

Memory efficient RNA energy landscape exploration

Martin Mann^{1,*}, Marcel Kucharík², Christoph Flamm²
and Michael T. Wolfinger^{2,3,4}

¹ Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, D-79110 Freiburg, Germany

² Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Vienna, Austria

³ Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories, University of Vienna & Faculty of Computer Science, University of Vienna, Dr. Bohr-Gasse 9, 1030 Vienna, Austria.

⁴ Department of Biochemistry and Molecular Cell Biology, Max F. Perutz Laboratories, University of Vienna, Dr. Bohr-Gasse 9, 1030 Vienna, Austria

Abstract

Energy landscapes provide a valuable means for studying the folding dynamics of short RNA molecules in detail by modeling all possible structures and their transitions. Higher abstraction levels based on a macro-state decomposition of the landscape enable the study of larger systems, however they are still restricted by huge memory requirements of exact approaches.

We present a highly parallelizable local enumeration scheme that enables the computation of exact macro-state transition models with highly reduced memory requirements. The approach is evaluated on RNA secondary structure landscapes using a gradient basin definition for macro-states. Furthermore, we demonstrate the need for exact transition models by comparing two barrier-based approaches and perform a detailed investigation of gradient basins in RNA energy landscapes.

Source code is part of the C++ Energy Landscape Library available at <http://www.bioinf.uni-freiburg.de/Software/>.

1 Introduction

The driving force of disordered systems in physics, chemistry and biology is characterized by coupling and competing interaction of microscopic components. At a qualitative level, this is reflected by the potential energy function and often results in complex topological properties induced by individual conformational degrees of freedom. It seems fair to say that it is

*to whom correspondence should be addressed: <http://www.bioinf.uni-freiburg.de>

practically impossible to compute dynamic and thermodynamic properties directly from the Hamiltonian of such a complex system. However, analyzing the underlying energy landscape and its features directly provides a valuable alternative.

Here, we focus on RNA molecules and their folding kinetics. RNAs are key players in cells acting as regulators, messengers, enzymes, and many more roles. In many cases, a specific structure is crucial for biological specificity and functionality. The formation of these functional structures, *i.e.* the folding process, can be studied at the level of RNA energy landscapes (Flamm and Hofacker, 2008; Geis *et al.*, 2008).

RNA is composed of the biophysical alphabet $\{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{U}\}$ and has the ability to fold back onto itself by formation of discrete base pairs, thus forming secondary structures. The latter provide a natural coarse-graining for the description of the thermodynamic and kinetic properties of RNA, because, in contrast to proteins, the secondary structure of RNA captures most of the folding free energy. This is accommodated by novel approaches for predicting three-dimensional RNA structures from secondary structures (Popenda *et al.*, 2012).

Formally, an RNA secondary structure is defined as a set of base pairs between the nuclear bases complying with the rules: (a) only $\mathbf{A-U}$, $\mathbf{G-C}$, and $\mathbf{G-U}$ pairings are allowed, (b) any base is involved in maximal one base pair, and (c) the structure is nested, *i.e.* there are no two base pairs with indices $(i, j), (k, l)$ with $i < k < j < l$. Summation over the individual base pair binding energies and entropic contributions for unpaired bases defines the energy function E (Hofacker *et al.*, 1994; Tinoco *et al.*, 1971; Freier *et al.*, 1986). The degeneracy of this energy definition is countered via a structure ordering based on their string encoding (Flamm *et al.*, 2002). We refer to the literature (Flamm and Hofacker, 2008; Chen, 2008) for more details.

In this work, we study the folding kinetics of RNA molecules by means of a discrete energy landscape approach. While stochastic folding simulations based on solving the Master equation are limited to relatively short sequence lengths (Flamm *et al.*, 2000b; Aviram *et al.*, 2012), a common approach to studying biopolymer folding dynamics is using a coarse grained model that partitions the energy landscape into distinct basins of attraction, thus assigning macro-states to each basin (Wolfinger *et al.*, 2004). The basin decomposition and computation has been described in different contexts, including Potential Energy Landscapes (Heuer, 2008), RNA kinetics (Flamm and Hofacker, 2008) and lattice protein folding (Wolfinger *et al.*, 2006; Tang and Zhou, 2012). Given appropriate transition rates between macro-states (optionally comprised of rates among micro-states that form a macro-state), the dynamics can be modeled as continuous-time Markov process and solved directly by numerical integration (Wolfinger *et al.*, 2004). While suitable for system sizes up to approx. 10,000 states, improvements to this approach are currently subject to our research, allowing investigation of up to a few hun-

dred thousand states by incorporating sparsity information and additional approximations.

The crucial step in the procedure sketched above is to obtain the transition rates between macro-states. Global methods for complete (Flamm *et al.*, 2002) or partial (Sibani *et al.*, 1999; Kubota and Hagiya, 2005; Wolfinger *et al.*, 2006) enumeration of the energy landscape are not applicable to large systems due to memory restrictions. On the other side, sampling with high precision requires long sampling times (Mann and Klemm, 2011). Therefore approximating the energy landscape by a subset of important local minima, gained via sampling approaches or spectroscopic methods (Fürtig *et al.*, 2007; Alemán *et al.*, 2008; Rinnenthal *et al.*, 2011), and transition paths between them (Noé and Fischer, 2008) has been investigated in the past (Tang *et al.*, 2005, 2008; Kucharik *et al.*, 2014).

We propose a novel, highly parallelizable and memory efficient local enumeration approach for computing exact transition probabilities. While the method is intrinsically generic and can be readily applied to other discrete systems, we exemplify the concept in the context of energy landscapes of RNA secondary structures, based on the Turner energy model (Xia *et al.*, 1998), as implemented in the **Vienna RNA Package** (Hofacker *et al.*, 1994; Lorenz *et al.*, 2011) and the Energy Landscape Library (Mann *et al.*, 2007). We evaluate the memory efficiency and dynamics quality for different RNA molecules and report features of gradient basin macro-states in RNA energy landscapes.

2 Discrete Energy Landscapes

In the following, we will define energy landscapes for two levels of abstraction: the *microscopic level* covers all possible (micro-) states of a system and its dynamics, while the *macroscopic level* enables a more coarse grained model of the system’s dynamics, based on a partitioning of all micro-states into macro-states. The macroscopic view is required when studying the dynamics of larger systems.

2.1 Microscopic Level

Discrete energy landscapes are defined by a triple (X, E, M) given a finite set of (micro-)states X , an appropriate energy function $E : X \rightarrow \mathbb{R}$, and a symmetric neighborhood relation $M : X \rightarrow \mathcal{P}(X)$ (also known as move set), where $\mathcal{P}(X)$ is the power set of X . The neighborhood $M(x)$ is the set of all neighboring states that can be directly reached from state x by a simple move set operation.

Consequently, RNA energy (folding) landscapes can be defined at the level of secondary structures, which represent the micro-states $x \in X$. An

RNA structure y is neighbored to a structure x ($y \in M(x)$) if they differ in one base pair only. While alternative move set definitions are possible (Flamm *et al.*, 2000b), they are not considered in this work for simplicity.

Within this work, we consider time-discrete stochastic dynamics based on Metropolis transition probabilities p at inverse temperature β :

$$\begin{aligned} p_{x \rightarrow y} &= \Delta^{-1} \min\{\exp(-\beta[E(y) - E(x)]), 1\} \\ &= \Delta^{-1} \min\{w(y)/w(x), 1\} \end{aligned} \quad (1)$$

$$\text{with } w(x) = \exp(-\beta E(x)) \quad (2)$$

$$\text{and } \Delta = \max_{x \in X} |M(x)|. \quad (3)$$

$w(x)$ is the Boltzmann weight of x . Normalization is performed via the constant Δ , which is the maximally possible number of neighbors/transitions of any state. The transition probability $p_{x \rightarrow y}$ is only defined for neighboring states, *i.e.* $y \in M(x)$.

2.2 Macroscopic Level

Although desirable, studying dynamic properties at the microscopic level is often not feasible due to the vastness of the state space X , even for relatively small systems. An alternative approach is coarse graining *i.e.* lumping many micro-states into fewer macro-states, such that the microscopic dynamics is resembled as closely as possible (Wolfinger *et al.*, 2004).

This can be achieved by partitioning of the state space X with a mapping function $F : X \rightarrow B$ that uniquely assigns any micro-state in X to a macro-state in B . With $F^{-1}(b)$ we denote the inverse function that gives the set of all F -assigned states for a macro-state $b \in B$. Following (Kramers, 1940; Wolfinger *et al.*, 2004; Flamm and Hofacker, 2008; Mann and Klemm, 2011), we will use the simplifying assumption that the probability of the system to be in micro-state x while it is in macro-state $b \in B$ is given by

$$P_b(x) = \begin{cases} w(x)Z_b^{-1} & \text{if } x \in F^{-1}(b) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$\text{with } Z_b = \sum_{y \in F^{-1}(b)} w(y). \quad (5)$$

Based on this, we can define the macroscopic transition probabilities $q_{b \rightarrow c}$ between macro-states $b, c \in B$ by means of the microscopic probabilities p

from Eq. 1 as follows:

$$\begin{aligned}
q_{b \rightarrow c} &= \sum_{x \in F^{-1}(b)} \left(P_b(x) \sum_{y \in M(x) \cap F^{-1}(c)} p_{x \rightarrow y} \right) \\
&= \sum_{(x,y)} P_b(x) p_{x \rightarrow y} \\
&= \sum_{(x,y)} \frac{w(x)}{Z_b} \Delta^{-1} \min\{w(y)/w(x), 1\} \\
&= Z_b^{-1} \sum_{(x,y)} \Delta^{-1} \min\{w(y), w(x)\} \\
&= Z_b^{-1} Z_{\{b,c\}} \text{ and thus} \tag{6} \\
q_{c \rightarrow b} &= Z_c^{-1} Z_{\{b,c\}}. \tag{7}
\end{aligned}$$

Equation (6) considers all microscopic transitions $x \rightarrow y$ from a micro-state x in b to a micro-state y in c , based on the probability of x ($P_b(x)$) and the transition probability $p_{x \rightarrow y}$. The energetically higher micro-state of each such transition contributes to the partition function of all transition states between b and c , $Z_{\{b,c\}}$ (Eq. 6 and 7). Consequently $Z_{\{b,c\}} \equiv Z_{\{c,b\}}$, *i.e.* the transition state partition function is direction-independent.

Within this work, we use the common gradient basin partitioning of X following (Doye, 2002; Flamm *et al.*, 2002; Flamm and Hofacker, 2008; Mann and Klemm, 2011). A gradient basin is defined as the set of all states who have a steepest descent (gradient) walk ending in the same local minimum, where $\tilde{x}$ is a local minimum if $\forall_{y \in M(\tilde{x})} : E(\tilde{x}) < E(y)$. In this context the set of macro-states B is given by the set of all local minima of the landscape, whose number is drastically smaller than that of all micro-states (Lorenz and Clote, 2011). The mapping function $F(x)$ applies a gradient walk starting in x , thus assigning it a local minimum $\tilde{x}$ and a macro-state b . Here, the minimum is used as a representative for the macro-state comprised of the gradient basin.

A coarse abstraction of the macro-state transition probabilities can be obtained by an Arrhenius-like transition model (Wolfinger *et al.*, 2004). Here, the transition probability is dominated by the minimal energy barrier that needs to be traversed in order to go from one state to another. Formally, given two states x and y , one has to identify the path $p = (x_1, \dots, x_l) \in X^l, l > 1$ with $x_1 = x, x_l = y$, and $\forall i < l : x_{i+1} \in M(x_i)$ with lowest energy maximum. Arrhenius barrier-based transition probabilities are thus defined by

$$a_{x \rightarrow y} = A \exp(-\beta(E(x, y) - E(x))) \text{ with} \tag{8}$$

$$E(x, y) = \min_{p \in X^*} \max_{x_i \in p} (E(x_i)) \tag{9}$$

where A is an intrinsically unknown pre-exponential factor. For macro-state transitions based on a gradient basin partitioning, transition probabilities can be approximated by Arrhenius probabilities among local minima of macro-states. In this context it is important to note that this transition model does not enforce neighborhood of the macro-states. The impact on modeling quality of such an Arrhenius-based model is evaluated in Sec. 4. We will now present approaches for the exact determination of the macro-state transition probabilities for a given landscape and partitioning.

3 Macro-state transition probabilities

Following the rationale presented above, all macroscopic transition rates need to be determined in order to study the coarse-grained dynamics. Given Eq. 6, the partition function Z_b (Eq. 5) and adjunct partition functions of transition states $Z_{\{b,c\}}$ to adjacent $c \neq b$ have to be computed for each macro-state b .

A direct approach is brute-force enumeration of X , computing $F(x)$ for each micro-state $x \in X$ and updating $Z_{F(x)}$ accordingly. Subsequently, all neighbors $y \in M(x)$ are enumerated in order to update $Z_{\{F(x),F(y)\}}$ if $F(x) \neq F(y)$. While this is the simplest and most general approach, it is not efficient for the majority of definitions of F . It can, however, be replaced with more efficient dedicated flooding algorithms and can be even more tuned for gradient basin definitions of F as we will discuss now.

3.1 Standard approach via global flooding

The **lid** method (Schön and Sibani, 1998; Sibani *et al.*, 1999) performs a “spreading” enumeration starting from a local minimum with an upper energy bound for micro-states to consider, the lid. Internally, two lists are hashed: The set **D** containing all micro-states that have been processed so far and the “todo-list” **T** comprised of states neighbored to **D** but not handled yet. Each processed micro-state x is assigned to its corresponding macro-state $b = F(x)$ during the enumeration process. b is stored along with x in **D** and **T** and the partition function Z_b is updated by $w(x)$ accordingly. Subsequently, all neighbors $y \in M(x)$ of x with $E(y) < \text{lid}$ -threshold are enumerated and either found in **D** or **T** (thus saving $F(x)$ computation) or added to **T**. If the macro-state assignment for x and y differs, *i.e.* $F(x) \neq F(y)$, the corresponding transition state partition function $Z_{\{F(x),F(y)\}}$ is increased by $\Delta^{-1} \min(w(x), w(y))$. The method was reformulated by Kubota and Hagiya (2005) for DNA energy landscapes and Wolfinger *et al.* (2006) in the context of lattice proteins.

The **barriers** approach by Flamm *et al.* (2002) performs a “bottom-up” evaluation of energy landscape topology based on an energy-sorted list of all micro-states in X above the ground state up to a predefined energy

threshold. Here, the macro-state assignment F can be handled more efficiently compared to the `lid`-method if gradient basins are applied: Given that the steepest descent walk used for a gradient mapping F is recursive, *i.e.* the assignment $F(x)$ of a state x is known as soon as the assignment of its steepest descent neighbor $m_{\min} \in M(x)$, $F(m_{\min})$, is known, the macro-state assignment is accomplished by a single hash lookup: Since the processed set of states $\mathbf{D}$ already contains all states with energy less than $E(x)$, looking up $m_{\min}$ and its corresponding macro-state $F(m_{\min})$ in $\mathbf{D}$ yields $F(x) \equiv F(m_{\min})$. The energy of the micro-state currently processed marks the “flood level”, *i.e.* all states in X with energy below have been processed. Consequently, the macro-state partition functions Z_b are collected as soon as the flood level reaches the according local minimum defining b .

Both methods perform a massive hashing of processed states and are thus restricted by memory, *i.e.* the number of micro-states that can be stored in $\mathbf{D}$ and $\mathbf{T}$ is constrained to the available memory resources. Considering the exponential growth *e.g.* of the RNA structure space X (Hofacker *et al.*, 1998), the memory is easily exhausted for relatively short sequence lengths. As the memory limit is approached, both methods result in incomplete macro-state transition data.

The `barriers` approach ensures a “global picture” of the landscape since it covers the lower parts of all macro-states up to the reached flood level exhaustively, missing all macro-states above the limit. In case the transition states connecting the macro-states are above the flood level, no transition information is available. This can be approached by heuristics approximating the transition barrier (Morgan and Higgs, 1998; Flamm *et al.*, 2000a; Wolfinger *et al.*, 2004; Richter *et al.*, 2008; Bogomolov *et al.*, 2010), however the outcome is still not reflecting the true targeted macro-state dynamics. In contrast, the `lid` method will always result in connected macro-states but only a restricted part of the landscape is covered. Furthermore, each macro-state is enumerated up to different (energy) heights resulting in varying quality of the collected partition function estimates, which further distorts the dynamics.

3.2 Memory efficient local flooding

To overcome the memory limitation of global flooding approaches, we introduce a local flooding scheme. It enables parallel identification of the partition function Z_b and all transition state partitions $Z_{\{b,c\}}$ for a macro-state b without the need of full landscape enumeration.

Similar to global flooding, the *local* approach uses a set $\mathbf{D}$ of already processed micro-states that are *part of* b , *i.e.* $\forall x \in \mathbf{D} : F(x) = b$, and a set $\mathbf{T}$ of micro-states that might be part of b or adjacent to it.

The algorithm starts in the local minimum $l \in X$ of b , *i.e.* $F(l) = b$ and $\forall x \neq l \in F^{-1}(b) : E(x) > E(l)$, and does a local enumeration of micro-

states in increasing energy order starting from b . Thus, Z_b is initialized with $Z_b = w(l)$, all neighbors $m \in M(l)$ of the minimum are pushed to $\mathbf{T}$, and l is added to $\mathbf{D}$. Afterwards the following procedure is applied until $\mathbf{T}$ is empty.

1. get energy minimal micro-state x from $\mathbf{T}$ with $\forall_{x' \neq x \in \mathbf{T}} : E(x) < E(x')$
2. identify steepest descent neighbor $m_{\min} \in M(x)$ with $\forall_{m \neq m_{\min} \in M(x)} : E(m_{\min}) < E(m)$
3. if $m_{\min} \in \mathbf{D} \rightarrow F(x) = b$:
 - x is added to $\mathbf{D}$,
 - $Z_b = Z_b + w(x)$,
 - all neighbors $m \in M(x)$ with $E(m) > E(x)$ are added to $\mathbf{T}$, and
 - descending transitions leaving b are handled:
 x is transition state for all $m \in M(x)$ with $E(m) < E(x)$ and $m \notin \mathbf{D}$:
 $Z_{\{b, F(m)\}} = Z_{\{b, F(m)\}} + \Delta^{-1}w(x)$
- else $\rightarrow F(x) \neq b$:
 - descending transitions entering b are handled:
 x is transition state for all $m \in M(x)$ with $E(m) < E(x)$ and $m \in \mathbf{D}$:
 $Z_{\{F(x), b\}} = Z_{\{F(x), b\}} + \Delta^{-1}w(x)$

We use a data structure for $\mathbf{T}$ that is automatically sorted by increasing energy in order to boost performance of step 1.

The algorithm computes Z_b and $Z_{\{b, c\}}$, which are required for deriving the macro-state transition rates $q_{b \rightarrow c}$ (Eq. 6) from one macro-state b to adjacent macro-states $c \neq b$. It is individually applied to all macro-states in order to get the full transition rate information of the energy landscape. Evidently, the transition state partition function $Z_{\{b, F(x)\}}$, covering states between two macro-states b and c , has to be computed only once for each pair (see Eq. 6 and 7).

The major advantage of the local flooding method compared to global flooding approaches is an extremely reduced memory consumption. This is achieved by only storing the micro-states part of the current macro-state b (set $\mathbf{D}$) plus all member and transition state candidates (set $\mathbf{T}$). The reduction effect is studied in detail in the next section and an implementation of the presented local flooding has been added to the Energy Landscape Library (ELL) (Mann *et al.*, 2007). The ELL provides a generic platform for an independent implementation of algorithms and energy landscape models to be freely combined (Mann *et al.*, 2008; Mann and Klemm, 2011).

Within this work, we tested our new method using the ELL-provided RNA secondary structure model as discussed in the following section.

The reduced memory consumption of the local flooding scheme comes at the cost of increased computational efforts for the assignment of states that are not part of macro-state b . The above workflow does an explicit computation of F for all these states. Here, more sophisticated methods can be applied that either do a full or partial hashing of states partaking in steepest descent walks to increase the performance.

Another advantage is the inherent option for distributed computing since the local flooding is performed independently for each macro-state. As such, we can yield a highly parallelized transition rate computation not possible in the global flooding scheme. This can be combined with an automatic landscape exploration approach where each local flooding instance identifies neighboring, yet unexplored macro-states that will be automatically distributed for processing until the entire energy landscape is discovered.

We will now investigate the requirement and impact of our local flooding approach in the context of folding landscapes of RNA molecules.

4 Folding landscapes of RNA molecules

In the following, we will study the energy landscapes for the bistable RNA d33 from (Mann and Klemm, 2011) and the iron response element (IRE) of the Homo sapiens L-ferritin gene (GenBank ID: KC153429.1) in detail. The sequences are GGGAAUUAUUGUCCCCUGAGAGCGGUAGUUCUC and CUGUCUCUUGCUUCAACAGUGUUUGGACGGAACAG, respectively. In addition, and in order to evaluate the general character of our results, we generated 110 random RNA sequences with uniform base composition, 10 for each length from 25nt to 35nt. For this set average values are reported. The length restriction was a requirement for comparison to exhaustive methods.

4.1 Exact vs. approximated transition models

We will first investigate whether exact macro-state transition probabilities are essentially required for computing a coarse-grained dynamics or if an approximated model is providing similar results. To address this question, we performed an exhaustive enumeration of the RNA energy landscapes for d33 and ire, resulting in approximately 30 and 21 million micro-states, respectively, that are clustered into approximately 2,900 gradient basin macro-states for each sequence. These basins are connected by approximately 60,000 macro-state transitions, representing only a fraction of 1.5% of all possible pairwise transitions.

The concept of barrier trees (Flamm *et al.*, 2002; Flamm and Hofacker, 2008) represents a straightforward approach for modelling the coarse-grained

folding dynamics of an RNA molecule without explicit knowledge of the exact pairwise microscopic transition probabilities. In this context, transition probabilities between any two gradient basin macro-states b and c are defined via an Arrhenius-like equation. The latter is given in Eq. 8, considering the energy difference ΔE between the local minimum of macro-state b and the lowest saddle point of any path to the target macro state c (which may traverse some other macro-states). The saddle point can be identified either via exhaustive enumeration (Flamm *et al.*, 2002) or estimated by path sampling techniques (Richter *et al.*, 2008; Lorenz *et al.*, 2009; Bogomolov *et al.*, 2010; Li and Zhang, 2012; Kucharič *et al.*, 2014). Energy barriers can be visualized in a tree-like hierarchical data structure, the barrier tree, resulting in all n^2 pairwise transition probabilities for n macro-states. Coarse-grained folding kinetics based on this framework have been shown to resemble visual characteristics of the exact macro-state kinetics (Flamm *et al.*, 2002; Wolfinger *et al.*, 2004).

The supplementary material provides a visual comparison of coarse-grained folding dynamics for RNA d33, based on two different transition models. While the pure barrier tree dynamics resembles the overall dynamics of the two energetically lowest macro-states of the exact model quite well, it shows significant differences for states populated at lower extent. Given these visual discrepancies, we are interested in measuring the modelling quality of the barrier tree-based transition model vs. the exact configuration. To this end, we will analyze mean first passage times (FPT) and their correlations. The FPT $\tau(b, t)$, also termed exit time (Freier *et al.*, 1986), is the expected number of steps to reach the target state $t \in B$ from a start state $b \in B$ for the first time (Grinstead and Snell, 1997). The first passage time for a state to itself is 0 per definition, *i.e.* $\tau(b, b) = 0$. For all other cases, it is defined by the recursion

$$\tau(b, t) = 1 + \sum_{c \in B} q_{b \rightarrow c} \tau(c, t). \quad (10)$$

We are focused on folding kinetics, *i.e.* we compute the FPT from the unfolded state to all other macro-states using (a) the exact macro-state transition probabilities (Eq. 6) (full model) and (b) the barrier tree-based transition probabilities based on the Arrhenius equation (Eq. 8, barrier model).

First passage time values depend on the intrinsically unknown Arrhenius prefactor. As such, we will compare the two models using a Spearman rank correlation of the FPT, *i.e.* we compare the relation between FPTs rather than final values.

For d33 and ire the Spearman rank correlation coefficients is 0.28 and -0.12, respectively, indicating no correlation. The random sequence set shows a mean coefficient of 0.2, indicating no correlation either. No length-dependent bias was found (see suppl. material). Results are summarized in Table 1.

Sequence(s)	Spearman corr. exact – barrier	Spearman corr. exact – merged
d33	0.28	0.85
ire	-0.12	0.64
random	0.20	0.71

Table 1: Spearman rank correlation of different macro-state transition models. Comparison of the Arrhenius barrier-based and the exact model (Eq. 6) shows almost no correlation, while the merged model of both (see text) is highly correlated to the exact model.

The barrier model is a simplification of the full model in two aspects: 1.) *loss of precision* – the computation of transition rates based on Arrhenius-like equations is less accurate and 2.) *loss of topology* – the barrier model allows for all possible pairwise transitions, which may lead to an overestimation of transitions. To further distinguish between these two transition approaches, we have derived a merged transition model with modified transition probabilities q' . Within this merged model, $q'_{b \rightarrow c}$ is given by the Arrhenius-like equation (Eq. 8) for all exact macro-state transitions ($q_{b \rightarrow c} \neq 0$, Eq. 6) and zero otherwise. Investigating the Spearman rank correlation of the merged model’s FPTs with the exact FPTs, an increased correlation coefficient (0.85 and 0.64 for d33 and ire, resp.), is observed. This is supported by a robust average coefficient of 0.71 for the set of random sequences (see suppl. material).

These results clearly show two key aspects of reduced folding dynamics: First, importance of the underlying topology of the landscape, *i.e.* the necessity to identify sparse exact transitions between macro-states, and second the reduced modeling quality when restricting the computation of transition probabilities to energy barrier-based (Arrhenius-like) approximations. The importance of the topology information for kinetics is partly studied in the supplementary material of Kucharík *et al.* (2014).

4.2 Reduction of memory requirement

Given the need for an exact computation of macro-state transition probabilities, we will now evaluate the impact of a local flooding scheme compared to the standard global flooding approach. In this context, we will investigate the memory footprint, which is the central bottleneck of global flooding methods.

As outlined above, global flooding schemes keep track of all micro-states $x \in X$ within the energy landscape. As such, the global flooding memory consumption is dominated by $\text{mem}(\text{G}) = |X|$.

In contrast to that, all micro-states $x \in F^{-1}(b)$ of b in the local flooding scheme have to be stored in order to compute Z_b (Eq. 5) as well as the

Memory Reduction Local vs. Global Flooding

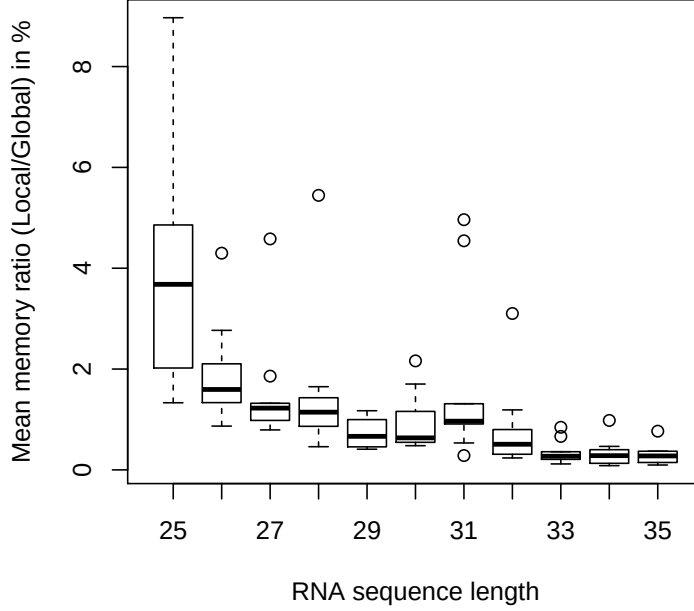


Figure 1: Memory consumption comparison of local vs. global flooding for the random sequence set. For each RNA sequence length, 10 mean ratios of local vs. global flooding memory requirement are measured and visualized in a box plot. The box covers 50% of the values and shows the median as horizontal bar. A similar picture is obtained when plotting the mean gradient basin size for each sequence.

set of all micro-state transitions leaving macro-state b , denoted $T(b)$, for computing $Z_{\{b,*\}}$ (Eq. 6). The memory consumption of local flooding of b is thus ruled by $\text{mem}(L) = |F^{-1}(b)| + |T(b)|$.

Investigating the ratio of $\text{mem}(L)/\text{mem}(G)$ for all macro-states, we find a mean value of 0.0015 and a median of < 0.0001 for both the d33 and the ire landscape. In other words, the memory footprint of local flooding comprises less than 0.005 (0.5%) compared to global flooding for almost all macro-states (99%). For approximately 80% of the macro-states, the footprint drops even lower to less than 0.01%. Similar numbers are observed within the random set for sequences of same lengths. Most notably, we see a logarithmic decrease of the average memory reduction with growing sequence length (see Fig. 1). We find only three large macro-states with $\text{mem}(L)/\text{mem}(G) > 10\%$ in both landscapes.

These numbers give evidence for the memory efficiency of a local flood-

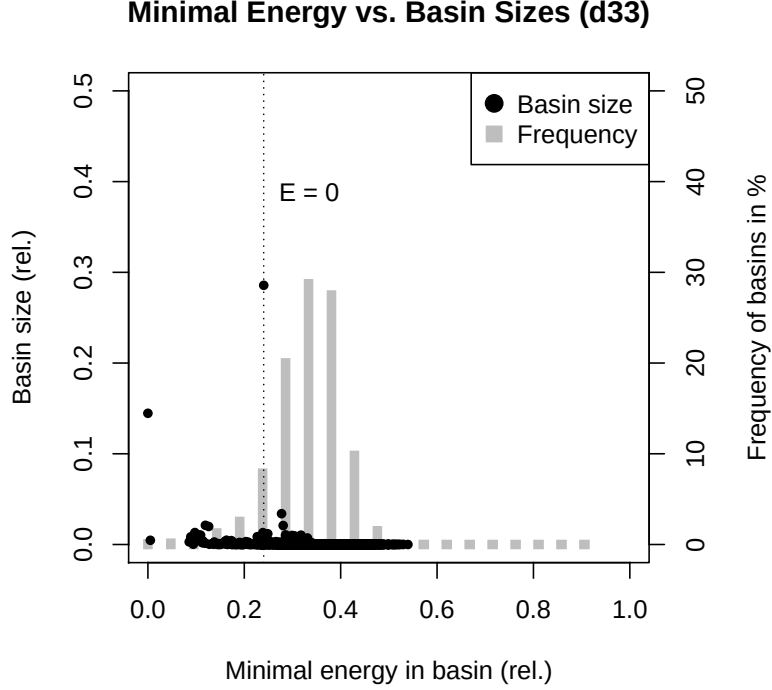


Figure 2: Distribution of basin sizes (dots) and frequency histogram of basins (bars) over the energy range within the energy landscape of RNA d33. Relative energies are given by $E_{\text{rel}}(x) = (E(x) - E_{\text{min}})/(E_{\text{max}} - E_{\text{min}})$ where $E_{\text{min}}/E_{\text{max}}$ denote the energy boundaries over X . The dotted line marks the position of the unstructured state with energy 0.

ing scheme. Within the context of extensive parallelization, such a scheme can be applied to large energy landscapes, since the individual memory consumption is several orders of magnitudes lower compared to a global flooding scheme. The remaining set of few large macro-states can be handled at the cost of longer runtimes by using the efficient local sampling scheme for macro-state transition probabilities presented in (Mann and Klemm, 2011).

4.3 Properties of gradient basins

In the following, we will work out various properties of gradient basins, since they are commonly used as macro-state abstraction in RNA energy landscapes. We will give examples for RNA d33, however the results can be generalized to other RNAs as shown for the random sequence set.

We have shown in the context of local flooding memory consumption that the overwhelming majority of gradient basins is small, whereas there are only

a few densely populated gradient basins. Most importantly, the basin of the open, unstructured state, which is a local minimum according to the Turner energy model (Xia *et al.*, 1998) and the selected neighborhood relation M allows for the largest neighborhoods. Consequently, its gradient basin is the largest for all RNAs studied and wraps about 20-30% of the state space. In the random data set, the open state covers on average approximately 40% of the landscape and we see a decrease of this fraction with increasing sequence length. The same tendency applies to the average relative basin size (see Fig. 1). Other large gradient basins are usually centered at energetically low local minima and their basin size is in general highly specific for the underlying sequence. We do observe a correlation of basin size with the energy of its local minimum (Spearman corr. -0.73), which is in accordance to the findings of Doye *et al.* (1998) for Lennard-Jones clusters.

When investigating the distribution of the energetically lowest micro-states in each gradient basin, *i.e.* the local minima, we find that most minima have positive energies (see histogram in Fig. 2). Minima are distributed over the lower 40-50% of the energy range for all sequences studied. The number of minima with negative energy, *i.e.* stable secondary structures, is approximately 100 for d33 and ire and is in the range of approximately 5% in general for the random set studied here. The majority of the state space of RNA energy landscapes shows positive energies, resulting from destabilizing energy terms for unstacked base pairs in the Turner energy model (Xia *et al.*, 1998). This is in accordance with the results from Cupal *et al.* (1997) who found that only $\sim 10^6$ of $\sim 10^{16}$ structures of a tRNA show an energy of less than zero.

The energy range of most gradient basins, *i.e.* minimal to maximal energy of any micro-state in the basin as plotted in Fig. 3, covers almost the entire range above a local minimum. This is generally independent of the basin size (compare Fig. 2 and 3), only for energetically high basins a lower maximal energy is observed. This might be a result of the accompanying basin size decrease or an artifact of the energy model. The gradient basin of the unstructured state covers the energetically highest states.

As mentioned above, only few of the possible $|B|^2$ macro-state transitions are observed. We find that more than 50% of the basins show less than 10 neighboring basins and almost all (98%) have transitions to less than 2% of the basins. The gradient basin of the unstructured state shows the highest number of macro-state transitions and is connected to more than 20% of the macro-states. We find that few large basins serve as hub nodes with high connectivity. This is in accordance to findings of Doye (2002) for Lennard-Jones polymers. Consequently, the number of transitions is highly correlated to the basin size, as one would expect. This is supported by a Spearman rank correlation coefficient of approx. 0.8 for all RNAs studied. The correlation to the basin’s minimal energy, as found by Doye (2002), is not as significant (Spearman corr. -0.6).

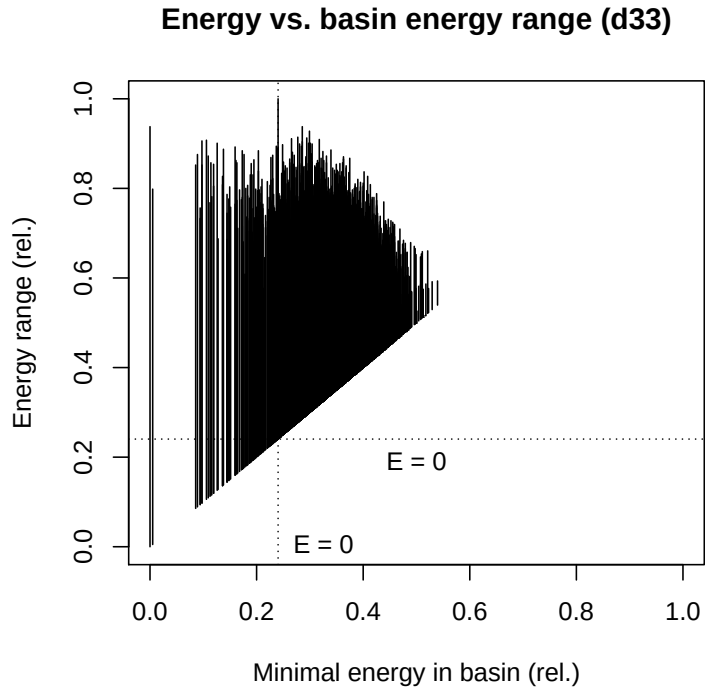


Figure 3: The energy range covered by each basin (Y-axis) sorted by the minimal energy within the basin (X-axis) over the whole energy range of the energy landscape of RNA d33. Relative energies are given by $E_{\text{rel}}(x) = (E(x) - E_{\text{min}}) / (E_{\text{max}} - E_{\text{min}})$ where $E_{\text{min}}/E_{\text{max}}$ denote the energy boundaries over X . The dotted lines mark the position of the unstructured state with energy 0.

5 Conclusion

We have introduced a local flooding scheme for computing the exact macro-state transition rates for arbitrary discrete energy landscapes provided some macro-state assignment is available. The approach has been evaluated on RNA secondary structure energy landscapes in the context of modeling coarse-grained RNA folding kinetics based on gradient basins. We have demonstrated the need for exact macro-state transition models via comparison to a simpler, barrier tree-based Arrhenius-like model. The latter resulted in significantly different dynamics measured by mean first passage times.

We showed that the local flooding scheme requires several orders of magnitude less memory compared to the standard global flooding scheme. Furthermore, it is intrinsically open to vast parallelization, which should also result in significant runtime reduction, given that the global flooding can not be easily parallelized.

Finally, we performed a thorough investigation of gradient basins in RNA energy landscapes, since they are commonly used as macro-state abstraction in the field. Gradient basins have been shown to be generally small, which is the reason for the tremendously reduced memory requirement of the local flooding scheme. The basin of the unstructured state has been shown to be special since it is the largest, most connected macro-state and covers the energetically highest micro-states. Independent of their size, most basins contain micro-states of almost the entire energy range above their respective local minimum. The majority of the gradient basins covers only states with positive energy. We found a low average connectivity between gradient basins, the existence of few highly connected hub nodes, and a high correlation of connectivity with basin size.

Acknowledgement

This work was partly funded by the Austrian Science Fund (FWF) project “RNA regulation of the transcriptome” (F43), the EU-FET grant RiboNets 323987, the COST Action CM1304 “Emergence and Evolution of Complex Chemical Systems” and by the IK Computational Science funded by the University of Vienna.

References

- Alemán, E. A., Lamichhane, R., and Rueda, D. (2008). Exploring RNA folding one molecule at a time. *Curr Opin Chem Biol*, **12**, 647–654.
- Aviram, I., Veltman, I., Churkin, A., and Barash, D. (2012). Efficient procedures for the numerical simulation of mid-size RNA kinetics. *Algorithms for Molecular Biology*, **7**, 24.

- Bogomolov, S., Mann, M., Voss, B., Podelski, A., and Backofen, R. (2010). Shape-based barrier estimation for RNAs. In *In Proceedings of German Conference on Bioinformatics GCB'10*, volume 173 of *LNI*, pages 42–51. GI.
- Chen, S.-J. (2008). RNA folding: Conformational statistics, folding kinetics, and ion electrostatics. *Annual Review of Biophysics*, **37**(1), 197–214.
- Cupal, J., Flamm, C., Renner, A., and Stadler, P. F. (1997). Density of states, metastable states, and saddle points exploring the energy landscape of an RNA molecule. In *Proc Int Conf Intell Syst Mol Biol.*, volume 5, pages 88–91. AAAI Press.
- Doye, J. P. K. (2002). Network topology of a potential energy landscape: A static scale-free network. *Phys. Rev. Lett.*, **88**, 238701.
- Doye, J. P. K., Wales, D. J., and Miller, M. A. (1998). Thermodynamics and the global optimization of Lennard-Jones clusters. *The Journal of Chemical Physics*, **109**(19), 8143–8153.
- Flamm, C. and Hofacker, I. L. (2008). Beyond energy minimization: Approaches to the kinetic folding of RNA. *Chemical Monthly*, **139**(4), 447–457.
- Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F., and Zehl, M. (2000a). Design of multi-stable RNA molecules. *RNA*, **7**, 254–265.
- Flamm, C., Fontana, W., Hofacker, I., and Schuster, P. (2000b). RNA folding kinetics at elementary step resolution. *RNA*, **6**, 325–338.
- Flamm, C., Hofacker, I. L., Stadler, P. F., and Wolfinger, M. T. (2002). Barrier trees of degenerate landscapes. *Z.Phys.Chem*, **216**, 155–173.
- Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T., and Turner, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the United States of America*, **83**(24), 9373–9377.
- Fürtig, B., Buck, J., Manoharan, V., Bermel, W., Jäschke, A., Philipp, W., Pitsch, S., and Schwalbe, H. (2007). Time-resolved NMR studies of RNA folding. *Biopolymers*, **86**(5-6), 360–383.
- Geis, M., Flamm, C., Wolfinger, M. T., Tanzer, A., Hofacker, I. L., Middendorf, M., Mandl, C., Stadler, P. F., and Thurner, C. (2008). Folding kinetics of large RNAs. *J. Mol. Biol.*, **379**, 160–173.
- Grinstead, C. M. and Snell, J. L. (1997). *Introduction to Probability*. American Mathematical Soc.
- Heuer, A. (2008). Exploring the potential energy landscape of glass-forming systems: from inherent structures via metabasins to macroscopic transport. *Journal of Physics: Condensed Matter*, **20**(37), 373101 (56pp).
- Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Chemical Monthly*, **125**, 167–188.
- Hofacker, I. L., Schuster, P., and Stadler, P. F. (1998). Combinatorics of RNA secondary structures. *Discr Appl Math*, **88**, 207–237.
- Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, **7**(4), 284–304.
- Kubota, M. and Hagiya, M. (2005). Minimum basin algorithm: An effective analysis technique for dna energy landscapes. In *DNA Computing*, volume 3384 of *LNCS*, pages 202–214. Springer Berlin Heidelberg.
- Kucharík, M., Hofacker, I. L., Stadler, P. F., and Qin, J. (2014). Basin hopping graph: A computational framework to characterize RNA folding landscapes. *Bioinformatics*. Accepted and online.
- Li, Y. and Zhang, S. (2012). Predicting folding pathways between RNA conformational structures guided by RNA stacks. *BMC Bioinformatics*, **13**(Suppl 3), S5.
- Lorenz, R., Flamm, C., and Hofacker, I. L. (2009). 2D projections of RNA folding landscapes.

- 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. *Algorithms Mol Biol*, **6**(1).
- Lorenz, W. A. and Clote, P. (2011). Computing the partition function for kinetically trapped RNA secondary structures. *PLoS ONE*, **6**(1), e16178.
- Mann, M. and Klemm, K. (2011). Efficient exploration of discrete energy landscapes. *Phys. Rev. E*, **83**(1), 011113.
- Mann, M., Will, S., and Backofen, R. (2007). The energy landscape library - a platform for generic algorithms. In *Proc. of BIRD'07*, volume 217, pages 83–86. OCG.
- Mann, M., Maticzka, D., Saunders, R., and Backofen, R. (2008). Classifying protein-like sequences in arbitrary lattice protein models using latpack. *HFSP Journal*, **2**(6), 396.
- Morgan, S. R. and Higgs, P. G. (1998). Barrier heights between ground states in a model of RNA secondary structure. *J Phys A: Math Gen*, **31**(14), 3153–3170.
- Noé, F. and Fischer, S. (2008). Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struc Biol*, **18**, 154–162.
- Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K., Lukasiak, P., Bartol, N., Blazewicz, J., and Adamiak, R. (2012). Automated 3D structure composition for large RNAs. *Nucleic Acids Research*, **40**(14), e112.
- Richter, A. S., Will, S., and Backofen, R. (2008). A sampling approach for the exploration of biopolymer energy landscapes. In *Proceedings of the European Conference on Metallobiology (HMI Berlin, Germany, 2007)*, pages 27–38. Herbert Utz Verlag, München.
- Rinntenthal, J., Buck, J., Ferner, J., Wacker, A., Fürtig, B., and Schwalbe, H. (2011). Mapping the landscape of RNA dynamics with NMR spectroscopy. *Acc Chem Res*, **44**(12), 1292–1301.
- Schön, J. C. and Sibani, P. (1998). Properties of the energy landscape of network models for covalent glasses. *J. Physics A: Mathematical and General*, **31**(40), 8165–8178.
- Sibani, P., van der Pas, R., and Schön, J. C. (1999). The lid method for exhaustive exploration of metastable states of complex systems. *Computer Physics Communications*, **116**(1), 17–27.
- Tang, W. and Zhou, Q. (2012). Finding multiple minimum-energy conformations of the hydrophobic-polar protein model via multidomain sampling. *Phys. Rev. E*, **86**(3).
- Tang, X., Kirkpatrick, B., Thomas, S., Song, G., and Amato, N. M. (2005). Using motion planning to study RNA folding kinetics. *J. Comp. Biol.*, **12**(6), 862–881.
- Tang, X., Thomas, S., Tapia, L., Giedroc, D. P., and Amato, N. M. (2008). Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.*, **381**(4), 1055–1067.
- Tinoco, I., Uhlenbeck, O. C., and Levine, M. D. (1971). Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.
- Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L., and Stadler, P. F. (2004). Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.*, **37**, 4731–4741.
- Wolfinger, M. T., Will, S., Hofacker, I. L., Backofen, R., and Stadler, P. F. (2006). Exploring the lower part of discrete polymer model energy landscapes. *Europhys. Lett.*, **74**, 726–732.
- Xia, T., SantaLucia, Jr, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**(42), 14719–35.

Supplementary Material

A Exact vs. approximated transition models

Figure 4 presents the Spearman rank correlation of the mean first passage times (FPT) for the different transition probability models studied. The plot is based on the random data set and grouped by sequence length.

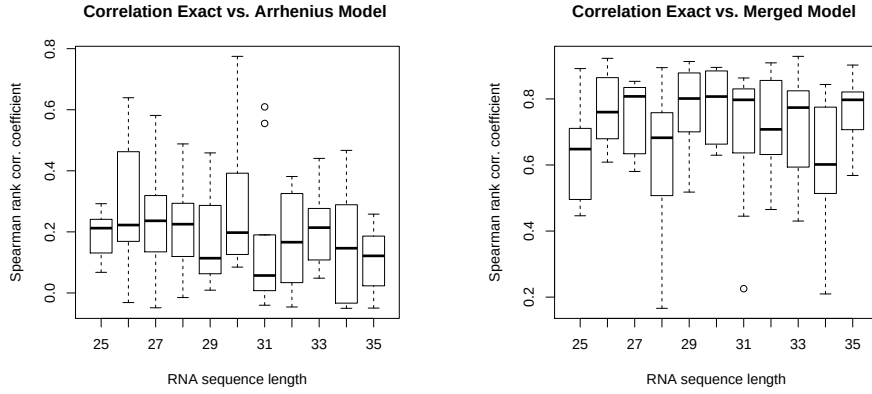


Figure 4: Spearman rank correlation coefficients of the mean first passage times (FPT) for the random data set grouped by sequence length. Correlation of the exact model (left) with the Arrhenius barrier-based transition model (right) and the merged transition probability model.

Figure 5 provides a visual comparison of coarse-grained folding dynamics for RNA d33, based on two different transition models. While the pure barrier tree dynamics (lower plot) resembles the overall dynamics of the two energetically lowest macro-states of the exact model (upper plot) quite well, it shows significant differences for states populated at lower extent (*e.g.* at rank 5 or 6).

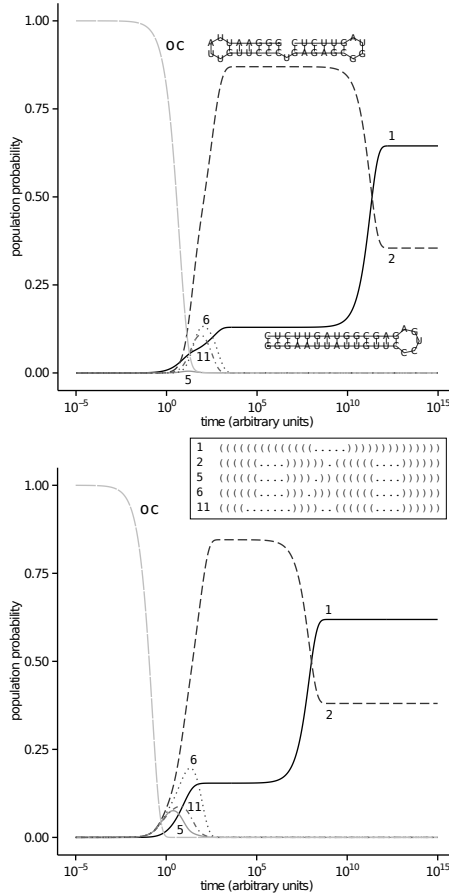


Figure 5: Coarse-grained folding dynamics of RNA d33 showing the five most populated gradient basins. Each curve represents the population probability of a gradient basin macro state, depicted by the secondary structure of its local minimum. Numbers correspond to energy sorted ranks. Simulations were started from the unstructured open chain macro-state (oc curve) and let evolve until a stationary distribution of the underlying Markov process was reached, see Wolfinger *et al.* (2004) for details. We compare the dynamics from exact transition probabilities (left) to those from a barrier tree-based Arrhenius transition model (right).

B Memory Consumption Local vs. Global Flooding

In Figure 6 on the left, we present the memory consumption of the local vs. the global flooding approach in terms of number of structures to be kept in memory for the random RNA sequence set. The local flooding requires several orders of magnitude less memory compared to global flooding. As expected, a growth in sequence length is visible.

The right side of Figure 6 presents the distribution of gradient basin sizes over the energy range for RNA d33. A decrease in basin size is observed with increasing minimal energy. A similar result was found in the context Lennard-Jones clusters by Doye *et al.* (1998).

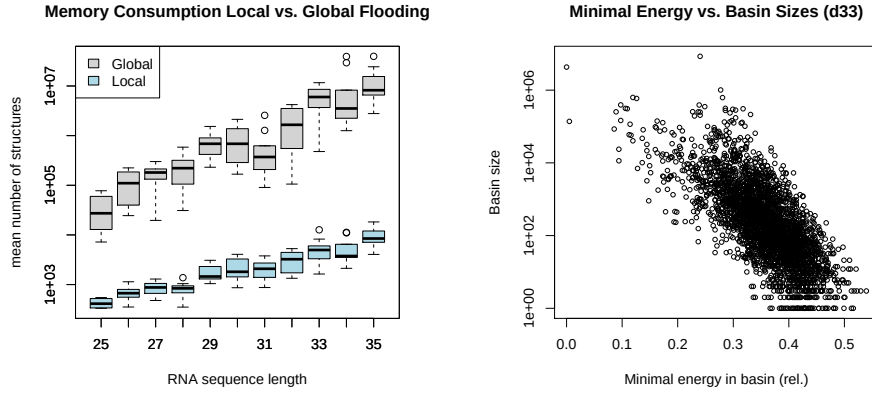


Figure 6: Memory consumption of global and local flooding for different RNA lengths within the random data set (left). Distribution of gradient basin sizes on a logarithmic scale over the energy range for RNA d33 (right).