BarMap: RNA Folding on Dynamic Energy Landscapes

Ivo L. Hofacker,[*][a] Christoph Flamm[a] Christian Heine[c]
Michael T. Wolfinger[a,b] Gerik Scheuermann[c]
Peter F. Stadler[d,e,f,a,g]

[a]*Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria* {`ivo,xtof,mtw`}`@tbi.univie.ac.at`

[b]*incore IT Solutions OG, Grundlgasse 3/1, 1090 Wien, Austria*

[c]*Signal Processesing and Visualization Group, Department of Computer Science, University of Leipzig, Johannisgasse 26, D-04103 Leipzig, Germany* {`heine,scheuermann`}`@informatik.uni-leipzig.de`

[d]*Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany* `studla@bioinf.uni-leipzig.de`

[e]*Max-Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany*

[f]*Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany*

[g]*Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

Abstract

Dynamical changes of RNA secondary structures play an important role in the function of many regulatory RNAs. Such kinetic effects, especially in time-variable and externally triggered systems, are usually investigated by means of extensive and expensive simulations of large sets of individual folding trajectories. Here we described the theoretical foundations of a generic approach that not only allows the direct computation of approximate population densities but also reduces the efforts required to analyse the folding energy landscapes to a one-time preprocessing step. The basic idea is to consider the kinetics on individual landscapes and to model external triggers and environmental changes as small but discrete changes in the landscapes. A "barmap" links macrostates of temporally adjacent landscapes and defines the transfer of population densities from one "snapshot" to the next. Implemented in the `BarMap` software, this approach makes it feasible to study folding processes at the level of basins, saddle points, and barriers for many non-stationary scenarios, including temperature changes, co-transcriptional folding, re-folding in consequence to degradation, and mechanically constrained kinetics as in the case of pulling polymers through a pore.

---

1 Introduction

Dynamic changes of protein structure play an important role in their cellular functions. This include in particular the process of folding itself but also the structural response to oligomerization, chemical modification, ligand binding, and changes in ambient temperature or pH. The investigation of these phenomena plays a central role in protein science in both theory and experiment. Large-scale Molecular Dynamics (MD) simulations of (re)folding trajectories constitute the major computational approach in this area (Rapaport, 2004).

Detailed case studies have demonstrated that nature also exploits the potential of RNA sequences to form multiple alternative metastable structures. These play a role in particular in regulating gene expression at the level of the mRNA. One widespread mechanism is the attenuation of transcription found in many bacterial operons related to the bio-synthesis of amino acids (Henkin and Yanofsky, 2002; Gollnick et al., 2005). Another impressive example is the control of plasmid R1 maintenance in *E. coli*, reviewed in Gerdes and Wagner (2007). RNA thermometers (Narberhaus et al., 2006) are temperature responsive structural elements located in the 5'-untranslated region of bacterial heat shock and virulence genes. Mechanistically, RNA thermometers regulate the transcription of their respective genes by undergoing temperature-induced structure changes, a widely used regulatory strategy in nature (Klinkert and Narberhaus, 2009). It has been shown repeatedly, furthermore, that alternative conformations of the same RNA sequence can perform completely different functions (Baumstark et al., 1997; Perrotta and Been, 1998; Schultes and Bartel, 2000).

A thorough analysis of the dynamics of RNA folding and re-folding is thus a necessary prerequisite for a detailed understanding of the functionality of many RNA molecules. In contrast to protein folding, the secondary structures of nucleic acids provide a level of description that is sufficient to understand the thermodynamics and kinetics of RNA folding (D. Thirumalai and Klimov, 2001) — a least in a useful approximation. Initially, kinetic folding was used as an attempt to improve RNA structure prediction (Martinez, 1984; Mironov et al., 1985; Abrahams et al., 1990; Gultyaev, 1991; Tacker et al., 1994). More recently, the focus has shifted towards understanding the conformational changes and the associated folding pathways themselves, recently reviewed in Flamm and Hofacker (2008).

Most kinetic folding algorithms for RNA are some form of discretized Monte Carlo simulations of folding trajectories. The direct analysis of the folding energy landscape presents a viable alternative (Flamm et al., 2002), due to the fact that the lower part of an energy landscape can be accessed efficiently by dynamic programming (Wuchty et al., 1998; Clote, 2005). Here, one first constructs a compact representation of the energy landscape in the form of a hierarchical structure termed *barrier tree*. Recently, coarse grained landscapes have also been used in conjunction with stochastic sampling algorithms (Tang et al., 2008).

Barrier trees and related tree structures have been developed independently for different classes of disordered systems, including spin glasses (Klotz and Kobe, 1994), potential energy surfaces in protein folding (Becker and Karplus, 1997; Garstecki et al., 1999), molecular clusters (Wales et al., 1998; Doye et al., 1999), and RNA secondary structures (Flamm et al., 2000). Assuming that the basins of individual local minima are in quasi-equilibrium, the rates between all local minima can be calculated during barrier tree construction, providing an approximated master equation that can be solved explicitly (Wolfinger et al., 2004). This observation provides the starting point for the present contribution.

Often, one is most interested in the re-folding of an RNA in response to an external signal. Such a "signal" can be the binding of a ligand, a nucleolytic cleavage, the elongation of the RNA during transcription, a change of the environmental temperature, or some form of mechanical stress. We show here that all these scenarios can be treated within a single coherent framework, namely as (a series of) perturbations of the energy landscape on which the folding process operates. This observation will allow us to develop generic tools that allow the efficient evaluation of the re-folding kinetics by connecting the coarse-grained tree representations of perturbed landscapes in a suitable way. Before we proceed to three illustrative applications, we will develop the associated theory in detail in the following section.

2 Theory

2.1 Energy Landscapes for RNA Folding

The *energy landscape* of an RNA molecule is, for our purposes, defined on the set X_σ of all secondary structures that can be formed by the sequence σ in such a way that base pairs obey the usual base pairing rules. As usual, we disregard pseudoknots. It is well known that the size of the set X_σ grows exponentially with the chain length n, see e.g. Hofacker et al. (1998) and the
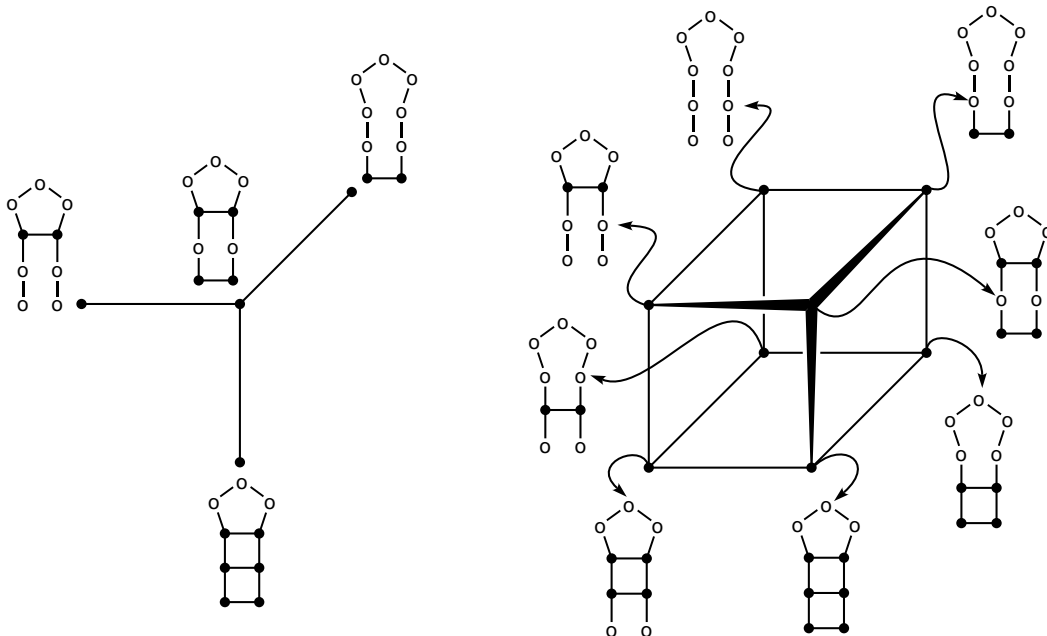
Fig. 1. Move sets for simulations of RNA folding kinetics at secondary structure level. Adjacent conformations differ by insertion or deletion of a single base pair, arranging the secondary structures in an undirected graph.

references therein. The Turner energy rules (Mathews et al., 1999) allow us to compute the energy $f(x)$ for each given secondary structure $x \in X_\sigma$.

This set of discrete conformations is arranged as a graph by defining a "move set", i.e., by specifying which pairs of secondary structures can be interconverted in a single step, see e.g. Reidys and Stadler (2002) and the references therein. Fig. 1 gives a simple example. In Flamm et al. (2000), two move sets are considered for RNA: the simpler case allows only the opening or closing of a single base pair, the more complex approach allows the sliding on one endpoint of a pair to a new pairing partner. In both cases neighboring structures differ by adding and/or removing a single base pair, hence the size D of the neighborhood of a conformation is at most quadratic in sequence length. This small size of the neighborhoods relative to the huge set of all conformations is crucial for the computational feasibility of our approach.

We remark in passing that lattice models of protein folding have the same formal properties (Wolfinger et al., 2006). The entire machinery described here for RNA folding kinetics can thus be applied also to this class of models.

The set of conformations, the move set, and the energy function together define the energy landscape of our molecule. Conceptually, this energy landscape is closely related to *potential energy surfaces* (Mezey, 1987; Heidrich et al., 1991), which describe the system at the level of spatial coordinates of individual atoms.
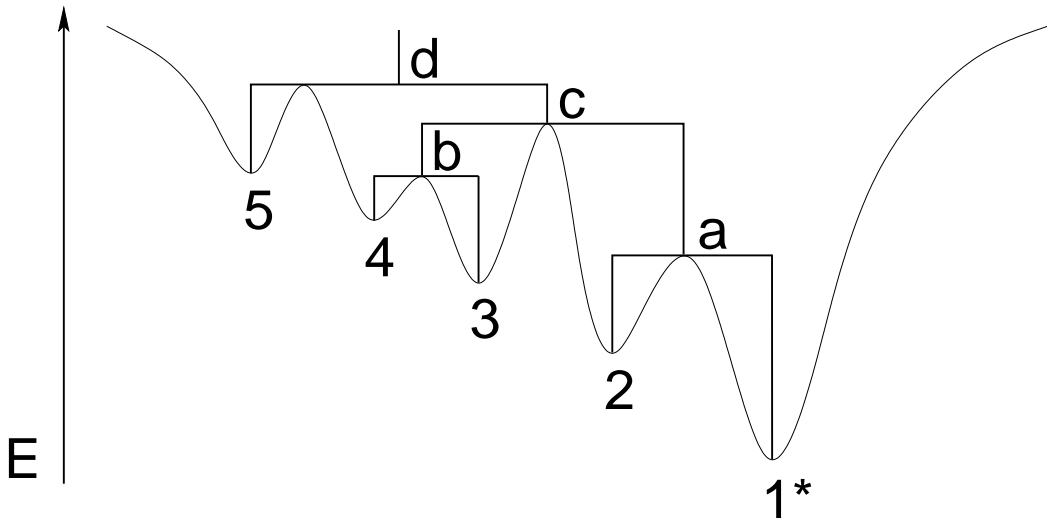
Fig. 2. Schematic representation of an energy landscape and its associated barrier tree. Local minima are labeled with numbers (1-5), saddle points with lowercase letters (a-d). The global minimum is marked with an asterisk.

2.2 Level Sets and Barrier Trees

A *cycle* or *level set* at energy level η can be defined as a maximal connected set $C \subseteq X$ such that $f(x) \leq \eta$ for all structures $x \in C$. Intuitively, one can interpret the level sets as basins of attraction. When the energy level η is increased, level sets grow and new level sets emerge. More formally, let A_η and $B_{\eta'}$ be two level sets at levels $\eta \geq \eta'$. Then either $B_{\eta'} \subseteq A_\eta$ or $A_\eta \cap B_{\eta'} = \emptyset$. This hierarchical structure is naturally represented by a tree. The leafs of this tree are the *local minima* of the landscape, i.e., those configurations x which do not have neighbors with lower energy. With each leaf/local minimum $\hat{x}$ and each energy level we can thus associate the connected level set $X_\eta[x]$. For consistency, we set $X_\eta[x] = \emptyset$ if $f(x) > \eta$. The level sets of two local minima $\hat{x}$ and $\hat{y}$ thus *merge* at the level η if $X_\eta[\hat{x}] = X_\eta[\hat{y}]$ and $X_{\eta'}[x] \cap X_{\eta'}[y] = \emptyset$ for all $\eta' < \eta$. The interior nodes of the barrier tree correspond to these "merging points". In the following, we write $\mathfrak{B}(X, f, \mathcal{M})$ for the barrier tree of the landscape $(X, f, \mathcal{M})$. Fig. 2 shows a simple example. For further formal details we refer to Flamm et al. (2002).

Given an energy-sorted listing of the L lowest energy configurations of the landscape, the barrier tree can be computed in $\mathcal{O}(L \times D)$ time and space (Flamm et al., 2000), where $D = \mathcal{O}(n^2)$ is the number of neighbors according to the move set. In the case of RNA secondary structures, our model at hand, the energy sorted list can in turn be computed in $\mathcal{O}(n^3 + nL + L \ln L)$ time and $\mathcal{O}(n^2 + nL)$ space using `RNAsubopt` (Wuchty et al., 1998). It is therefore feasible in practice to compute the barrier tree for RNAs of interesting sizes ($n \approx 100$) with moderate computational resources.

2.3 Macrostates

Let Π be a partition of X. The classes of Π can be seen as a coarse-graining of the configuration space. For our purposes, it will be of particular interest to consider partitions that are *consistent* with the energy function in the following sense:

If $Q \in \Pi$ then $Q_\eta := \{x \in Q | f(x) \leq \eta\}$ is either empty or a connected set.

It follows that every level-set is the union of such "lower parts" of macrostates. In the non-degenerate case, furthermore, each consistent macrostate has a unique local minimum $\hat{x}_Q$ that may serve as its representative.

For example, we can associate the conformation $x \in X$ with the local minimum $\gamma(x)$ that is reached from x by gradient decent. Again, in non-degenerate landscapes, γ is well-defined and the collection

$$\Pi^\gamma = \{\gamma^{-1}(\hat{z}) | \hat{z} \text{ is a local optimum}\} \tag{1}$$

of the *gradient basins* of local optima forms a partition of X. In degenerate landscapes we can break ties e.g. stochastically, see Flamm et al. (2002) for further details. Clearly, Π^γ is consistent with the energy function and hence also with the barrier tree. The local minima of the energy landscape thus act as representatives of the macrostates in this case.

2.4 Kinetics on Barrier Trees

This construction allows us to associate with each local minimum not only its "basin" in the barrier tree but also a macrostate that is consistent with the energy function and hence with the barrier tree. In particular, we use here the *gradient basins* Π^γ defined in the previous paragraph.

The dynamics of biopolymer folding, in our discrete picture, is given as a Markov process on X with transition rates of the form

$$p_{xy} \propto \begin{cases} \exp\left(-\frac{f(y)-f(x)}{RT}\right) & \text{if } x \in \mathcal{M}_y \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

As demonstrated in Wolfinger et al. (2004), one can approximate this dynamics by a dynamics on the set of macrostates *provided* one can argue that the process is approximately equilibrated within each class of Π. A slightly cruder,

but computationally much more efficient approximation entails an Arrhenius ansatz using the barrier tree to estimate the activation energies. For any two local minima $\hat{x} \neq \hat{y}$ we define their transition state energy $f[\hat{x}, \hat{y}]$ as the energy level at which their associated macrostates merge in the barrier tree, i.e.,

$$f[\hat{x}, \hat{y}] = \min \left\{ \eta | [\gamma^{-1}(\hat{x})]_\eta \cap [\gamma^{-1}(\hat{y})]_\eta \neq \emptyset \right\} . \tag{3}$$

Note that this expression coincides with the more "usual" definition of the barrier height as the minimum of the maximal height of paths connecting $\hat{x}$ and $\hat{y}$, see e.g. Nemoto (1988). The advantage of equ.(3) is that it emphasizes that the saddle height $f[\hat{x}, \hat{y}]$ can be computed as the *merging* of cycles within a flooding algorithm (Flamm et al., 2000), instead of the (algorithmically infeasible) optimization over all paths.

Transition rates between macrostates, represented here by the local minima that define them, are then given by the Arrhenius law

$$p_{\hat{x}, \hat{y}} = A \exp \left(-\frac{f[\hat{x}, \hat{y}] - f(\hat{y})}{RT} \right) \tag{4}$$

where A is normalization constant. For further details we refer to Wolfinger et al. (2004).

2.5   BarMaps

Given a landscape $(X, f, \mathcal{M})$ we now may ask how the folding behavior changes if we perturb the landscape. Such perturbations can take a wide variety of forms:

(1) $(X, \mathcal{M})$ remains the same, only the energy function is perturbed, $f \to g$. This is the case e.g. when temperature or ionic strength of the system is changed.
(2) (X, f) remains the same, but the move set changes $\mathcal{M} \to \mathcal{M}'$. This case is of interest when one is interested in the sensitivity of folding kinetics to changes in the underlying mechanistic models, e.g. to assess the impact of shift moves (Wuchty et al., 1998)
(3) X, f, and $\mathcal{M}$ change systematically. Examples are co-transcriptional folding or for experimental manipulations such as pulling an RNA molecule through a pore.

Our goal is to consider these types of changes in a coherent way in the framework of barrier trees. This will allow us to approximate the folding dynamics in time-variable landscapes of various types. Since we model the dynamics
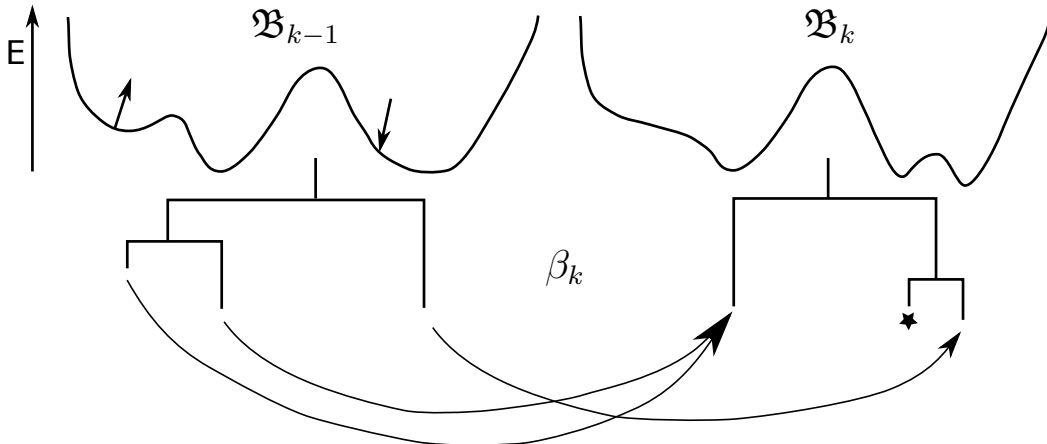
Fig. 3. Schematic of the bar map between two consecutive landscapes. Three types of events may occur: (i) two (here the two leftmost) local minima in $\mathfrak{B}_{k-1}$ merge into one (ii) a new minimum (marked by $\star$) appears in $\mathfrak{B}_k$ (iii) one to one correspondence between minima (as for the rightmost minimum here).

at the level of macrostates, we need to investigate how the perturbation of the landscape translates into changes of the barrier trees and their associated macrostates. In other words, we need to construct a map $\beta : \Pi \to \Pi'$ from the macrostates of $(X, f, \mathcal{M})$ to the macrostates of $(X', f', \mathcal{M}')$.

From a mathematical point of view, we first of all need a map $\xi : X \to X'$: $x \mapsto x'$ which specifies how the perturbation affects an individual conformation x before it "relaxes" in the modified landscapes. In the first two cases, this map is trivial: it coincides with the identity map, $\imath : x \mapsto x$, since the set of conformations does not change.

In the case of co-transcriptional folding it is also quite simple: When the next nucleotide is appended to a growing chain, it initially does not interact with the already folded "head" of the molecules, so that x' is x with an unpaired base appended, $x' = x +\!\!+' \bullet '$, where $+\!\!+$ denotes concatenation of strings.

The situation is a bit more complex in applications such as pulling macro-molecules through pores, or other mechanical constraints. In the pore case, the RNA structure is composed of two independently folding parts $x[1..k]$ and $x[k + \ell + 1..n]$, while the interval $[k + 1, k + \ell]$ is located within the pore and hence inaccessible to base-pairing. In the next step, the 5' part is $x[1..k-1]$; if k was paired, the base pair (j, k) now has been opened because nucleotide k is now covered by the pore. The other part is $x[k + \ell..n]$ where the first position, $k + \ell$ emerges unpaired from the pore. Note that in the pore case, the gradient descent operator γ also needs to be restricted to producing independent structures on both sides of the pore.

8

In the landscape $(X', f', \mathcal{M}')$ we have again well-define gradient basins by means of the steepest descent operator γ' on this landscape. The concatenation $\gamma'(\xi(z))$ thus maps every local minimum of $(X, f, \mathcal{M})$ to a local minimum of the perturbed $(X', f', \mathcal{M}')$ by first re-interpreting z in the new context and then relaxing it to the local minimum of the associated basin. It therefore implies the desired mapping β that maps macrostates of $(X, f, \mathcal{M})$ to the macrostates of $(X', f', \mathcal{M}')$. In other words, β maps the leafs of the barrier tree $\mathfrak{B}(X, f, \mathcal{M})$ to the leafs of the barrier tree $\mathfrak{B}(X', f', \mathcal{M}')$. We thus refer to β as the barrier tree map, or *bar map* for short, Fig. 3.

Note that, in general, the bar map is neither injective nor surjective: There may be local minima in $(X', f', \mathcal{M}')$ that are not the image of any local minimum of $(X, f, \mathcal{M})$, while multiple local minima of $(X, f, \mathcal{M})$ may be merged into a single minimum of $(X', f', \mathcal{M}')$.

2.6 Kinetics on Time-Variable Landscapes

The formalism developed in the previous subsections can be exploited to approximate RNA (re)folding kinetics on time-variable landscapes. The idea is to first determine a sequence of barrier trees $\{\mathfrak{B}_k\}$ together with barmaps $\beta_k : \mathfrak{B}_{k-1} \to \mathfrak{B}_k$. These data have to be determined only once. We are then free to choose a sequence $\{T_k\}$ of time points at which the system proceeds from $\mathfrak{B}_k$ to $\mathfrak{B}_{k+1}$. This allows us to explore the effects of variations in the speed of transcription, the rate temperature changes, or the pulling force in a manner that is independent of the computationally expensive analysis of the energy landscapes.

Denote by $\pi(\hat{x}, 0)$ the initial condition, i.e., the population densities in macrostate $\hat{x}$ on barrier tree $\mathfrak{B}_1$ at time 0. The population density on $\mathfrak{B}_1$ just before the transition to $\mathfrak{B}_1$ is $\pi(\hat{x}, T_1)$. The initial condition on the next barrier tree $\mathfrak{B}_{k+1}$ is obtained by collecting for each macrostate $\hat{y}$ the population densities of all those macrostates $\hat{x}$ of the previous barrier tree $\mathfrak{B}_k$ that map to $\hat{y}$ under the barmap β_k. In symbols:

$$\pi(\hat{y}, T_k) = \sum_{\hat{x}:\beta_k(\hat{x})=\hat{y}} \pi(\hat{x}, T_k) \tag{5}$$

Within the time interval $[T_k, T_{k+1}]$ we simply have to solve the master equation

$$\dot{\pi}(\hat{x}) = \sum_{\hat{y}} p_{\hat{x},\hat{y}} \pi(\hat{y}) \tag{6}$$

with $p_{\hat{x},\hat{y}} = -\sum_{\hat{z}} p_{\hat{z},\hat{x}}$ and the initial conditions described above. Note that the

transition matrix $P = (p_{\hat{x},\hat{y}})$ is by assumption independent of time for each fixed barrier tree. Thus the expensive part of solving the Master equation, namely the diagonalization of P, is also independent of the time intervals, and thus has the be performed only once for each barrier tree. After these preparatory computations have been performed, the population dynamics for a given schedule $\{T_k\}$ can be evaluated with a few matrix and vector multiplications. This sets the stage for an in-depth analysis of the interplay of folding dynamics and changes in the energy landscapes without substantial computational costs.

3   Results

3.1   The `BarMap` Software

The `BarMap` software is implemented as a combination of C programs and Perl scripts that form a pipeline for simulating folding time-dependent energy landscapes. In the first step of the pipeline, all low-energy structures of a landscape are computed using `RNAsubopt` from the Vienna RNA package. Subsequently, they are analyzed by the `barriers` program (Flamm et al., 2000). This is done separately for each landscape in the time series and yields both a barrier tree and a matrix of effective transition rates. The `bar_map` Perl program then computes the barmap β between consecutive barrier trees. Folding dynamics on each landscape are computed by the `treekin` program (Wolfinger et al., 2004). The final population on the landscape at time step $k-1$ is mapped to the the initial population on the landscape at time step k using the barmap β_k. A helper Perl script, `barmap_simulator`, is available that automatically generates the necessary `treekin` command lines. In order to plot folding dynamics as shown in Figs. 4 and 5 the `treekin` trajectory for the time intervals in which the landscapes is fixed are stitched together using the barmaps β_k, $k \geq 1$. This is accomplished by the final `bmjoin` script. An accompanying visualization tool, `BarMapViz` (Heine et al., 2006) can be used to create movies of a barrier tree sequences, facilitating the analysis of the landscape features that are responsible for particular kinetic effects.

Source code for the `barriers` and `treekin` programs, as well as the `bar_map`, `barmap_simulator`, and `bmjoin` Perl programs is available from `http://www.tbi.univie.ac.at/RNA/Barriers/`.
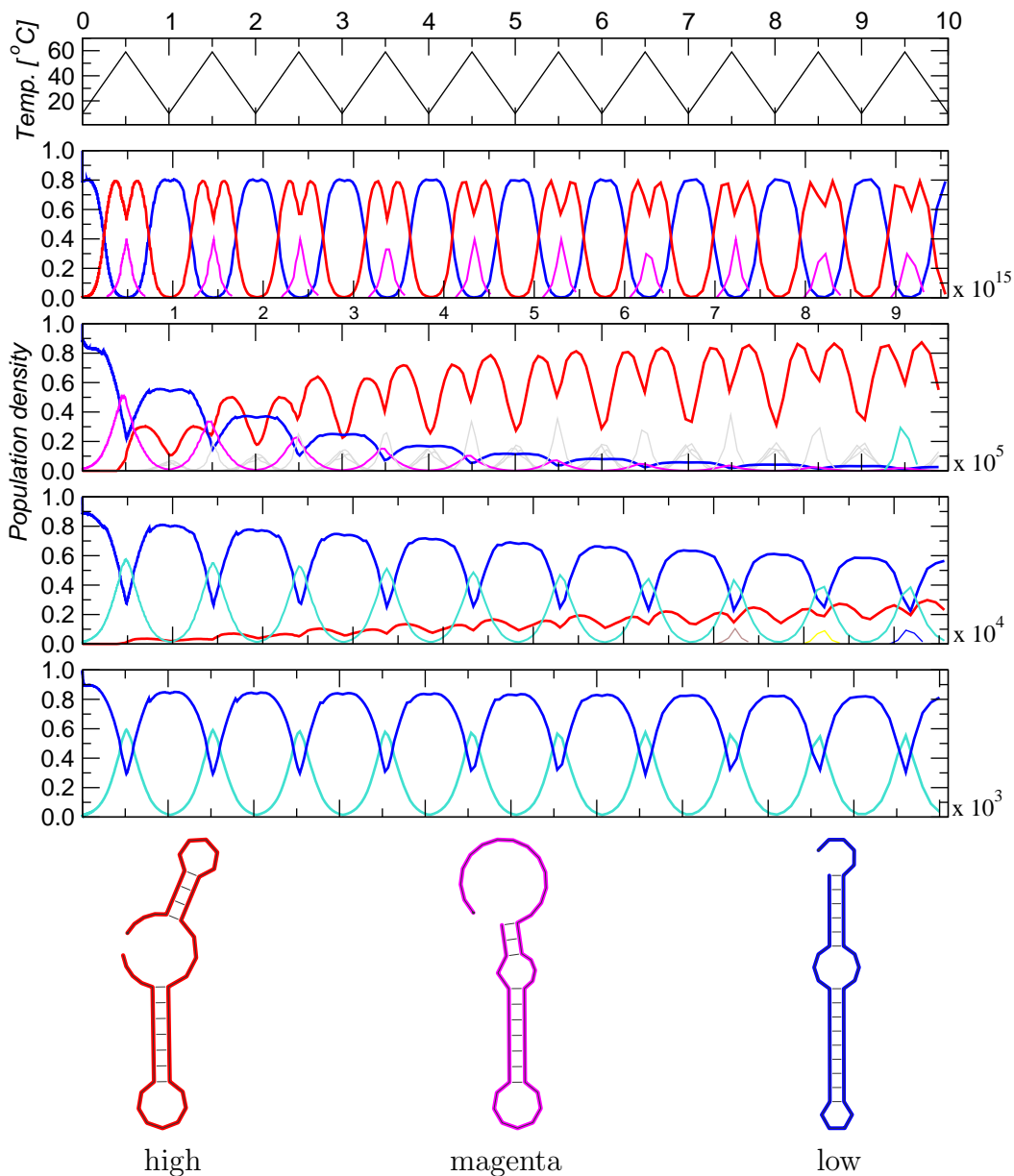
10

Fig. 4. Hysteresis effects in an thermo-sensitive RNA.

In this example, the temperature cycles periodically from 10 to $59°C$. The RNA has different optimal conformations at $10°C$ (blue) and $59°C$ (red), respectively. At high temperatures, furthermore, the minimum energy structure is nearly degenerate, so that an alternative structure (magenta) is populated substantially. For very fast temperature cycles, the only structural change that is fast enough is the opening of a GU:GU stack (cyan). These pairs are marked in the secondary structure diagrams.

3.2 Application 1: A RNA thermometer

Figure 4 shows the refolding dynamics of an artificial RNA thermometer when cycling between a high and low temperature regime. The sequence was de-

signed using the RNA switch designer described in Flamm et al. (2001), taking into account the sequence and structure constraints listed in Waldminghaus et al. (2008). This study demonstrated that *in silico* design with subsequent *in vivo* fine-tuning can produce temperature-controlled RNA elements with efficiencies comparable to their natural counterparts. For very slow temperature cycles (top), the molecule behaves adiabatically, effectively reaching thermodynamic equilibrium at each time step. The dynamics is therefore determined entirely by the barrier trees and the connecting barmaps. For intermediate cycling frequencies ($10^4 - 10^5$ time units per cycle), the system prefers the high-temperature structure. The relaxation time increases with cycling frequency. At even faster cycles, the system is trapped close to the (low temperature) starting conformation, since it does not have sufficient time to refold before the temperature drops again.

3.3 Application 2: Co-Transcriptional Folding

Under cellular conditions, RNA molecules start to fold before transcription is completed. This phenomenon is exploited by many bacteria to regulate the expression of amino acid biosynthesis genes (Yanofsky, 2000; Vitreschak et al., 2004; Merino et al., 2008). This RNA-based regulatory strategy by premature termination of transcription, often called *transcription attenuation* (Henkin and Yanofsky, 2002), relies on the selective formation of either of two mutually exclusive RNA secondary structures (the anti-terminator and the terminator) in the nascent transcript. The terminator structure causes premature termination of transcription.

We investigated the co-transcriptional folding dynamics of the leader RNA of the phenylalanine tRNA synthetase operon from `E. coli` (Fayat et al., 1983) under different transcription speeds, see Figure 5. For slow transcription, when the full-length chain is produced after $\sim 10^5$ arbitrary time units, the anti-terminator structure is formed (green curve top left panel). In contrast, under fast transcription conditions (transcription completed already after $\sim 10^4$ arbitrary time units), the terminator structure is formed (red curve bottom left panel). Since transcription attenuation operates far from the thermodynamic equilibrium, the kinetic competition between two small stem-loop structures (see blow-up panels on the right) decides whether the full-length leader RNA will eventually end up in the terminator or the anti-terminator structure. This competition early in the folding process is highly sensitive to the speed of transcription. Note, that for very long folding times ($\sim 10^{11}$) both co-transcriptional folding scenarios converge, as expected, to the thermodynamic equilibrium, which is dominated by the more stable terminator structure.

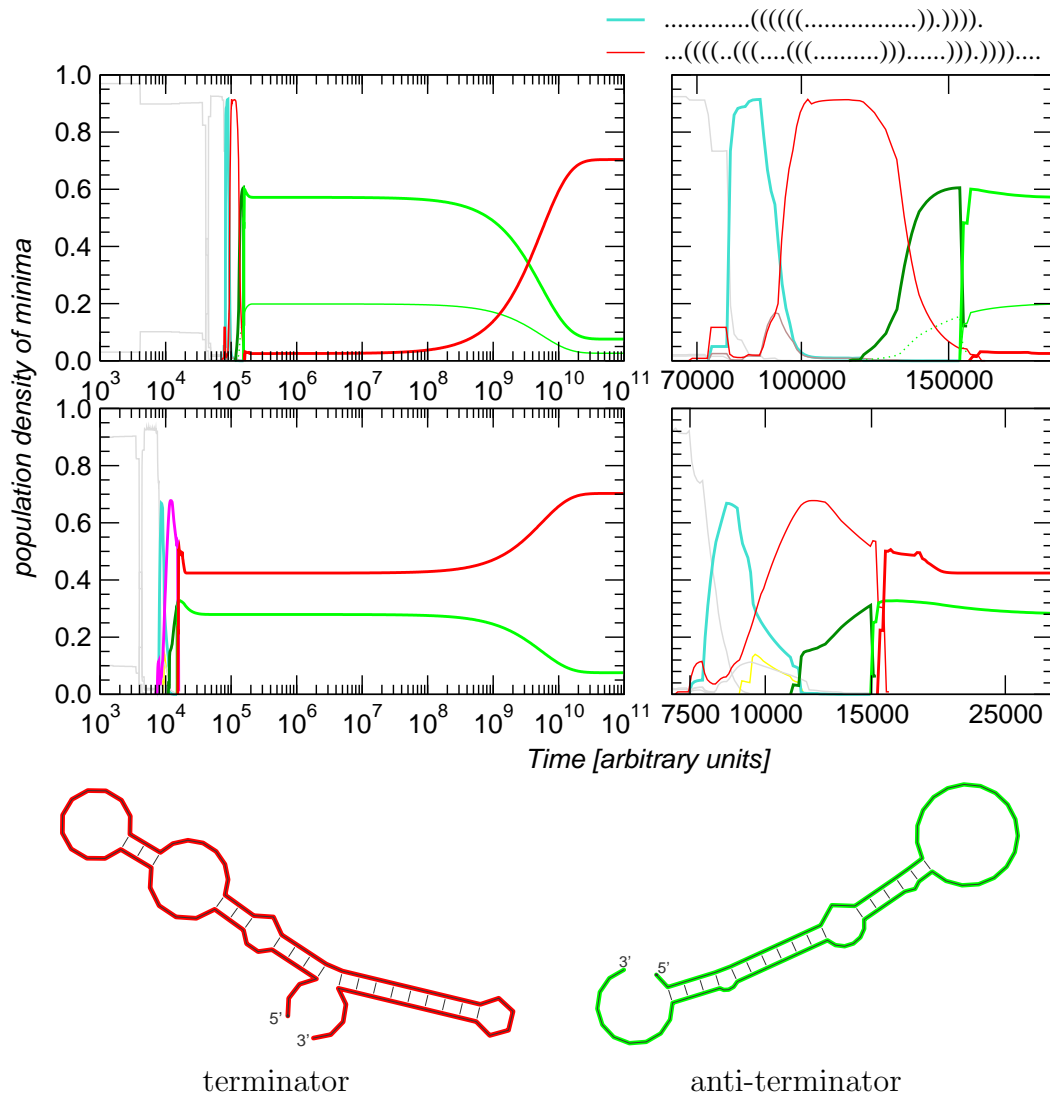*In vivo*, elongation speed is not constant, but influenced by site-specific paus-

Fig. 5. Co-transcriptional folding of the *E. coli* leader RNA of the tRNA[phe] synthetase operon for two different transcription speeds. For slow transcription (top) the completely transcribed chain shows a nearly zero density for the terminator structure, and transcription of the full length operon proceeds. For fast transcription, most of the fully elongated molecules form the terminator structure. Thermodynamic equilibrium is reached only on very long ($> 10^{10}$) time scales.

ing of the RNA polymerase and interactions of the nascent RNA with proteins (Pan and Sosnick, 2006). The effect of pause sites can easily be included in our approach. One simply needs to specify an appropriate elongation speed profile, i.e., an explicit list of time-points $\{T_k\}$ for the transitions from one landscape to the next.
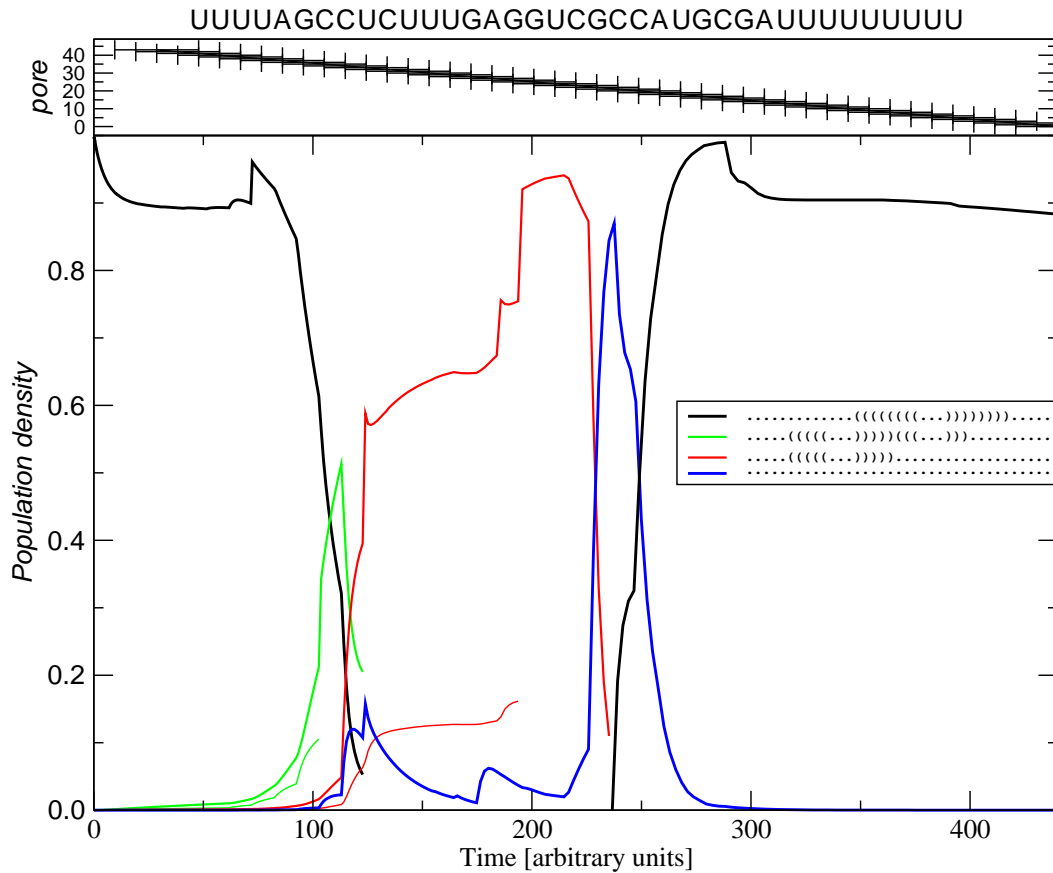
13

Fig. 6. Translocation of the artificial RNA sequence UUUUAGCCUCUUUGAGGUCGCCAUGCGAUUUUUUUU through a pore with a length of 5*nts*. The RNA enters the pore with its 3' end. As more and more of the minimum free energy (MFE) structure (black) is occluded by the pore, the RNA refolds into alternative structures (green and red). At the mid point, the most likely structure is the open chain (blue). Note how the probability of the MFE almost reaches 100% at $t \approx 290$, when the structure is fully formed, but alternative structures are inhibited by the pore.

3.4 Application 3: Re-folding during Pore Translocation

The transport of biopolymers through narrow pores is a fundamental process in life which is often coupled to the dynamics of biopolymer structure formation e.g. the base pair unfolding and folding dynamics while an mRNA passes through the ribosome during translation. Translocation of polymers is hindered by an entropic barrier, since the narrow confinement of the pore effectively separates the biopolymer into two independent sections resulting in a reduction of the chain entropy and hence an increase of the free energy of the chain (Muthukumar, 2007).

For structured nucleic acids, further kinetic barriers arise since the molecule has to locally unfold while passing through the pore (Bundschuh and Gerland,

14

2005; McCauley et al., 2009). In recent years, the single-molecule techniques of driving biopolymers through nano-pores using electric fields have been used to explore experimentally the structural and dynamic properties of nucleic acids (Vercoutere et al., 2001; Sauer-Budge et al., 2003; Mathé et al., 2004; Dudko et al., 2007).

We model the effect of the pore by allowing only secondary structures that are unpaired within the pore and contain no base pairs crossing from one side of the pore to the other. Figure 6 shows the resulting translocation dynamics for an artificial RNA sequence. In this example we use a slow translocation rate which allows the base pairing pattern on both sides of the pore to almost equilibrate.

4 Discussion

We have introduced here a very generic approach to investigate in detail the dynamic aspects of RNA folding in scenarios that involve external stimuli and/or changes of environment. By separating changes *in* the energy landscapes from the dynamics *on* these landscapes it becomes possible to avoid the extensive simulation of individual trajectories altogether. Instead, transition matrices between macrostates in each fixed landscape and "barmaps" linking the macrostates of temporally adjacent landscapes are computed in a pre-processing step. The time course of the population densities of macrostates is then obtained by means of a few matrix and vector operations. This computational efficiency allows detailed numerical studies of externally guided kinetic effects.

The examples described in the previous section highlight the major advantage of the `BarMap` approach: each energy landscape and its barrier tree, and all the barmaps between adjacent landscapes need to be computed only once. The transition rate matrices between macrostates within a landscape also have to be computed and diagonalized only once. The systematic exploration of the effects of different rates of change in the environment can thus be conducted very efficiently without the need to recompute any landscape-specific data. Time series of population densities in fact can be obtained using a few simple matrix and vector multiplications. The `BarMap` approach is thus particularly suitable to study the subtle kinetic effect that arises from the intricate interplay of different time scales.

5 Methods

5.1 RNA Folding

All structure predictions were performed using the Vienna RNA package (Hofacker et al., 1994) version 1.8.3, using the Turner energy parameters as described in Mathews et al. (1999).

5.2 Visualization of Barrier Tree Series

In order to gain a thorough understanding of the effects of changes in the landscape one needs to comprehend how these changes affect the corresponding barrier trees. To this end, we have developed the `BarMapVis` tool to create an animation of a sequences of barrier trees and the leaf mappings between adjacent trees (Heine et al., 2006). In brief, `BarMapVis` is based on the *foresight layout with tolerance* algorithm (Diehl and Görg, 2002), a very general attempt to solve any offline dynamic graph drawing problem. First, a directed acyclic supergraph G^* is constructed that contains all barrier trees as subgraphs and reflects the topological properties of all energy landscapes. The supergraph G^* is then laid out in the plane using a modified version of `dot` (Gansner et al., 1993). Finally, the layout of the subgraphs is determined by using the layout of the supergraph as a template following static drawing esthetic criteria in a way that approximately preserves the mental map (Misue et al., 1995) between consecutive barrier trees.

Animations showing the sequence of barrier trees generated by `BarMapVis` for each of the three examples from the Results section can be found in the web supplement.

Supplemental Material

Machine readable files of the input sequences, barrier trees, and `BarMapVis` movies can be downloaded from `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/BarMap/`. The `barmap` software can be downloaded from the `barriers` website `http://www.tbi.univie.ac.at/RNA/Barriers/`.

References

Abrahams JP, van den Berg M, van Batenburg E, Pleij C. 1990. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.* 18:3035–3044.

Baumstark T, Schroder AR, Riesner D. 1997. Viroid processing: Switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation. *EMBO J.* 16:599–610.

Becker OM, Karplus M. 1997. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* 106:1495–1517.

Bundschuh R, Gerland U. 2005. Coupled dynamics of RNA folding and nanopore translocation. *Phys Rev Lett.* 95:208104.

Clote P. 2005. An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. *J Comput Biol.* 12:83–101.

D. Thirumalai SAW, N. Lee, Klimov DK. 2001. Early events in RNA folding. *Annu. Rev. Phys. Chem.* 52:751–762.

Diehl S, Görg C. Graphs, they are changing — dynamic graph drawing for a sequence of graphs. In Goodrich MT, Kobourov SG, editors, *Proceedings of 10th International Symposium on Graph Drawing*, number 2528 in Lecture Notes Comp. Sci., pages 23–31, Heidelberg, 2002. Springer.

Doye JP, Miller MA, Welsh DJ. 1999. Evolution of the potential energy surface with size for Lennard-Jones clusters. *J. Chem. Phys.* 111:8417–8429.

Dudko OK, Mathé J, Szabo A, Meller A, Hummer G. 2007. Extracting kinetics from single-molecule force spectroscopy: nanopore unzipping of dna hairpins. *Biophys J.* 92:4188–95.

Fayat G, Mayaux JF, Sacerdot C, Fromant M, Springer M, Grunberg-Manago M, Blanquet S. 1983. *Escherichia coli* phenylalanyl-tRNA synthetase operon region. Evidence for an attenuation mechanism. Identification of the gene for the ribosomal protein L20. *J Mol Biol.* 171:239–61.

Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M. Feb 2001. Design of multistable RNA molecules. *RNA.* pages 254–65.

Flamm C, Fontana W, Hofacker I, Schuster P. 2000. RNA folding kinetics at elementary step resolution. *RNA.* 6:325–338.

Flamm C, Hofacker IL. 2008. Beyond energy minimization: Approaches to the kinetic folding of RNA. *Monatsh. f. Chemie.* 139:447–457.

Flamm C, Hofacker IL, Stadler PF, Wolfinger MT. 2002. Barrier trees of degenerate landscapes. *Z. Phys. Chem.* 216:155–173.

Gansner ER, Koutsofios E, North SC, Vo KP. 1993. A technique for drawing directed graphs. *IEEE Trans. Software Engineering.* 19:214–230.

Garstecki P, Hoang TX, Cieplak M. 1999. Energy landscapes, supergraphs, and "folding funnels" in spin systems. *Phys. Rev. E.* 60:3219–3226.

Gerdes K, Wagner GH. 2007. RNA antitoxins. *Curr. Op. Microbiol.* 10: 117–124.

Gollnick P, Babitzke P, Antson A, Yanofsky C. 2005. Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis. Annu. Rev. Genet.* 39:47–68.

Gultyaev AP. 1991. The computer simulation of RNA folding involving pseudoknot formation. *Nucl. Acids Res.* 19:2489–2493.

Heidrich D, Kliesch W, Quapp W. 1991. *Properties of Chemically Interesting Potential Energy Surfaces*, volume 56 of *Lecture Notes in Chemistry.* Berlin, Springer-Verlag.

Heine C, Scheuermann G, Flamm C, Hofacker IL, Stadler PF. 2006. Visualization of barrier tree sequences. *IEEE Trans. Vis. Comp. Graphics.* 12: 781–788. *Info-Vis 2006 contribution.*

Henkin TM, Yanofsky C. 2002. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *Bioessays.* 24:700–707.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125:167–188.

Hofacker IL, Schuster P, Stadler PF. 1998. Combinatorics of RNA secondary structures. *Discr. Appl. Math.* 88:207–237.

Klinkert B, Narberhaus F. Aug 2009. Microbial thermosensors. *Cell Mol Life Sci.* 66(16):2661–76.

Klotz T, Kobe S. 1994. "Valley Structures" in the phase space of a finite 3D Ising spin glass with $\pm i$ interactions. *J. Phys. A: Math. Gen.* 27:L95–L100.

Martinez HM. 1984. An RNA folding rule. *Nucl. Acid Res.* 12:323–335.

Mathé J, Visram H, Viasnoff V, Rabin Y, Meller A. 2004. Nanopore unzipping of individual dna hairpin molecules. *Biophys J.* 87:3205–12.

Mathews D, Sabina J, Zuker M, Turner H. 1999. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–940.

McCauley M, Forties R, Gerland U, Bundschuh R. 2009. Anomalous scaling in nanopore translocation of structured heteropolymers. *Phys Biol.* 6:036006.

Merino E, Jensen RA, Yanofsky C. 2008. Evolution of bacterial trp operons and their regulation. *Curr Opin Microbiol.* 11:78–86.

Mezey PG. 1987. *Potential Energy Hypersurfaces.* Amsterdam, Elsevier.

Mironov AA, Dyakonova LP, Kister AE. 1985. A kinetic approach to the prediction of RNA secondary structures. *Journal of Biomolecular Structure*

and Dynamics. 2:953.

Misue K, Eades P, Lai W, Sugiyama K. 1995. Layout adjustment and the mental map. *J. Visual Languages Comput.* 6:183–210.

Muthukumar M. 2007. Mechanism of dna transport through pores. *Annu Rev Biophys Biomol Struct.* 36:435–450.

Narberhaus F, Waldminghaus T, Chowdhury S. 2006. RNA thermometers. *FEMS Microbiol Rev.* 30:3–16.

Nemoto K. 1988. Metastable states of the SK spin glass model. *J. Phys. A.* 21:L287–L294.

Pan T, Sosnick T. 2006. RNA folding during transcription. *Annu Rev Biophys Biomol Struct.* 35:161–75.

Perrotta AT, Been MD. 1998. A toggle duplex in hepatitis delta virus self-cleaving RNA that stabilizes an inactive and a salt-dependent pro-active ribozyme conformation. *J. Mol. Biol.* 279:361–373.

Rapaport DC. 2004. *The Art of Molecular Dynamics Simulation.* Cambridge, UK, Cambridge University Press, 2nd edition.

Reidys CM, Stadler PF. 2002. Combinatorial landscapes. *SIAM Review.* 44: 3–54.

Sauer-Budge AF, Nyamwanda JA, Lubensky DK, Branton D. 2003. Unzipping kinetics of double-stranded dna in a nanopore. *Phys Rev Lett.* 90:238101.

Schultes EA, Bartel DP. 2000. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science.* 289:448–452.

Tacker M, Fontana W, Stadler PF, Schuster P. 1994. Statistics of RNA melting kinetics. *Eur. Biophys. J.* 23:29–38.

Tang X, Thomas S, Tapia L, Giedroc DP, Amato NM. 2008. Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.* 381:1055–1067.

Vercoutere W, Winters-Hilt S, Olsen H, Deamer D, Haussler D, Akeson M. Mar 2001. Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. *Nat Biotechnol.* 19(3): 248–252.

Vitreschak AG, Lyubetskaya EV, Shirshin MA, Gelfand MS, Lyubetsky VA. 2004. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis. *FEMS Microbiol Lett.* 234: 357–370.

Waldminghaus T, Kortmann J, Gesing S, Narberhaus F. 2008. Generation of synthetic rna-based thermosensors. *Biol Chem.* 389:1319–1326.

Wales DJ, Miller MA, Walsh TR. 1998. Archetypal energy landscapes. *Nature.* 394:758–760.

Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF. 2004. Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.* 37:4731–4741.

Wolfinger MT, Will S, Hofacker IL, Backofen R, Stadler PF. 2006. Exploring the lower part of discrete polymer model energy landscapes. *Europhys. Lett.* 74:726–732.

Wuchty S, Fontana W, Hofacker IL, Schuster P. 1998. Complete suboptimal folding of RNA and the stability of secondary structure. *Biopolymers.* 49: 145–165.

Yanofsky C. 2000. Transcription attenuation: once viewed as a novel regulatory strategy. *J Bacteriol.* 182:1–8.