Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

pourRNA - a time- and memory-efficient approach for the guided exploration of RNA energy landscapes

Gregor Entzian [1,*] and Martin Raden [2]

[1] University of Vienna, Faculty of Chemistry, Department of Theoretical Chemistry, Währingerstraße 17, 1090 Vienna, Austria, and [2] Bioinformatics Group, Department of Computer Science, University of Freiburg, 79110 Freiburg, Germany

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The folding dynamics of RNAs are typically studied via coarse-grained models of the underlying energy landscape to face the exponential growths of the RNA secondary structure space. Still, studies of exact folding kinetics based on gradient basin abstractions are currently limited to short sequence lengths due to vast memory requirements. In order to compute exact transition rates between gradient basins, state-of-the-art approaches apply global flooding schemes that require to memorize the whole structure space at once. *pourRNA* tackles this problem via local flooding techniques where memorization is limited to the structure ensembles of individual gradient basins.

Results: Compared to the only available tool for exact gradient basin based macro state transition rates (namely *barriers*), *pourRNA* computes the same exact transition rates up to ten times faster and requires two orders of magnitude less memory for sequences that are still computationally accessible for exhaustive enumeration. Parallelized computation as well as additional heuristics further speed up computations while still producing high quality transition model approximations. The introduced heuristics enable a guided trade-off between model quality and required computational resources. We introduce and evaluate a macroscopic direct-path heuristics to efficiently compute refolding energy barrier estimations for the co-transcriptionally trapped RNA sv11 of length 115 nt. Finally, we also show how *pourRNA* can be used to identify folding funnels and their respective energetically lowest minima.

Availability: *pourRNA* is freely available at https://github.com/ViennaRNA/pourRNA

Contact: entzian@tbi.univie.ac.at

Supplementary information: Supplementary data available at *Bioinformatics* online.

1 Introduction

The prediction of accurate RNA folding kinetics is still a computationally demanding problem despite decades of research. One of the main reasons is the exponential growth of an RNA's structure space with sequence length even in simple secondary structure models (Hofacker *et al.*, 1998). This growth directly relates to increasing runtimes when aggregating folding simulation statistics, e.g. done in Flamm *et al.* (2000) and Kirkpatrick *et al.* (2013). To compute exact dynamics, the master equation formalism can be used, but this typically requires a reduction of the system's size. To this end, a coarse-grained representation of the vast structural ensemble by a small set of macro-states with respective transition rates is used. Different macro-state definitions have been introduced, e.g. based on static properties of the respective energy landscape (Wolfinger *et al.*, 2004) or correlations of transition rates (Zhang and Chen, 2003).

A standard approach to compute exact macro-state transition rates, e.g. implemented in *barriers*, requires a (partial) structure space enumeration below a reasonable energy threshold (Wolfinger *et al.*, 2004). This procedure is currently limited to short sequence lengths up to 100 nt due to vast memory requirements.

To bypass this problem, different strategies to sample the structure space are combined with transition rate estimations. For instance, the recent Basin Hopping Graph (*BHG*) framework samples structures and

local minima and subsequently estimates the minimal energy barrier between them via direct folding pathway computations (Kucharík *et al.*, 2014). While this enables kinetics predictions for longer RNAs, the procedure might miss transition states (Kucharík *et al.*, 2014) and the dynamics can be distorted due to poor transition rate estimates based on saddle heights only (see supplementary material F).

Here, we introduce *pourRNA* for both transition rate estimation as well as local minima exploration of RNA secondary structure energy landscapes. The approach makes use of local flooding techniques as introduced by Mann *et al.* (2014). That is, starting from a given local minimum, the respective gradient basin macro-state is explored. This provides all data needed to compute exact transition rates to the (on-the-fly) identified neighbored macro-states. So far unexplored neighbored basins are queued for flooding and the process iterates until the queue is empty. Local flooding makes the approach (a) memory efficient, since only small parts of the structure space have to be memorized, and (b) enables parallelized macro-state processing, which results in much lower runtimes compared to global flooding approaches while still computing the same exact transition rates.

Besides the computation of exact transition rates, *pourRNA* also offers heuristics to further reduce its computational requirements while still providing highly accurate rates. This is mainly achieved by additional constraints on local flooding and results in a controlled trade-off between runtime and prediction quality. Furthermore, we can guide whether or not neighbored macro-states are processed with additional filters. This enables us to explore only specific parts of the energy landscape that are reached via the fastest folding pathways, which are related to folding funnels as defined by Klemm *et al.* (2008). Therein, a funnel is defined as a second level macro-state that fuses neighbored gradient basins based on a gradient that follows the highest macroscopic transition rate to a lower basin. We show that *pourRNA* can compute the mapping of individual structures or gradient basins to their folding funnels, since its heuristics can be set to compute macroscopic gradient walks following Klemm *et al.* (2008).

Finally, we introduce a new macro-state-based direct path heuristic to identify the energy barrier between two RNA structures. Besides its general application to study refolding pathways of RNA molecules, e.g. from meta-stable states into the global optimum, such heuristics are important to handle incomplete transition information. The latter is common when only the lower part of the energy landscape is computationally accessible.

We evaluate *pourRNA* both concerning its technical benefits in terms of runtime and memory consumption as well as its potential for high-quality rate models for a known set of RNAs from the literature. This is done both for the computation of exact transition rates as well as using the newly introduced heuristics.

2 Formal Preliminaries

Within the following, we shortly introduce *RNA energy landscapes* representing non-crossing secondary structures defined by the triple $(\mathcal{P}, N, E)$. The *structure space* $\mathcal{P}$ comprises all *non-crossing secondary structures* P for a given RNA sequence considering Watson-Crick and G-U base pairs. A structure is non-crossing, or nested or pseudoknot-free, if any two base pairs are either enclosing distinct subsequences or one is enclosing the other. The symmetric *neighborhood* $N(P)$ comprises the set of all structures $P' \in \mathcal{P}$ that differ in exactly one base pair from P, i.e. it holds that $P' \in N(P)$ if one can transform P into P' by inserting or deleting a single base pair. The (free) *energy* $E(P)$ of a structure in units of $\frac{kcal}{mol}$ is determined by the Nearest Neighbor Model (Tinoco *et al.*, 1973) using the parameters from Mathews *et al.* (2004). A structure P' is considered '*energetically smaller*' than P if its energy is either lower ($E(P') < E(P)$) or equal but its structure dot-bracket string encoding is lexicographically smaller, which is needed due to the degeneracy of the energy model (see Flamm *et al.* (2002) for further details). A structure is called *local minimum* $\check{P}$ if it is energetically smaller than all its neighbors. Detailed definitions and formalisms are provided in the supplementary material A.

Following Wolfinger *et al.* (2004), we partition $\mathcal{P}$ into gradient basin macro-states. Each gradient basin is associated with a single local minimum structure $\check{P}$ of the energy landscape. The *gradient basin* $B(\check{P}) \subseteq \mathcal{P}$ is defined recursively and contains all structures whose gradient neighbor is within B, where the *gradient neighbor* (if existent) is the smallest among all energetically smaller neighbors. Therefore, any local minimum $\check{P}$ does not have a gradient neighbor and is thus the minimal energy structure of a basin. The set of basins $\mathcal{B}$ thus comprises a partitioning of $\mathcal{P}$ and we denote with $b \in \mathcal{B}$ in short a basin for some local minimum. Two *basins* $b \neq b'$ *are neighbored* if two of their respective structures are neighbored. The *energy of a basin* is given by its ensemble energy $E(b) = -RT\log(Z_b)$, where Z_b *denotes its partition function* given by $\sum_{P\in b} w(P)$, i.e. the sum of Boltzmann weights $w(P) = \exp(-E(P)/RT)$ for gas constant R and temperature T (fixed to 37°C in this study).

Folding dynamics of an RNA are typically studied as a Markov process on $\mathcal{P}$ (Flamm and Hofacker, 2008) for which appropriate transition rates between the structure-representing states have to be defined. For kinetics on $\mathcal{P}$, in the following referred to by *micro-states*, transition rates are typically defined by Metropolis rates. That is for two neighbored structures P, P' (with $P{\in}N(P')$) the *transition rate* $k_{P\to P'}$ *from* P *to* P' is given by $\min\left(1, \exp(-(E(P') - E(P))/RT)\right)$; all other rates are 0. The *macroscopic transition rates on* $\mathcal{B}$ are aggregated from microscopic ones under the assumption that each basin is in thermodynamic equilibrium (and thus observing a structure $P{\in}b{\in}\mathcal{B}$ is given by its Boltzmann probability $Pr[P|b] = w(P)/Z_b$). For a transition from a basin b to one of its neighbors $b'{\in}N(b)$ the rate is therefore $k_{b\to b'} = \sum_{P\in b}\sum_{P'\in b'} Pr[P|b]k_{P\to P'}$, i.e. the weighted sum of inter-basin micro-state transitions.

The vast majority of structures from $\mathcal{P}$ have positive energy values (see (Lou and Clote, 2010) for an illustration) and thus extremely low probabilities $Pr[P|\mathcal{P}]$ within the structural ensemble. Therefore, we restrict w.l.o.g. the structure space to $\bar{\mathcal{P}} \subseteq \mathcal{P}$ via an *absolute upper energy threshold* of 5 $\frac{kcal}{mol}$ (i.e. $P{\in}\bar{\mathcal{P}} \leftrightarrow E(P){<}5\ \frac{kcal}{mol}$). Besides reducing the state space, this threshold also ensures that the probable parts of the landscape are represented and still connected, as discussed in the supplementary material B. A similar effect is obtained when excluding structures with unstacked base pairs from $\mathcal{P}$ in combination with an appropriate neighborhood definition as e.g. done by Kirkpatrick *et al.* (2013), since such base pairs are heavily penalized within the energy model.

Before discussing how macro-state transition rates are computed, we first point out that they can be expressed as $k_{b\to b'} = \hat{Z}_{\{b,b'\}}/Z_b$, where $\hat{Z}_{\{b,b'\}}$ denotes the sum of the minimal Boltzmann weights $(\min\left(w(P), w(P')\right))$ of all microscopic inter-basin transitions $P{\in}b \to P'{\in}b'$ with $P{\in}N(P')$ as shown in Mann *et al.* (2014), i.e.

$$\hat{Z}_{\{b,b'\}} = \sum_{P\in b}\sum_{P'\in N(P)\cap b'} \min\left(w(P), w(P')\right). \qquad (1)$$

We call the set of states contributing to $\hat{Z}_{\{b,b'\}}$ the *transition-state ensemble between* b *and* b', which is direction independent. Given this, the computation of macro-state rates reduces to the problem of (i) computing the partition function Z_b for each basin $b{\in}\mathcal{B}$ and (ii) the identification of all transition-state ensembles to derive respective $\hat{Z}_{\{b,b'\}}$ values.

3 Methods

In the following, we first shortly introduce the current standard approach to compute exact transition rates before introducing our *pourRNA* approach.

3.1 State of the Art – global flooding

barriers, introduced by Wolfinger *et al.* (2004), was the first tool for RNA research to compute exact macro-state transition rates up to a given energy threshold. The approach is based on two preliminaries: (1) the recursive definition of gradient basins and (2) the possibility to enumerate the structure space ordered by ascending energy. The latter was made possible for RNAs by the work of Wuchty *et al.* (1999), implemented in *RNAsubopt* from the Vienna RNA package (Lorenz *et al.*, 2011), and recently extended by Stone *et al.* (2015).

For each structure from an energy-sorted input P, *barriers* identifies its gradient neighbor $g(P)$. If none exists, P is a local minimum and thus is assigned to a new gradient basin b. Otherwise, P is assigned to the basin b of its gradient neighbor, which was already processed due to the ascending energy input. Furthermore, the partition functions Z_b and $\hat{Z}_{\{b,b'\}}$ are updated. The latter is done for all neighbors of P (not within b) with lower energy since their gradient basin assignment b' is known and that way no microscopic transition is counted twice.

This global flooding approach requires the memorization of *all* processed input structures. The exponential growth of the structure space thus makes it infeasible for sequences longer than 100 nt (Geis *et al.*, 2008). When only a low energy subspace up to some upper energy bound is enumerated, it is not ensured that the resulting (partial) basins are all transitively connected via identified micro-state transitions, which is needed for an exact Markov model.

A similar approach, *lid* by Sibani *et al.* (1999), requires no presorted input but faces the same memory problem. While *barriers*'s processing can be imagined as a rising global ground water level within the whole landscape, *lid* implements a local 'flood-and-overflow' strategy where transitions to neighbored basins directly trigger (recursively) their flooding until the basins' flooding levels are balanced again. When the method 'flows over the lid' into an unknown neighbored basin, a gradient walk is applied to identify the respective local minimum. Eventually, the *lid* approach can result in very distorted and incomplete transition rate estimates of the landscape when the available memory limit is exceeded too early. To our knowledge, no implementation for RNA energy landscapes is available. Thus, we restrict our comparison in the following to the *barriers* pipeline.

3.2 pourRNA – asynchronous, exhaustive, local flooding

Here, we introduce *pourRNA* that solves the memory problems of the global flooding methods. It fuses the explorative idea of the *lid* method with the memory-efficient local flooding approach introduced by Mann *et al.* (2014), which enables a controllable low-memory approach intrinsically open to asynchronous parallelization.

Local flooding enables the efficient computation of the partition function Z_b of a given basin b. To this end, a priority queue T of unprocessed structures (sorted by energy, see Sec. 2) is initialized with the neighbors of the local minimum of the basin. The minimum itself initializes the basin's list (or priority queue) D of already processed states. As for global flooding, iteratively the (energetically smallest) top element P from T is extracted and its neighbors are investigated. If its gradient neighbor is within D, it is part of the basin and thus Z_b is updated, all its neighbors with higher energy are added to T. In addition, partition functions $\hat{Z}_{\{b,b'\}}$ of macroscopic transitions are updated using hashed gradient walks if the gradient neighbor is unknown (and thus part of a neighbored basin b'). Given that this is only applied to the 'surface' of the current basin, gradient walk computations can be accelerated by storing the gradient neighbor relation of states observed along the gradient walks, which provides improvements for gradient walks with equal tails. Finally, P is added to D. Since D only contains information about the currently processed basin b, the identification of neighboring b' by means of an inter-basin micro-state transition $P \to P'$ (with $P \in b$ and $P' \in b'$) has to be determined by a gradient walk starting in P'. For further details, please refer to Mann *et al.* (2014).

Given this, the overall workflow of *pourRNA* for a given set of arbitrary input structures $\mathcal{P}_i \subseteq \mathcal{P}$ can be sketched as follows:

- initialize list of processed basins $\mathcal{B}_D = \emptyset$
- initialize list of unprocessed basins $\mathcal{B}_T$ (local minimum information) via gradient walks for all input structures from $\mathcal{P}_i$

1. extract next basin $b \in \mathcal{B}_T$
2. run local flooding for b
3. store Z_b and $\hat{Z}_{\{b,b'\}}$
4. update $\mathcal{B}_T$ with all newly identified neighbored basins b'
5. add b to $\mathcal{B}_D$ and mark as processed
6. go to (1.) if $\mathcal{B}_T \neq \emptyset$

As shown by Mann *et al.* (2014), the vast majority of gradient basins is small. Since this approach exhaustively processes all basins $b \in \mathcal{B}$ in an iterative scheme, *pourRNA* can compute all exact transition rates with the low memory requirements of local flooding.

In addition, this approach is intrinsically open to parallelization since local flooding of b is independent of the processing of any other basin. Only the update of $\hat{Z}_{\{b,b'\}}$ and $\mathcal{B}_T$ has to be synchronized. As soon as new neighbored basins are identified, their processing can be started asynchronously.

3.3 pourRNA – exploration heuristics

As already discussed, the RNA structure space and thus the size of the respective energy landscape grows exponentially in sequence length. Since the vast majority of RNA structures is energetically unfavorable (positive energy values) and thus unlikely to be formed, it is reasonable to restrict the energy landscape exploration to low energy conformation. In the following, two such heuristics implemented in *pourRNA* are discussed.

3.3.1 maxE – global absolute energy bound

A first and obvious way to constrain exploration, also e.g. available within the global flooding *barriers* pipeline, is to ignore RNA structures above a *global upper absolute energy bound (maxE)*, also discussed e.g. in (Hofacker *et al.*, 2010; Kucharík *et al.*, 2014). Due to the bell-shaped density-of-states distribution of RNAs (see e.g. (Lou and Clote, 2010)), a low maxE value will (i) ignore the vast majority of the structure space while (ii) still approximating the partition function (ensemble weight) with high precision. The latter results from the discussed inverse exponential relationship of energy terms and Boltzmann weights w. Eventually, the probability $(w(P)/Z)$ to observe a respective structure P drops exponentially with increasing energy.

When initialized with a single input structure, *pourRNA* identifies its respective local minimum via gradient walk computation and explores the accessible gradient basins (transitively) neighbored to the start basin. While this will eventually result in the full basin partitioning of the energy landscape for unconstrained exploration, *pourRNA* will produce only the cluster of accessible basins if the structure space is restricted by an upper absolute energy bound. In order to be exhaustive under such constraints, one can efficiently enumerate local minima via dynamic programming as introduced by Clote (2005); Lorenz and Clote (2011) or approximate results via sampling following, for example, Ding and Lawrence (2003); Lorenz and Clote (2011); Kucharík *et al.* (2014) or Michalik *et al.* (2017). Also we can initialize *pourRNA* with structure samples from co-transcriptional folding simulations (Danilova *et al.*, 2006; Hofacker *et al.*, 2010; Proctor and Meyer, 2013), since they resemble the structure space that initializes and thus guides subsequent global folding. A well studied example is the RNA `sv11` (Biebricher and Luce, 1992) also used for benchmarking within the Results section.

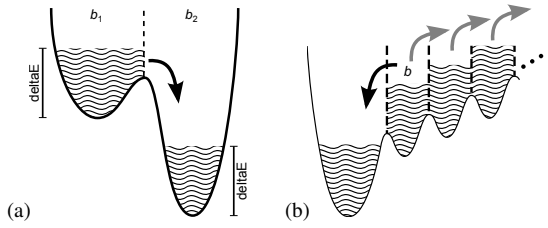

Fig. 1. (a) When restricting local flooding to a deltaE range above the local minimum, macro-state transitions might only be identified in one direction (arrow). (b) depiction of the "stair climb effect" (gray arrows; exploration starts in basin b), which can be undesired when interested in the kinetics following high rates to more stable basins (black arrow).

3.3.2 deltaE – local relative energy bound

In contrast to global flooding, *pourRNA* can also restrict the local flooding of individual macro-states by the *deltaE heuristic*. If enabled, only structures within the energy range of deltaE above the respective local minimal energy are investigated whether they are part of the gradient basin or mark a transition to a neighbored basin. This efficiently restricts the memory consumption of *pourRNA*, since basins typically cover a wide energy range (and also show a density-of-states distribution similar to the global one, see Mann *et al.* (2014)).

Since we are ignoring many (high energy) micro-states depending on the local flooding level defined by deltaE, the resulting basin's partition function Z_b is only an estimate of the true partition function. But since Z_b is dominated by the local minimum (and energetically close structures), highly accurate approximations can already be gained by low deltaE thresholds (see supplementary material E). Nevertheless, energetically lower basins will show larger effects than high energy basins, since we apply the same relative deltaE threshold to all.

Furthermore, since we ignore high energy micro-state transitions, we might miss respective macro-state transitions as well. This problem is partially amended by the *pourRNA* pipeline when both macro-states are processed. In that case, it is likely that micro-state transitions from the energetically higher to the lower basin are found (see Fig. 1(a)). Since $\hat{Z}_{\{b,b'\}}$ is symmetric, identification from one side is sufficient. If both directions are processed (independently), *pourRNA* stores the higher $\hat{Z}_{\{b,b'\}}$ value, since it resembles the better approximation. Note, when using the deltaE heuristics, exploration has to start at high energy basins. Otherwise, unfavorable "up hill" transitions and thus whole sublandscapes can be missed (see Fig. 1(a)).

The deltaE bound implies a maximal energy barrier for considered exit pathways for the basin. That is, deltaE equals the maximal energy difference between the local minimum $\check{P}$ and the highest transition state $\hat{P}$ and thus the smallest transition rate can be estimated by the Arrhenius equation, i.e. $\exp(-\text{deltaE}/RT)$ (see supplementary material E). For instance, a deltaE value of +5 $\frac{kcal}{mol}$ corresponds to a minimal transition rate of about 0.0003 to be considered for landscape exploration. Thus, depending on the deltaE value, the resulting sublandscape models the "fast folding" sublandscape (structure space) accessible for the molecule for the given start structure and can be used to study refolding processes as we exemplify later.

3.3.3 maxNeighE and kBest – gradient walks on macro-states

Gradient basins are not the most abstract representation used to study energy landscapes and folding kinetics. Folding funnels represent an even higher level of abstraction and cluster macro-states by their separating lowest barriers (to more stable structures) (Leopold *et al.*, 1992; Frauenfelder and Leeson, 1998; Karplus, 2011). Klemm *et al.* (2008) have introduced a formal framework building on gradient basins. Therein, the lowest barrier (and thus the highest macroscopic transition rate) to a neighbored basin with lower local minimum defines recursively a funnel partitioning of the energy landscape. The funnel assignment is eventually based on *gradient walks on macro-states*.

pourRNA can be used to identify such macroscopic gradient neighbors to compute the overall funnel partitioning or to identify the respective local minimum of the funnel. To this end, *pourRNA* allows to filter the set of neighbored basins that are considered for further exploration (step 4 in the *pourRNA* algorithm).

For a given gradient basin b, the *kBest filter* allows to restrict the neighbored basin exploration to the kBest basins b' with highest macro-state transition rates $k_{b\to b'}$, independently of the absolute rate values. Since the rate is inversely related to the respective energy barrier on folding paths, the highest rate corresponds to the lowest energy barrier.

The *maxNeighE filter* prunes all neighbored basins b' for which the energy difference of the respective minima $E(\check{P}') - E(\check{P})$ is below the user defined threshold. That is, if maxNeighE is set to 0 or below, only more stable (lower energy) and thus kinetically favored basins are considered for further exploration (this corresponds to the black arrow in Fig. 1(b)).

To compute a gradient walk on macro-states, we simply set maxNeighE=0 (to ensure "down hill climbs") and kBest=1 (to follow only the highest rate; assuming it to be unique). This is similar to ideas of Kühnl *et al.* (2017). Note, the order of the filters is important, since the highest rates are not necessarily leading to energetically lower neighbored gradient basins. Macroscopic gradient walks are of interest when studying higher level energy landscape organizations like folding funnels (Klemm *et al.*, 2008), which is discussed in detail in supplementary material H.

The combination of both filters is a powerful tool to restrict the search space of *pourRNA* when investigating fast refolding kinetics. For long RNAs, even the restriction of the maximal absolute energy (maxE) as well as the local flooding boundary (deltaE) will result in very large macro-state models of the landscape. Thus, we can face a "stair climb effect" during the exploration (depicted by gray arrows in Fig. 1(b)) that might be undesired when e.g. studying fast refolding pathways to more stable structures. This problem can be mitigated by *pourRNA*, since the consideration of "less stable" basins can be limited or excluded (via maxNeighE) and we can focus on the fastest folding routes (controlled by kBest).

3.3.4 Direct paths on macro-state level

When multiple input structures are provided and the exploration is constrained by maxE or deltaE, the resulting transition rate model can be non-ergodic, i.e. some (clusters of) macro-states are not (transitively) connected. The same applies to the state-of-the-art *barriers* pipeline, since it covers all local minima (and respective macro-states) below the maxE energy bound to which *pourRNA* input structures are initially mapped. To compute folding kinetics, additional post-processing is needed to heuristically estimate (high energy) transitions between connected macro-state clusters.

Kucharík *et al.* (2014) discuss and evaluate various ways how macro-state clusters can be connected. One of the first approaches was introduced by Morgan and Higgs (1998) and considers only direct paths of micro-states. That is, given a start and target structure P and P', resp., only structures $P_x \subseteq (P \cup P')$ that show a combination of start/target base pairs are considered (while keeping the shared base pairs, i.e. $(P \cap P') \subseteq P_x$). In the following, we refer to trajectories $P..P_x..P'$ via such structures as *microscopic direct paths*. *findpath* implements a fast bounded breadth-first search within this structure space (Flamm *et al.*, 2001), which is employed by Kucharík *et al.* (2014) to connect macro-states and to identify the interjacent basins. The latter is supported by local flooding techniques to optimize the barrier estimation.

Here, we generalize the notion of *direct paths to macro-states*, which is so far, to our knowledge, only defined for micro-states as given above.

To this end, we first map start and target structure to their respective local minima $\check{P}$ and $\check{P}'$. Subsequently, we restrict *pourRNA*'s landscape exploration to basins with local minima $\check{P}_x$ along direct paths between the $\check{P}$ and $\check{P}'$ (analogously to the microscopic direct path definition). To allow for some variations (e.g. due to helix extensions within the local minima of basins), we allow for some (small) base pair deviation Δ_{bp} from microscopic direct paths, i.e. $|\check{P}_s \setminus (\check{P} \cup \check{P}')| < \Delta_{\text{bp}}$. This macroscopic path exploration can be combined with the introduced heuristics to speed up the search.

Note, in contrast to the approach of Kucharík *et al.* (2014), we are not biased to macro-states along the microscopic direct path optimized by *findpath*. Note further, since we are only constraining the base pair distance of local minima and not the micro-states considered during local flooding, our macroscopic direct paths will explore a larger structure space compared to microscopic direct-path search.

4 Results and Discussion

4.1 Data set

For analyzing and benchmarking *pourRNA*, RNAs used in other RNA kinetics studies of varying lengths were extracted from literature. The shorter molecules were used to compare to exhaustive methods, while we exemplify the use of *pourRNA*'s heuristics for longer RNAs. The supplementary material C lists all RNAs with sequence and meta information. Unless stated differently, *pourRNA* uses the unstructured open chain as initial state for its exploration.

4.2 Exact transition rate computation

In the following, we compare *pourRNA*'s runtime and maximal memory consumption to the state-of-the-art *barriers* pipeline using the 56 nt `SL` RNA. Both approaches are considering only structures with an absolute energy below $5\ \frac{kcal}{mol}$. The *barriers* approach required in total 45 min (*RNAsubopt+barriers*) with a memory peak of 8.3 GB; both numbers define the reference for subsequent comparisons. Note, this pipeline cannot profit from multi-threading and requires a sorted list of all suboptimal structures up to a certain threshold (computed by the tool *RNAsubopt*).

In order to benchmark *pourRNA*'s computational performance for *exact kinetics studies, no additional heuristics* (beside maxE) are applied. Thus, *pourRNA* produces exactly the same macroscopic transition rates compared to the *barriers* pipeline. Single-thread computations with *pourRNA* (solid lines in Fig. 2(a)) took 6.3 min and 60 MB; using 10 threads reduced the runtime to 3.6 min but increased maximal memory to 190 MB (averages for two repetitions). We observe a non-linear correlation of the number of used threads and runtime, which converges already for about 4 threads. In contrast, memory consumption almost linearly increases with thread number (see also supplementary material D).

To investigate the source of the runtime improvement, we first recap that the *RNAsubopt+barriers* pipeline can be decomposed into three steps: (i) unsorted structure enumeration via dynamic programming, (ii) sorting of the structures, and (iii) computation of the macroscopic transition rates, where (i)+(ii) are done by *RNAsubopt*. Given that the latter requires less than 0.5 min for sorted enumeration of the considered `SL` structure space, the global flooding of *barriers* defines the overall runtime that is mainly governed by microscopic rate computations. Furthermore, *barriers* implements its own neighborhood routines; *pourRNA* uses methods from the recent Vienna RNA package v2.4.10. While this might explain some runtime difference, we assume that the speedup mainly stems from the different rate computation implementations used in *pourRNA*. Finally, *barriers* enumerates a large number of high energy basins that are not connected to the minimum free energy basin and are irrelevant for the overall

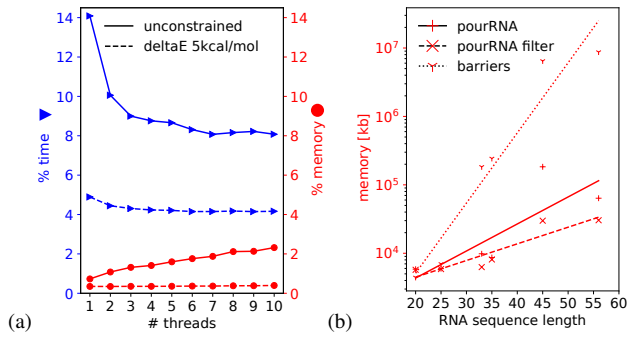

Fig. 2. (a) Averaged relative time (triangle) and memory (circle) consumption of our pourRNA approach (1-10 threads) compared to RNAsubopt+barriers (1 thread) for `SL` RNA. See text for absolute time and memory values.
(b) Memory consumption of the tools barriers and pourRNA for RNA sequences of different length. Each marker corresponds to the memory size for the tool that processed one or more sequences of the given length. Each tool uses 1 thread and pourRNA applies also the delta energy micro state filter, which additionally decreases the memory consumption, indicated by the trend lines. For barriers the line is dotted, for pourRNA it is solid and for pourRNA with deltaE=5 filter it is dashed.

folding dynamics. Thus, *pourRNA* benefits from the reduced memorization (and thus allocation and hashing efforts) needed for local compared to global flooding. The fast convergence results from a few large basins that consume most of the time (compare basin size statistics by Mann *et al.* (2014)). While all other basins are smaller and already completed, the large basins are still in the flooding process.

Figure 2(b) compares the trends of peak memory consumption for increasing sequence lengths using *barriers* and *pourRNA*. The consumption of *pourRNA* is much lower, because it holds only the structures for one basin in memory.

4.3 Parameter optimization for the deltaE heuristic

To assess the quality of the kinetic models produced with *pourRNA*'s deltaE heuristic, we investigate the overall partition function $Z = \sum_{b \in \mathcal{B}} Z_b$, the ensemble energy $E_{\text{all}} = -RT \log(Z)$, and the number of identified macro-state transitions for different deltaE values (see supplementary material E). The closer these values are to the exact values (without additional heuristics), the better is the approximation. For deltaE values above $5\ \frac{kcal}{mol}$, we observe no significant difference. For values below 4-5 $\frac{kcal}{mol}$, the explored sublandscape is increasingly less representative. This can be seen both at the decrease of relative Z as well via the strong reduction of the number of identified macro-state rates, which is observed independently of the sequence lengths. Given our experiments, we thus consider a deltaE value of $5\ \frac{kcal}{mol}$ (or slightly above) as a useful threshold for local flooding restrictions to provide a good balance of time consumption and the quality of the resulting kinetic model.

4.4 Approximate transition rates using the deltaE heuristic

Here, we reevaluate the performance of *pourRNA* compared to the *barriers* pipeline when restricting local flooding with deltaE = $5\ \frac{kcal}{mol}$. Results are depicted by dashed lines in Fig. 2. By applying the deltaE filter, both runtime and memory consumption are significantly further reduced (compare Fig. 2(a)). However, memory requirements still grow slowly exponentially with the sequence length (i.e. a linear trend in log scale plotting in Fig. 2(b)).

Since *pourRNA*'s runtime is dominated by the processing of the largest basins, restricting this local flooding results in both time and memory reduction. As can be seen from Fig. 2(a), this leads to an even faster convergence of the time requirement. Notably, memory requirement becomes

nearly constant for different thread numbers when deltaE is applied. This results from the extremely small basin fractions that are enumerated for the given deltaE bound.

To evaluate the quality of the approximated transition rates produced with the deltaE heuristic, we computed population trajectories using *treekin* (Wolfinger *et al.*, 2004) and compared them to the kinetics when using exact macroscopic rates as computed by *pourRNA* without filters (or *barriers*). The resulting kinetics are depicted within the supplementary material F and show no visual difference for exact and approximate rates.

4.5 Energy barrier estimation of RNA refolding

As discussed already, co-transcriptional folding can guide the folding process to meta-stable conformations that are structurally quite different from the energetic optimum (global minimum free energy structure). A well studied example is the `sv11` RNA that is kinetically "trapped" in a local minimum (see supplementary material G.1) due to co-transcriptional substructure formation (Biebricher and Luce, 1992). To study the refolding process from the kinetic trap to the global optimum (which is dominant in thermodynamic equilibrium), estimates of the relative energy barrier that needs to be overcome are needed. The energy barrier is defined as the maximal energy $E(\hat{P})$ of the transition state $\hat{P}$ on the energetically lowest (microscopic) path that connects the meta-stable start state with the global optimum (see e.g. (Wolfinger *et al.*, 2004)). The Arrhenius equation inversely relates the energy difference to the barrier (i.e. the activation energy within a high-level abstraction) with the respective transition rate on an exponential scale. Thus, the energy barrier has to be identified as well as possible to enable accurate rate and timing assessment of the refolding.

Since the energy landscape of `sv11` is already too large for global flooding (Kucharík *et al.*, 2014), so far only estimates of the refolding energy barrier are known. Kucharík *et al.* (2014) have shown for `sv11` that the energy barrier estimates from microscopic direct path optimization (using *findpath*) provide only a rough approximation. Also applying heuristics implemented in *PathFinder* (Lorenz *et al.*, 2009) or *TabuPath* (Dotu *et al.*, 2009) yield no improvements. They could show via their *BHG* sampling approach that a "detour" within the energy landscape improves the barrier estimates (compare *findpath* and *BHG* in Tab. 1). In the following, we will study this refolding problem using *pourRNA* (always using up to 8 cores).

To emulate a proper experimental setup, we first "rediscovered" the meta-stable conformation known from the literature using the co-transcriptional folding simulation webserver *KineFold* (Xayaphoummine *et al.*, 2005) (see supplementary material G.1). This provides us with the start structure for our explorative local flooding approach.

Refolding along fast macro-state transitions

First, we investigate the energy landscape accessible via fast macro-state transitions using the deltaE heuristic. To enable comparisons with results from (Kucharík *et al.*, 2014) we are using the `turner-99` energy parameters from the Vienna RNA package 2.4.10 (see supplementary material G.3 for `turner-04` results). We also use an extended micro-state neighborhood definition (as done for *BHG* by Kucharík *et al.* (2014)) that also connects structures that differ in exactly one position in one base pair, which is called a shift move. Note, so far shift moves are not available for *findpath*.

Since we know already an upper bound on the energy barriers from the microscopic path optimization (see *findpath* in Tab. 1), we set maxE to -56 $\frac{kcal}{mol}$. Starting from the meta-stable conformation, we cannot reach the global minimum free energy structure using a deltaE value of 4 $\frac{kcal}{mol}$. Using deltaE=5 $\frac{kcal}{mol}$, *pourRNA* explores a cluster of 143,032 gradient basins within 6.3 h of computation time. The cluster includes the macro-state of the global minimum free energy structure. Thus, using a modified Dijkstra algorithm, we find an energy barrier estimation of -62.3 $\frac{kcal}{mol}$ (compare Tab. 1). The respective refolding barrier relates to the path identified by Kucharík *et al.* (2014), as discussed within the supplementary material G.2.

method (parameters)	$E(\hat{P})$	time	$\lvert\mathcal{B}_p\rvert$
findpath (width=1000)	-56.1	<1 m	
BHG [from (Kucharík *et al.*, 2014)]	[-59.2]	[∼20 h]	
pourRNA			
(maxE=-56, deltaE=5)	-62.3	6.3 h	143,032
(maxE=-56, deltaE=6)	-62.3	7.8 d	714,359
(maxE=-56, deltaE=9)	-62.3	10.6 d	771,300
(maxE=-56, deltaE=5, kBest=8)	-62.3	4.9 h	111,777
(maxE=-56, deltaE=6, kBest=5)	-62.3	2.2 h	40,880
pourRNA - macroscopic direct paths			
(maxE=-56, deltaE=6, Δ_{bp}=6)	-59.4	1.2 h	20,607
(maxE=-56, deltaE=6, Δ_{bp}=8)	-62.3	2 h	44,130
(maxE=-56, deltaE=6, Δ_{bp}=10)	-62.3	4.3 h	89,598
(maxE=-56, deltaE=6, Δ_{bp}=10, kBest=6)	**-62.3**	**2.8 m**	2,361

Table 1. Barrier estimations (energy of lowest transition states $\hat{P}$ in $\frac{kcal}{mol}$ using `turner-99` energy parameters from the Vienna RNA package 2.4.10) and runtime for `sv11` refolding using different methods: microscopic direct path search (findpath), Basin-Hopping-Graph (BHG) and pourRNA. For pourRNA also the number of processed macro-states $b \in \mathcal{B}_p$ is reported.

To test whether the identified barrier can be improved, we run higher local flooding thresholds of up to 9 $\frac{kcal}{mol}$. While we are exploring a much larger part of the energy landscape no lower energy barrier could be found. We thus conclude that -62.3 $\frac{kcal}{mol}$ is the true energy barrier of the refolding.

While superior to both *findpath* as well as *BHG*, *pourRNA* still requires much more time than *findpath*. To further speed up the computation, we next tested the impact of the kBest heuristics, i.e. we restrict exploration to neighbored basis that are reached via the highest rates (and thus lowest local energy barriers). Setting kBest to 5, reduces the runtime to 2.2 hours while we find the same energy barrier. This results from the (expected) strong reduction of explored basins (see Tab. 1).

Refolding along macroscopic direct paths

As shown by Kucharík *et al.* (2014), the energy barrier identified by *BHG* can be found by nearly direct microscopic paths. Thus, we expect to find our identified energy barrier via macroscopic direct path exploration.

To this end, we first investigate the impact on runtime and barrier estimation when comparing macroscopic direct paths with the exhaustive exploration reported in the last section for deltaE=6 $\frac{kcal}{mol}$. When not allowing for some structural flexibility ($\Delta_{bp} = 0$), the meta-stable state can not be connected to the global optimum. When local minima of basins are allowed to differ in up to 6 base pairs from both start and target minimum ($\Delta_{bp} = 6$), we find a first energy barrier estimate of -59.4 $\frac{kcal}{mol}$ within 1.2 hours. Further relaxations to $\Delta_{bp} = 8$ reproduces in 2 hours the already known barrier of -62.3 $\frac{kcal}{mol}$. The runtime differences correlate well with the number of processed macro-states (compare Tab. 1).

Since the overall runtime is still high, we restrict the exploration to fast transitions along macroscopic direct paths by setting kBest. As before, focussing on fast transitions vastly reduces the explored sublandscape. This enables a much faster identification of the optimal energy barrier within less than three minutes, which is close to the *findpath* heuristic.

Note, beside being orders of magnitude faster compared to *BHG*, *pourRNA* provides deterministic results while *BHG* is using randomized sampling strategies to cover the energy landscape.

5 Conclusion

The computation of gradient basin macro-state transition rate models to study the folding kinetics of RNA molecules is a hard computational problem. So far, exact computations were mainly limited by the extensive memory consumption resulting from the exponential growths of the RNA structure space with sequence length.

Here, we have introduced *pourRNA* that implements an explorative local flooding strategy rather than applying a global flooding scheme as done by state-of-the-art approaches. Local flooding enables faster kinetics model computation while its memory footprint is orders of magnitude smaller compared to the *barriers* pipeline. Furthermore, it enables the application of restricted local flooding schemes. *pourRNA*'s deltaE heuristic limits flooding to a given energy range above the basin's local minimum. That way, the resulting kinetics will reflect the fast refolding transition. Due to this limitation, transitions exceeding the deltaE limit will be missed. However, since the transition state ensembles are symmetric, they can be identified also in the reverse direction and are thus often still available for kinetics computation.

Further filters that restrict the exploration of neighbored basins e.g. to more stable structures (maxNeighE) or just to the most likely transitions (kBest) enable a further reduction of the computational cost but restrict the study e.g. to fast refolding pathways and kinetics. In its extreme, these filters can be used to identify gradient walks on a macroscopic level, needed e.g., to study folding funnels.

We emulated a refolding use case for the 115 nt long RNA `sv11`. This RNA is co-transcriptionally trapped in a meta-stable conformation. We have shown that *pourRNA* is able to identify better energy barrier estimates compared to results from literature. We could show that restricting the exploration to fast folding macroscopic direct paths provides a powerful filter to identify high accuracy energy barriers with low runtimes.

For such refolding experiments, it is often hard to choose a well working parameterization without additional knowledge. Thus, *pourRNA* enables dynamic adaptation e.g. of the kBest filter value if the start structure can not be connected either to the global minimum free energy structure or a provided target for the initial parameter setup.

Our ongoing work focuses on a dynamic deltaE heuristic that will ensure that the local partition function is well approximated using the ideas by Mann and Klemm (2011). Furthermore, we are investigating possibilities to parallelize local flooding to further speed up the investigation of very large basins. Finally, we will test other heuristics, e.g. following Bogomolov *et al.* (2010) or Huang and Voß (2014), to further improve the barrier estimates between macro-state clusters.

Acknowledgements

We thank S. Will for our fruitful discussions.

Funding

This work was supported by German Research Foundation [*BA2168/16-1*] and by the Austrian science fund FWF project SFP F43 Regulation of the RNA transcriptome.

References

Biebricher, C. and Luce, R. (1992). In vitro recombination and terminal elongation of RNA by Q beta replicase. *The EMBO Journal*, **11**(13), 5129–5135.

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.

Clote, P. (2005). An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. *Journal of Computational Biology*, **12**(1), 83–101.

Danilova, L. V., Pervouchine, D. D., Favorov, A. V., and Mironov, A. A. (2006). RNA-kinetics: a web server that models secondary structure kinetics of an elongating RNA. *J. Bioinform. Comput. Biol.*, **04**(02), 589–596.

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

Dotu, I., Van Hentenryck, P., Clote, P., and Lorenz, W. A. (2009). Computing folding pathways between RNA secondary structures. *Nucleic Acids Research*, **38**(5), 1711–1722.

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., Fontana, W., Hofacker, I. L., and Schuster, P. (2000). RNA folding at elementary step resolution. *RNA*, **6**(3), 325–38.

Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F., and Zehl, M. (2001). Design of multistable RNA molecules. *RNA*, **7**(2), 254–265.

Flamm, C., Hofacker, I. L., Stadler, P. F., and Wolfinger, M. T. (2002). Barrier trees of degenerate landscapes. *Z.Phys.Chem*, **216**, 155–173.

Frauenfelder, H. and Leeson, D. T. (1998). The energy landscape in non-biological and biological molecules. *Nature*, **5**, 757–759.

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. *Journal of Molecular Biology*, **379**(1), 160 – 173.

Hofacker, I. L., Schuster, P., and Stadler, P. F. (1998). Combinatorics of RNA secondary structures. *Discrete Applied Mathematics*, **88**(1), 207 – 237. Computational Molecular Biology DAM - CMB Series.

Hofacker, I. L., Flamm, C., Heine, C., Wolfinger, M. T., Scheuermann, G., and Stadler, P. F. (2010). BarMap: RNA folding on dynamic energy landscapes. *RNA*, **16**(7), 1308–1316.

Huang, J. and Voß, B. (2014). Analysing RNA-kinetics based on folding space abstraction. *BMC Bioinformatics*, **15**(1), 60.

Karplus, M. (2011). Behind the folding funnel diagram. *Nature Chemical Biology*, **7**, 401–404.

Kirkpatrick, B., Hajiaghayi, M., and Condon, A. (2013). A new model for approximating RNA folding trajectories and population kinetics. *Computational Science & Discovery*, **6**(1), 014003.

Klemm, K., Flamm, C., and Stadler, P. F. (2008). Funnels in energy landscapes. *The European Physical Journal B*, **63**(3), 387–391.

Kucharík, M., Hofacker, I. L., Stadler, P. F., and Qin, J. (2014). Basin Hopping Graph: a computational framework to characterize RNA folding landscapes. *Bioinformatics*, **30**(14), 2009–2017.

Kühnl, F., Stadler, P. F., and Will, S. (2017). Tractable RNA–ligand interaction kinetics. *BMC Bioinformatics*, **18**(12), 424.

Leopold, P. E., Montal, M., and Onuchic, J. N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, **89**(18), 8721–8725.

Lorenz, R., Flamm, C., and Hofacker, I. L. (2009). 2D projections of RNA folding landscapes. In *In: German Conference on Bioinformatics 2009*, volume 157 of *Lecture Notes in Informatics*, pages 11–20.

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**, 26.

Lorenz, W. A. and Clote, P. (2011). Computing the partition function for kinetically trapped RNA secondary structures. *PLOS ONE*, **6**(1), 1–17.

Lou, F. and Clote, P. (2010). Thermodynamics of RNA structures by Wang-Landau sampling. *Bioinformatics*, **26**(12), i278–i286.

Mann, M. and Klemm, K. (2011). Efficient exploration of discrete energy landscapes. *Phys. Rev. E*, **83**(1), online.

Mann, M., Kucharik, M., Flamm, C., and Wolfinger, M. T. (2014). Memory efficient RNA energy landscape exploration. *Bioinformatics*, **30**(18), 2584–2591.

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

Michalik, J., Touzet, H., and Ponty, Y. (2017). Efficient approximations of RNA kinetics landscape using non-redundant sampling. *Bioinformatics*, **33**(14), i283–i292.

Morgan, S. R. and Higgs, P. G. (1998). Barrier heights between ground states in a model of RNA secondary structure. *Journal of Physics A: Mathematical and General*, **31**(14), 3153.

Proctor, J. R. and Meyer, I. M. (2013). CoFold : an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Research*, **41**(9), e102.

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.

Stone, J. W., Bleckley, S., Lavelle, S., and Schroeder, S. J. (2015). A parallel implementation of the Wuchty algorithm with additional experimental filters to more thoroughly explore RNA conformational space. *PLOS ONE*, **10**(2), 1–20.

Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., and Gralla, J. (1973). Improved estimation of secondary structure in ribonucleic acids. *Nature New Biology*, **246**(150), 40–41.

Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L., and Stadler, P. F. (2004). Efficient computation of RNA folding dynamics. *Journal of Physics A: Mathematical and General*, **37**(17), 4731–4741.

Wuchty, S., Fontana, W., Hofacker, I. L., and Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**(2), 145–65.

Xayaphoummine, A., Bucher, T., and Isambert, H. (2005). Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, **33**(suppl_2), W605–W610.

Zhang, W. and Chen, S.-J. (2003). Analyzing the biopolymer folding rates and pathways using kinetic cluster method. *The Journal of Chemical Physics*, **119**(16), 8716–8729.