the conformations within each macro-state such that the partition function Z_α can be used to calculate the probability of being in conformation x in macro-state α. The Metropolis rule is used to assign the micro-state transition probabilities $k(x \rightarrow y)$. The macro-state transition matrix and a vector of initial populations is then handed to the program `treekin` Wolfinger et al. (2004), which numerically integrates the master equation for arbitrary long times t by computing the matrix exponential. The time evolution of the population density is returned as a result (see Figure 5). The folding dynamics of RNA molecules up to the size of tRNAs can therefore easily be computed for arbitrary long time scales using the `barriers/treekin` approach.

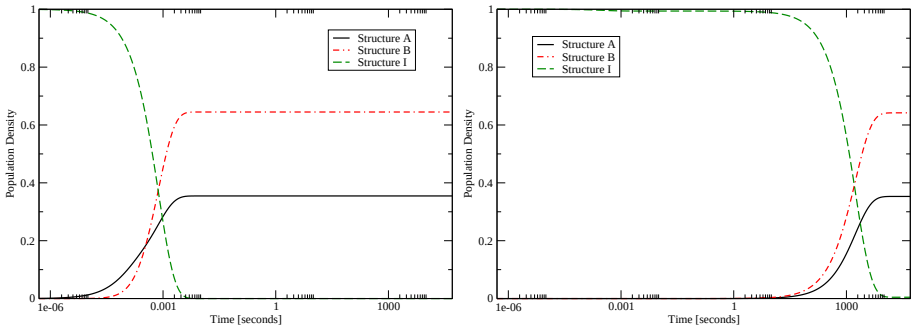

Figure 5: `treekin` simulation of the unbound RS10 riboswitch (left) and the theophylline bound switch (right). We set the start conditions of the simulation to structure I being the only population. While the refolding to the MFE structure and a close neighbor happens really fast in the unbound condition, the molecule is trapped for a long time ($> 10^3$ seconds) in the ligand bound state.

For illustration, we again use the RS10 riboswitch and compute folding kinetics starting at the aptamer conformation on either the undisturbed energy landscape or the landscape corrected for ligand binding energies. We predict that the unbound RS10 riboswitch re-folds to the off-state in about one hundredth of a second, while the theophylline bound off-state remains stable for more than 15 minutes, see Figure 5.

A general problem with such simulations is that the computation uses an in-

ternal time-scale whose relation to actual wall clock time is unknown. Recently, Sauerwine & Widom (2013) performed `Kinfold` simulations on short RNAs and compared `Kinfold` time to re-folding times determined from NMR experiments and determined that 1 `Kinfold` time-step corresponds to roughly $5\mu s$. We performed `Kinfold` and `treekin` simulations on the same molecule in order to verify that `treekin` and `Kinfold` time units are approximately equivalent and used this number to convert simulation time to seconds.

We note, that the computation of barrier trees and `treekin` trajectories can be performed using the Vienna RNA web services Gruber et al. (2008) at `http://www.tbi.univie.ac.at/`. The web version, however, does not support the inclusion of ligand binding energies.

5. RNA folding kinetics on dynamic landscapes

In a cellular context the nascent RNA molecule starts folding before the transcription process is completed Lai et al. (2013) and the folded structure may therefore depend on the speed of elongation, on site-specific pausing of the RNA polymerase Wong et al. (2007), and interactions of the nascent RNA molecule with proteins or small-molecule metabolites Pan & Sosnick (2006). Many riboswitches are thought to co-transcriptionally fold into their on- or off-state, depending on the presence of their trigger, and will then stay trapped in that conformation even if the trigger is removed.

The hybrid-simulation framework `BarMap` Hofacker et al. (2010) enables to study the interplay between the kinetic folding process and time dependent changes of the folding landscape. The main idea is to compute a mapping between macro-states of successive folding landscapes and use this information to determine the initial population densities for successive kinetic simulations. In the case of co-transcriptional folding, an energy landscape for each RNA elongation step (adding a single nucleotide) is computed using `barriers`.

`BarMap` then constructs a mapping between the energy landscapes $\mathcal{L}_n \rightarrow \mathcal{L}_{n+1}$. Since a newly transcribed nucleotide cannot initially interact with the

previously transcribed part, every minimum in $\mathcal{L}_n$ is appended by an unpaired base. In the easiest case, this new structure is a minimum in landscape $\mathcal{L}_{n+1}$, then it can be directly mapped. Alternatively, a heuristic is used to compute the next best local minimum conformation which, if still not found in $\mathcal{L}_{n+1}$, is mapped to the state with the least base-pair distance. The three possible cases that result from this mapping are illustrated in Figure 6.

Folding kinetics can now be simulated using `treekin` starting with the first landscape that has more than one macro-state. The amount of time should correspond to the elongation time of the polymerase. The distribution of populated minima after the simulation is then transfered to the successive landscape according to the mapping computed by `BarMap`. Again, a folding simulation is performed starting from these conditions. This interleaving sequence of kinetic folding and transfer of the population density to the successive landscape is done until the folding landscape of the full length sequence is reached. The amount of time the folding chain spends on a particular landscape (in the series) before being re-mapped allows to handle any type of coupling between the dynamics of the folding chain and the dynamics of the changing landscape.

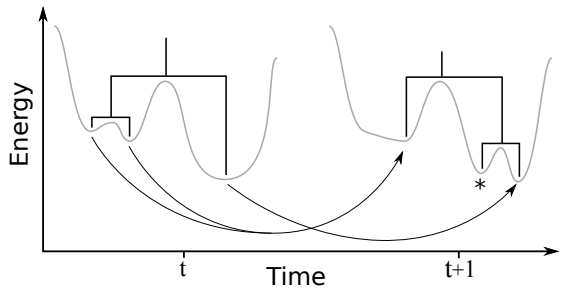

Figure 6: Schematic representation of the mapping process between two consecutive landscapes at time t and $t+1$. Three types of events need to be distinguished: (i) A simple one-to-one correspondence between two local minima (right), (ii) two minima are merged into one (left) and (iii) a new minimum appears in $t+1$ (*).

The RS10 riboswitch introduced above is a good example for a system whose function can only be understood in view of its co-transcriptional folding behav-

ior. A terminator hairpin can only be effective if it is formed almost immediately after the transcript reaches the poly-U tract adjacent to the hairpin. The interplay between the height of energy barriers and the speed of transcription is therefore crucial for riboswitch function. This fact also makes the design of such switches especially challenging. Wachsmuth et al. (2013) thus designed a series of candidates among which RS10 was the most effective.

Following Bremer & Dennis (1996), the transcription rate of *E. coli* polymerase is around 50 nt/second. Hence, we used the `BarMap` framework with an elongation time of 4000 in `treekin` units to compute co-transcriptional folding dynamics of RS10 in presence / absence of theophylline. The resulting population density for the different conformational states as a function of time can be seen in Figure 7.

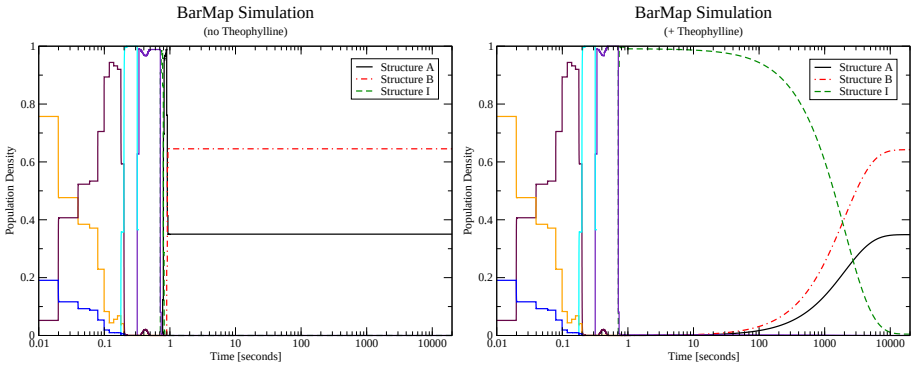

Figure 7: `BarMap` simulation of co-transcriptional folding with a transcription speed of approximately 50nt per second. Structures A, B and I correspond to the ones shown in Figure 4. Both simulations start after transcribing the first 16nt. The simulation on the landscape without theophylline reaches the equilibrium (with the riboswitch in the off state) as soon as the last nucleotide is added. With theophylline, almost 100% of the RNA is in the on-state (structure I) at the end of the elongation period. The molecule then needs on the order of 1000 seconds to re-fold into the equilibrium off-state.

6. Conclusion

The well-known RNA structure prediction methods assume an RNA in thermodynamic equilibrium and are therefore of limited use for studying the conformational switching at the heart of riboswitch function. A number of approaches exist that characterize the folding landscape and the resulting dynamics of RNA molecules. While these are well suited to study mechanisms of riboswitch function, they are both computationally more demanding as well as more challenging for the user.

In particular, many riboswitches can only be understood in the context of co-transcriptional folding. The recently developed `BarMap` approach treats co-transcriptional folding as a process on a time-varying landscape. Using the example of a recently designed theophylline riboswitch, we show that this approach predicts riboswitch behavior in good agreement with experimental observations.

Several limitations remain in the computational approaches. (i) The energetics of RNA-ligand interactions cannot be predicted within the secondary structure model, although binding energies from experiments can be incorporated. (ii) In general, the structural prerequisites for ligand binding are not precisely known, making it difficult to judge which conformations along a folding pathway are binding competent. (iii) Secondary structure prediction ignores pseudo-knots and tertiary interactions which can be essential for aptamer function. (iv) The most accurate computational methods are expensive and limited to moderate sequence lengths.

Nevertheless, the examples presented here illustrate that the secondary structure model captures enough detail of the molecular mechanism to provide a realistic picture of riboswitch function.

7. Acknowledgments

This work was supported in part by the FWF International Programme I670, the DK RNA program FG748004, the EU-FET grant RiboNets 323987,

and the COST Action CM1304 “Emergence and Evolution of Complex Chemical Systems”.

References

Bernhart, S., Tafer, H., Mückstein, U., Flamm, C., Stadler, P., & Hofacker, I. (2006). Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, *1*, 3. doi:10.1186/1748-7188-1-3.

Bremer, H., & Dennis, P. P. (1996). Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella: cellular and molecular biology*, *2*, 1553–1569.

Danilova, L. V., D., P. D., Favorov, A. V., & Mironov, A. A. (2006). RNAKinetics: A web server that models secondary structure kinetics of an elongating RNA. *J. Bioinform. Comput. Biol.*, *4*, 589–596.

Dawid, A., Cayrol, B., & Isambert, H. (2009). RNA synthetic biology inspired from bacteria: construction of transcription attenuators under antisense regulation. *Physical Biology*, *6*, 025007. URL: http://iopscience.iop.org/1478-3975/6/2/025007. doi:10.1088/1478-3975/6/2/025007.

Flamm, C., Fontana, W., Hofacker, I. L., & Schuster, P. (2000). RNA folding at elementary step resolution. *RNA*, *6*, 325–338.

Flamm, C., & Hofacker, I. L. (2008). Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatshefte für Chemie*, *139*, 447–457. doi:doi:10.1007/s00706-008-0895-3.

Flamm, C., Hofacker, I. L., Stadler, P. F., & Wolfinger, M. T. (2002). Barrier trees of degenerate landscapes. *Z Phys Chem*, *216*, 155–173. doi:10.1524/zpch.2002.216.2.155.

Giegerich, R., Haase, D., & Rehmsmeier, M. (1999). Prediction and visualization of structural switches in rna. *Pac Symp Biocomput*, (pp. 126–37).

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, *81*, 2340–2361.

Gouda, H., Kuntz, I. D., Case, D. A., & Kollman, P. A. (2003). Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers*, *68*, 16–34. URL: http://onlinelibrary.wiley.com/doi/10.1002/bip.10270/abstract. doi:10.1002/bip.10270.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., & Hofacker, I. L. (2008). The Vienna RNA websuite. *Nucl. Acids Res.*, *36*, W70–W74. doi:10.1093/nar/gkn188.

Hofacker, I. L., Flamm, C., Heine, C., Wolfinger, M. T., Scheuermann, G., & Stadler, P. F. (2010). BarMap: RNA folding on dynamic energy landscapes. *RNA*, *16*, 1308–1316. doi:doi:10.1261/rna.2093310.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures (the Vienna RNA Package). *Monatsh. Chem.*, *125*, 167–188. doi:10.1007/BF00818163.

Huang, J., & Voß, B. (2014). Analysing RNA-kinetics based on folding space abstraction. *BMC Bioinformatics*, *15*, 60. doi:doi:10.1186/1471-2105-15-60.

Isaacs, F. J., Dwyer, D. J., Ding, C., Pervouchine, D. D., Cantor, C. R., & Collins, J. J. (2004). Engineered riboregulators enable post-transcriptional control of gene expression. *Nature Biotechnology*, *22*, 841–847. URL: http://www.nature.com/nbt/journal/v22/n7/abs/nbt986.html. doi:10.1038/nbt986.

Isambert, H., & Siggia, E. D. (2000). Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. USA*, *97*, 6515–6520.

Jenison, R. D., Gill, S. C., Pardi, A., & Polisky, B. (1994). High-resolution molecular discrimination by RNA. *Science*, *263*, 1425–1429. URL: http://www.sciencemag.org/content/263/5152/1425. doi:10.1126/science.7510417.

Jucker, F. M., Phillips, R. M., McCallum, S. A., & Pardi, A. (2003). Role of a heterogeneous free state in the formation of a specific RNA-theophylline complex. *Biochemistry*, *42*, 2560–2567.

Kuchark, M., Hofacker, I. L., Stadler, P. F., & Qin, J. (2014). Basin hopping graph: a computational framework to characterize RNA folding landscapes. *Bioinformatics*, *30*, 2009–17. doi:10.1093/bioinformatics/btu156.

Lai, D., Proctor, J. R., & Meyer, I. M. (2013). On the importance of cotranscriptional RNA structure formation. *RNA*, *19*, 1461–1473. doi:10.1261/rna.037390.112.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). Viennarna package 2.0. *Algorithms Mol Biol*, *6*, 26. doi:10.1186/1748-7188-6-260.

Lorenz, R., Flamm, C., & Hofacker, I. L. (2009). 2D projections of RNA folding landscapes. In I. Grosse, S. Neumann, S. Posch, F. Schreiber, & P. Stadler (Eds.), *German Conference on Bioinformatics 2009* (pp. 11–20). Bonn: Gesellschaft f. Informatik volume 157 of *Lecture Notes in Informatics*.

Markham, N. R., & Zuker, M. (2008). Unafold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, *453*, 3–31. doi:10.1007/978-1-60327-429-6_1.

Mathews, D. H. (2014). Rna secondary structure analysis using rnastructure. *Curr Protoc Bioinformatics*, *46*, 12.6.1–12.6.25. doi:10.1002/0471250953.bi1206s46.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., & Turner, D. H. (2004). Incorporating chemical modification constraints into

a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 7287–7292.

Mathews, D. H., Sabina, J., Zuker, M., & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology*, *288*, 911–940.

Mironov, A. A., & Lebedev, V. F. (1993). A kinetic modle of RNA folding. *BioSystems*, *30*, 49–56.

Mückstein, U., Tafer, H., Bernhart, S. H., Hernandez-Rosales, M., Vogel, J., Stadler, P. F., & Hofacker, I. L. (2008). Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. In M. Elloumi, J. Küng, M. Linial, R. Murphy, K. Schneider, & C. Toma (Eds.), *Bioinformatics Research and Development* (pp. 114–127). Springer volume 13 of *Communications in Computer and Information Science*. doi:10.1007/978-3-540-70600-7_9.

Nudler, E., & Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. *TRENDS Biochem. Sci.*, *29*, 11–17. doi:10.1016/j.tibs.2003.11.004.

Pan, T., & Sosnick, T. (2006). RNA folding during transcription. *Annu. Rev. Biophys. Biomol. Struc.*, *35*, 161–175. doi:10.1146/annurev.biophys.35.040405.102053.

Qi, L., Lucks, J. B., Liu, C. C., Mutalik, V. K., & Arkin, A. P. (2012). Engineering naturally occurring trans-acting non-coding RNAs to sense molecular signals. *Nucleic Acids Research*, *40*, 5775–5786. URL: http://nar.oxfordjournals.org/content/40/12/5775. doi:10.1093/nar/gks168.

Rodrigo, G., Landrain, T. E., Majer, E., Dars, J.-A., & Jaramillo, A. (2013). Full design automation of multi-state RNA devices to program

gene expression using energy-based optimization. *PLoS Comput Biol*, *9*, e1003172. URL: http://dx.doi.org/10.1371/journal.pcbi.1003172. doi:10.1371/journal.pcbi.1003172.

Sauerwine, B., & Widom, M. (2013). Folding kinetics of riboswitch transcriptional terminators and sequesterers. *Entropy*, *15*, 3088–3099.

Serganov, A., & Nudler, E. (2013). A decade of riboswitches. *Cell*, *152*, 17–24. doi:10.1016/j.cell.2012.12.024.

Turner, D. H., & Mathews, D. H. (2009). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, (p. gkp892).

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

Wachsmuth, M., Findeiß, S., Weissheimer, N., Stadler, P. F., & Mörl, M. (2013). De novo design of a synthetic riboswitch that regulates transcription termination. *Nucleic acids research*, *41*, 2541–2551.

Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L., & Stadler, P. F. (2004). Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.*, *37*, 4731–4741. doi:10.1088/0305-4470/37/17/005.

Wong, T. N., Sosnick, T. R., & Pan, T. (2007). Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proc. Natl. Acad. Sci. USA*, *104*, 17995–18000. doi:doi:10.1073/pnas.0705038104.

Wuchty, S., Fontana, W., Hofacker, I. L., & Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, *49*, 145–165.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, *244*, 48–52.