RNAxplorer: Harnessing the Power of Guiding Potentials for Sampling of RNA Landscapes

Gregor Entzian [1], Ivo Hofacker [1,2], Yann Ponty [3], Ronny Lorenz [1]* and Andrea Tanzer [1,4]*

[1]University of Vienna, Faculty of Chemistry, Department of Theoretical Chemistry, Währingerstraße 17, 1090 Vienna, Austria

[2]University of Vienna, Faculty of Computer Science, Bioinformatics and Computational Biology, Vienna, Austria

[3]LIX, CNRS UMR 7161, Ecole Polytechnique, Institut Polytechnique de Paris, France

[4]Medical University of Vienna, Center for Anatomy and Cell Biology, Division of Cell and Developmental Biology, Vienna, Austria

Abstract

Motivation: Predicting the folding dynamics of RNAs is a computationally difficult problem, first and foremost due to the combinatorial explosion of alternative structures in the structure space. Abstractions are therefore needed to simplify downstream analyses, and make them computationally tractable. This can be achieved by various structure sampling algorithms. However, current sampling methods are still time consuming and frequently fail to represent key elements of the folding space.

Method: We introduce `RNAxplorer`, a novel adaptive sampling method which uses dynamic programming to perform an efficient Boltzmann sampling in the presence of guiding potentials reflecting the similarity to already well-sampled structures. These potentials are accumulated into pseudo-energy terms that effectively steer sampling towards underrepresented or unexplored regions of the structure space.

Results: `RNAxplorer` allows us to efficiently explore RNA state space. It yields rare conformations that may be inaccessible to other sampling methods. We developed and applied different measures to benchmark our sampling methods against its competitors. Most of the measures show that `RNAxplorer` produces more diverse structure samples and is better at finding the most relevant kinetic traps in the landscape. Thus, it produces a more representative coarse graining of the landscape that is well suited to compute better approximations of RNA folding kinetics.

Availability: https://github.com/ViennaRNA/RNAxplorer/

Contact: ronny@tbi.univie.ac.at

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

1 Introduction

Over the past two decades, our understanding of the roles and functions of RNAs has fundamentally changed. With the advent of next-generation sequencing a plethora of non-coding RNAs were discovered, along with specific expression patterns that support a diversity of functions within cellular compartments and molecular mechanisms. Accordingly, genome-wide bioinformatics studies (Eddy, 1999; Saito *et al.*, 2009) have confirmed the dense population of the intergenic space with ncRNAs, and comparative genomics approaches have revealed evolutionary conservation of structured ncRNAs. Even protein coding mRNAs often rely on specific structural arrangements to control their own splicing, transcription, translation, or degradation, where structure elements often serve as recognition sites for binding partners such as proteins. Modeling the structure(s) of RNA is therefore an important step towards understanding its function.

*To whom correspondence should be addressed.

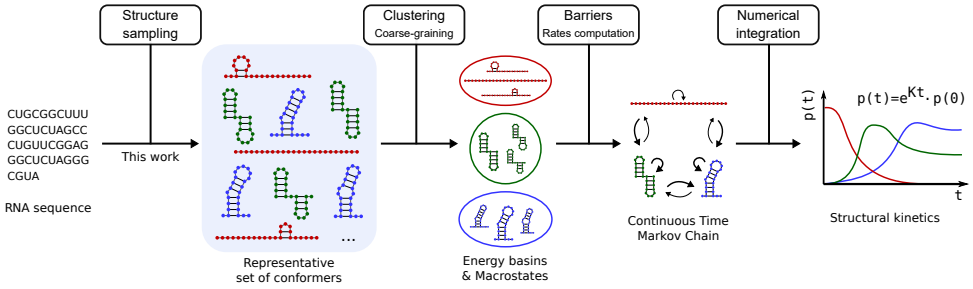

Figure 1: The RNA folding kinetics workflow starts with a sequence. For this sequence all secondary structures or a good approximation has to be sampled. This is followed by a coarse graining, rate computation and the final kinetics computation as a Markov Process.

At the secondary structure level, efficient dynamic programming (DP) algorithms enable the computation of various RNA structural properties at the thermodynamic equilibrium. Software suits such as `RNAstructure` (Reuter and Mathews, 2010), `UNAFold` (Markham and Zuker, 2008), or the `ViennaRNA` package (Lorenz *et al.*, 2011), enable the computation of minimum free energy (MFE), base pairing probabilities, consensus structures, RNA-RNA interactions and beyond using the Turner nearest neighbor model (Turner and Mathews, 2009). However, RNA folding is a dynamic process that already starts during transcription. While an RNA molecule tends to adopt a stable structural conformation, i.e. one that decreases its free energy, along the way it may be trapped in local minima. Depending on the height of (energy) barriers to escape such local minima, an RNA may only explore a negligible fraction of its conformation space, and never reach its ground state, the minimum free energy (MFE) structure, within its life time. Concrete instances of out-of-equilibrium kinetics notoriously include RNAs whose function is mediated by co-transcriptional folding and, in most cases, conformational switching (riboswitches).

A general framework for studying kinetics relies on an abstraction of the folding process as a Continuous-Time Markov Chain (CTMC) over a discretized conformational space. Properties of the CTMC can be derived from stochastic simulations of single trajectories within the folding landscape (Flamm *et al.*, 2000). However, many trajectories are then needed to estimate population densities, i.e. the probabilities/concentrations associated with most relevant conformations, hindering the kinetics analysis for RNAs beyond modest lengths. For these reasons, recent popular methods rely on a coarse-graining of the folding landscape, in which a subset of representative conformations is first identified, followed by the numerical resolution of the differential equation describing the time-resolved evolution of the population densities. Figure 1 illustrates the general principle of such a prediction workflow. The choice of a suitable coarse-graining is then critical to allow the omission of large parts of the conformational space, while maintaining key states in the coarse-grained RNA landscape, leading to an accurate approximation of RNA kinetics. Available approaches for coarse-graining include flooding strategies (Wolfinger *et al.*, 2004; Entzian and Raden, 2019), whose enumerative nature makes them unsuitable for RNAs beyond 100 nt. For longer RNAs, methods combining sampling with a reconstruction of the CTMC, such as the Basin Hopping Graph (Kucharík *et al.*, 2014), currently represent the only realistic option.

To identify important (meta)stable secondary structures within folding landscapes, the dominant approach usually resorts to structure sampling followed by a clustering step, as introduced by Ding *et al.* (2005). However, classified DP approaches have been proven useful to yield structure representatives from partitions of the ensemble that share a common feature, for instance their abstract shape (Giegerich *et al.*, 2004) or their base-pair distance to one or two reference structures (Freyhult *et al.*, 2007; Lorenz *et al.*, 2009). Other DP algorithms reduce the state space *ab initio* to draw (random) samples that constitute locally optimal structures, i.e. where no structural neighbor has lower free energy (Lorenz and Clote, 2011; Li and Zhang, 2011; Kucharík *et al.*, 2014; Michálik *et al.*, 2017).

However, the accuracy of virtually all the aforementioned methods is hindered by a strong bias towards low-free energy structures. This situation leads such methods to overlook important regions of the folding landscapes, or induces unreasonable computational costs due to precomputations (Michálik *et al.*, 2017), lack of diversity, forcing

further rounds of sampling (Kucharík *et al.*, 2014), or the downstream reconstruction of the coarse-grained CTMC model. Indeed, the clustering of structures, and computation of (pairwise) transition rates between the structures are the computationally most demanding steps. Computing such pairwise transition rate requires approximating the energy barrier between two secondary structures, a NP-hard problem even under simplistic assumptions (Maňuch *et al.*, 2009). Consequently, the structure sampling step is the most crucial, as a good balance between the size of the sample set and the coverage of important parts of the energy landscape is required.

In this work, we present a novel method to construct accurate approximations of kinetics landscapes. To this end, we iteratively utilize an efficient DP algorithm to compute the partition function (McCaskill, 1990) including pseudo-energies (Lorenz *et al.*, 2016), subsequently draw random samples using stochastic backtrack (Ding and Lawrence, 2003) and, iteratively, refine guiding potentials to (dis-)favor particular substructures, similar to the local elevation ideas introduced in *metadynamics* (Huber *et al.*, 1994). Our strategy provides a fast and effective means to discover local minima that may be far away from the ground state in terms of free energy but represent important landmarks of the energy landscape due to their impact on folding dynamics.

2 Methods

Formally, given an RNA sequence σ of length n, a secondary structure $s(\sigma) = \{(i,j) \mid (\sigma[i], \sigma[j]) \in BP)\}$ is sets of base pairs (i,j) compatible with σ. Here, one often restricts interacting nucleotides to the canonical Watson-Crick pairs (A, U) and (G, C), as well as the Wobble pair (G, U), i.e. $BP = \{(\mathsf{A},\mathsf{U}), (\mathsf{G},\mathsf{C}), (\mathsf{G},\mathsf{U})\}$. The generally accepted definition of secondary structures also excludes pseudo-knots and assumes a minimal backbone length between any two pairing bases due to sterical reasons. A detailed definition is given in Supp. Sec. 1.1.

The ensemble of all secondary structures compatible with an RNA σ defines its conformation space $\Omega(\sigma) = \{s(\sigma)\}$. Note, that in the following, we always assume a fixed sequence σ and will therefore omit its use for the sake of convenience. In conjunction with (i) a move set $\mathcal{M}$ that specifies elementary transitions to transform one structure s_i into one of its neighbors s_j, and (ii) the energy function $E : s \to \mathcal{R}$ that assigns each structure $s \in \Omega$ a real numbered value, one obtains the notion of the energy landscape $\mathcal{L} = \{\Omega, \mathcal{M}, E\}$. Over the past decades, different move sets $\mathcal{M}$ have been used (Flamm *et al.*, 2000; Xayaphoummine *et al.*, 2003), mostly to restrict the size of their induced neighborhood. The most commonly utilized move set is the difference of exactly one base pair between neighbored structures.

Local minima are defined via steepest descent trajectories $\gamma^{\infty}(s)$ of subsequent single base pair moves. These trajectories are called gradient walks and always end in a local minimum. Structures for which a gradient walk ends in the same local minimum belong to the same gradient basin of attraction $\mathcal{B}(s)$. Performing gradient walks for all structures results in a unique partitioning of the state space. This is often used as a most natural coarse graining in RNA folding kinetics simulations (Wolfinger *et al.*, 2004). Definitions of gradient walks can slightly differ in resolving ambiguity. We refer to the definition used by Entzian and Raden (2019).

Moreover, gradient basins and the minimal saddle points connecting them can be used to conveniently visualize and compare high-dimensional energy landscapes in the form of barrier trees or disconnectivity graphs (DG) (Becker and Karplus, 1997; Flamm *et al.*, 2002). However, computing the barrier tree for a particular RNA sequence typically relies on exhaustive enumeration of Ω which becomes impractical for sequence lengths of about $100\ nt$ or longer as Ω grows exponentially with sequence length n (Hofacker *et al.*, 1998).

Equilibrium Ensemble Properties Most RNA secondary structure prediction methods borrow a key concept of statistical mechanics, namely that structures s in thermal equilibrium are Boltzmann distributed, hence $p(s) \propto \exp(-\beta E(s))$ with $\beta := 1/kT$ for k the Boltzmann constant and T the temperature. For a particular RNA sequence this immediately suggests an obvious structure representative: the one with minimal free energy (MFE), i.e. $s_{\text{MFE}} = \arg\min_{s\in\Omega} E(s)$ since it has the highest probability among all other structures of the conformation space. Efficient DP algorithms exist that compute s_{MFE} in $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory for sequences of length

n (Zuker and Stiegler, 1981; Lorenz *et al.*, 2011). A small change in this DP concept leads to an efficient method to compute the partition function $Z = \sum_{s \in \Omega} \exp(-\beta E(s))$, with the same asymptotic complexities (McCaskill, 1990). Using Z many thermodynamic equilibrium properties can be derived, e.g. probabilities

$$p(s) = \frac{e^{-\beta E(s)}}{Z} \tag{1}$$

for any structure s or $p_{ij} = \sum_{s|(i,j)\in s} p(s)$ for base pairs (i, j). The DP algorithm to compute Z can also be adapted to perform a Boltzmann sampling, i.e. to draw structures s randomly from the ensemble according to their probability $p(s)$. This can be regarded a (random) backtracing in the DP matrices with worst case time complexity of $\mathcal{O}(n \log(n))$ per sample (Ding and Lawrence, 2003; Ponty, 2008).

2.1 The `RNAxplorer` Method

The `RNAxplorer` method approximates RNA energy landscapes using an iterative scheme which samples random structures using guiding potentials. Guiding potentials are updated after each iteration to avoid a concentration of samples within low free-energy basins, thus ensuring maximal coverage of the landscape. In order to steer sampling away from a given structure s° that has already been sampled repeatedly, we consider two types of guiding potentials:

$$E_c^{bp}(s^\circ, s) := \alpha \cdot \frac{|s \cap s^\circ|}{|s^\circ|} \tag{2}$$

$$E_c^{d}(s^\circ, s) := \alpha \cdot \frac{d_{\max}(s^\circ) - d_{\mathrm{BP}}(s, s^\circ)}{d_{\max}(s^\circ)} \tag{3}$$

where $d_{\max}(s^\circ) = \max_{s\in\Omega} d_{\mathrm{BP}}(s, s^\circ)$ and $d_{\mathrm{BP}}(s, s^\circ)$ represents the classic base pair distance (i.e. the number of base pairs that are present in one structure, but not in the other). The weight factor α is typically close to the thermal energy ($\alpha \approx kT$), effectively yielding moderate pseudo-energy penalties. The E_c^{bp} potential penalizes a structure s based on the number base pairs shared with s°, while E_c^d penalizes a structure depending on its base pair distance to s°. Note that by changing the sign of α, these potentials can also be used as attractive potentials guiding the sampling towards a region of interest.

For each structure guiding potentials are initially set to 0 and incrementally updated and accumulated within pseudo-energy terms E_Ψ over the course of sampling. At each round i, a multiset $\mathcal{S}_i \subseteq \Omega$ of structures is sampled from a (distorted) Boltzmann distribution. Through gradient descent, each structure $s \in \mathcal{S}_i$ is mapped to its local minimum $\gamma^\infty(s)$, used as a representative for its energy basin. The resulting set of local minima is then analyzed to identify the most over-represented basin, denoted as

$$s^\circ = \arg\max_{\hat{s}\in\Omega} |\mathcal{B}_{\mathcal{S}_i}(\hat{s})|, \quad \text{with} \quad \mathcal{B}_{\mathcal{S}}(\hat{s}) := \{s \in \mathcal{S} \mid \gamma^\infty(s) = \hat{s}\}.$$

In other words, s° is the local minima that attracts the most samples in $\mathcal{S}_i$. The method then updates the pseudo-energy term E_Ψ for iteration $i + 1$, based on the structural features of s°, through:

$$E_{\Psi_{i+1}}(s) := E_{\Psi_i}(s) + E_c(s^\circ, s). \tag{4}$$

The method stops after a user-defined maximum number $i_{\max}$ of iterations.

Additionally, we use a strategy that after every iteration determines whether E_Ψ needs an update. This avoids unnecessary recomputation of the rather costly partition function (see Sec. 2.2). The general idea is that the depth of a sampling is sufficient if collisions pervasively occur, i.e. most structures are observed multiple times (Sahoo and Albrecht, 2012). Similar to Kucharík *et al.* (2014), we compare the number of local minima observed only once against those observed multiple times. Given a saturation threshold μ, the algorithm only updates E_Ψ if

$$\frac{|\{\hat{s} \in \Omega : |\mathcal{B}_{\mathcal{S}_{\le i}}(\hat{s})| = 1\}|}{|\{\hat{s} \in \Omega : |\mathcal{B}_{\mathcal{S}_{\le i}}(\hat{s})| > 1\}|} \le \mu. \tag{5}$$

where $\mathcal{S}_{\le i} = \cup_{j=1}^{i} \mathcal{S}_j$ is the superset of structures from all iterations.

2.2 Sampling with Guiding Potentials

To mitigate oversampling, we introduce a focused approach based on (directed) guiding potentials, i.e. pseudo-energy terms that supplement the free-energy, and steer the sampling towards or away from a (set of) structure(s). This allows a finer level of control over the redistribution of the emission probabilities than previous alternatives, such as the temperature elevation method of Kucharík *et al.* (2014) (see Supp. Mat. 1.2).

Namely, given a pseudo energy $E_\Psi(s)$, our sampling procedure considers a pseudo-energy function $\hat{E}(s) = E(s) + E_\Psi(s)$ that includes the guiding potential, where $E(s)$ is the classic Turner free-energy. Our goal is then to sample from the distribution

$$\hat{p}(s) = \frac{e^{-\beta \hat{E}(s)}}{\hat{Z}} \quad \text{with} \quad \hat{Z} = \sum_{s \in \Omega} e^{-\beta \hat{E}(s)}. \tag{6}$$

However, Boltzmann sampling requires the precomputation of $\hat{Z}$, not through a exhaustive summation due to the combinatorial explosion of Ω, but rather using a recursive DP scheme. Thus in order to benefit from efficient algorithms, we restrict our attention to guiding potentials E_Ψ such that, for any structure s, $E_\Psi(s)$ can be written as a sum of contributions associated to derivations of the underlying folding grammar. Sampling under such guiding potentials is generically supported by the soft constraints framework of Lorenz *et al.* (2016)

In a minimalist setting, let us consider the case where energy terms $E_{i,j}$ are associated to base pairs such that, for any structure s, one has

$$E_\Psi(s) = \sum_{(i,j) \in [1,n]^2} \theta_{(i,j)\in s} \cdot E_{i,j} \tag{7}$$

where $\theta_{(i,j)\in s}$ is the indicator function, taking value 1 if $(i,j) \in s$ and 0 otherwise. Despite the simplicity of this terms, E_Ψ can be used to steer the sampling towards/away from one or several reference structure(s). For instance, the individual base pairs of a reference structure s° can be penalized/promoted by setting $E_{i,j} = \theta_{(i,j)\in s^\circ} \cdot \alpha$, leading to $E_\Psi(s) = |s \cap s^\circ| \cdot \alpha$. Setting $\alpha < 0$ will decrease the expected distance between sampled structures to $\hat{s}$, while $\alpha > 0$ will increase it.

This idea can be further generalized to capture the distance to multiple structures simultaneously, to express the guiding potentials E_c^{bp} of Equation (2). After k updates, associated to most represented structures $s_1^\circ, \ldots, s_k^\circ$, it suffices to set

$$E_{i,j} = \alpha \sum_{\ell=1}^{k} \frac{\theta_{(i,j)\in s_\ell^\circ}}{|s_\ell^\circ|}$$

which gives

$$\begin{aligned} E_{\Psi_{k+1}}(s) =& \sum_{(i,j)} \theta_{(i,j)\in s} \alpha \sum_{\ell=1}^{k} \frac{\theta_{(i,j)\in s_\ell^\circ}}{|s_\ell^\circ|} \\ =& \alpha \sum_{\ell=1}^{k} \frac{|s \cap s_\ell^\circ|}{|s_\ell^\circ|} = \sum_{\ell=1}^{k} E_c^{bp}(s, s_\ell^\circ) \end{aligned}$$

in which one recognizes the intended guiding potential after k updates.

Distance-based guiding potentials, such as $E_c^d(s^\circ, s)$ described in Equation (3), can also be expressed by assigning energy terms to derivations. Our approach in this case generalizes a strategy introduced by Freyhult *et al.* (2007) and Lorenz *et al.* (2009) respectively for a single and two reference structures. Its overarching principle is to keep track of the base pairs, in the reference structures, that are conclusively ruled out by the choice of a DP derivation. However, due to the technicality of its implementation, and its tight embedding within the DP scheme, we reserve a full exposition of its principle to Supp. Sec. 2.1.

Finally, complex pseudo-energy guiding potentials can be constructed, e.g. through variations of the energy values and/or a combination of using individual base pairs and structures as targets (see Supp. Sec. 2.2 and 2.3).

2.3 Quality assessment

To judge whether or not a landscape $\mathcal{L}$ is adequately approximated by a set of structures is a difficult task and dependens on particular applications. Here, we use (thermodynamic) ensemble measures to analyze a set with respect to its inherent diversity. The approximated shape of $\mathcal{L}$ is important for the overall dynamical behavior of subsequent folding simulations. We therefore analyze sample sets for the presence of certain key structures.

Mean pairwise distance. One of the most simple measures of structure diversity within any set of structures Ω is the (unweighted) *mean pairwise distance* $\langle d\_u \rangle = \frac{1}{|\Omega|} \sum_{s,t \in \Omega} d_{\text{BP}}(s,t)$. The smaller this value, the more similar the structures in Ω are. In equilibrium, one further weights the structures according to their probability in Ω. The weighted *mean pairwise distance* $\langle d \rangle = \sum_{s,t \in \Omega} p(s)p(t)d_{\text{BP}}(s,t) = \sum_{ij} p_{ij}(1 - p_{ij})$ can then be conveniently computed from the pairing probabilities p_{ij}.

Distance Classes. Single numbered average diversity measures, such as $\langle d \rangle$, usually conceal whether or not Ω is dominated by only a small number of representatives. A more detailed analysis can be done by partitioning Ω into distance classes with respect to one or many reference structures. Following the lines of Lorenz *et al.* (2009), with two fixed references $\hat{s}_1, \hat{s}_2$ each structure $s \in \Omega$ is assigned to its corresponding class $\mathcal{C}^{d_1,d_2}$ where $d_1 = d_{\text{BP}}(s, \hat{s}_1)$ and $d_2 = d_{\text{BP}}(s, \hat{s}_2)$. Each class can then be represented by

$$\text{MFE}^{d_1,d_2} = \min_{s \in \mathcal{C}^{d_1,d_2}} E(s), \tag{8}$$

or the corresponding ensemble free energy

$$G^{d_1,d_2} = -\beta \ln Z^{d_1,d_2}, \quad \text{with} \quad Z^{d_1,d_2} = \sum_{s \in \mathcal{C}^{d_1,d_2}} e^{-\beta E(s)}. \tag{9}$$

Finally, the resulting projections can be conveniently visualized in Cartesian coordinates and dimensions d_1 and d_2, for instance in the form of a heat map (see Fig. 2 or Fig. 5).

As a proxy of diversity, we count the number of distance classes $\mathcal{C}^{d_1,d_2}$ that are adequately covered by a sample set $\mathcal{S}$. We assume coverage to be sufficient, if any sampled structure s mapped to $\mathcal{C}^{d_1,d_2}$ is within an energy margin ϑ around MFE^{d_1,d_2} of the full ensemble, i.e. $\min_{s \in \mathcal{S} \cap \mathcal{C}^{d_1,d_2}} E(s) - MFE^{d_1,d_2} \leq \vartheta$, cf. Supp. Sec. 5.5.

Density of States. A useful measure to assess the free energy distribution within a set of structures is the *density of states* (DOS) (Cupal *et al.*, 1997). Here, structures $s \in \Omega$ are classified according to their free energy $E(s)$ to obtain densities, i.e. the number of structures, at particular energy levels. Most sampling methods are prone to over-sample the low free energy regime, but structures at higher energy levels might be important for modeling the dynamic behavior of RNA folding. For our benchmark, we therefore compute the DOS for each sample set obtained by the different methods. A visual comparison to the ground truth, i.e. the DOS of the full ensemble, can then be used as a quality measure.

Energy barriers. Comparing folding simulations between different tools is not trivial. This can be attributed to inherently different coarse graining of the methods, fast fluctuations, and slow folding components that only change close to thermodynamic equilibrium. To circumvent this problem, we first consider both, the computation of transition rates and the actual simulation, separate tasks, that are independent from the step that generates the sample sets. Assuming that we can solve these tasks, we can investigate whether the sampled structures correspond to those that determine the overall folding kinetics behavior. Thus, samples must not only cover the lowest energy states of $\mathcal{L}$, but also refolding events with large energy barriers, i.e. those associated with slow rates that effectively determine the long time behavior of the folding dynamics (indicated by Becker and Karplus (1997); Flamm *et al.* (2002)). For that purpose, structures of a sample set can be mapped into a barrier tree representation of the full

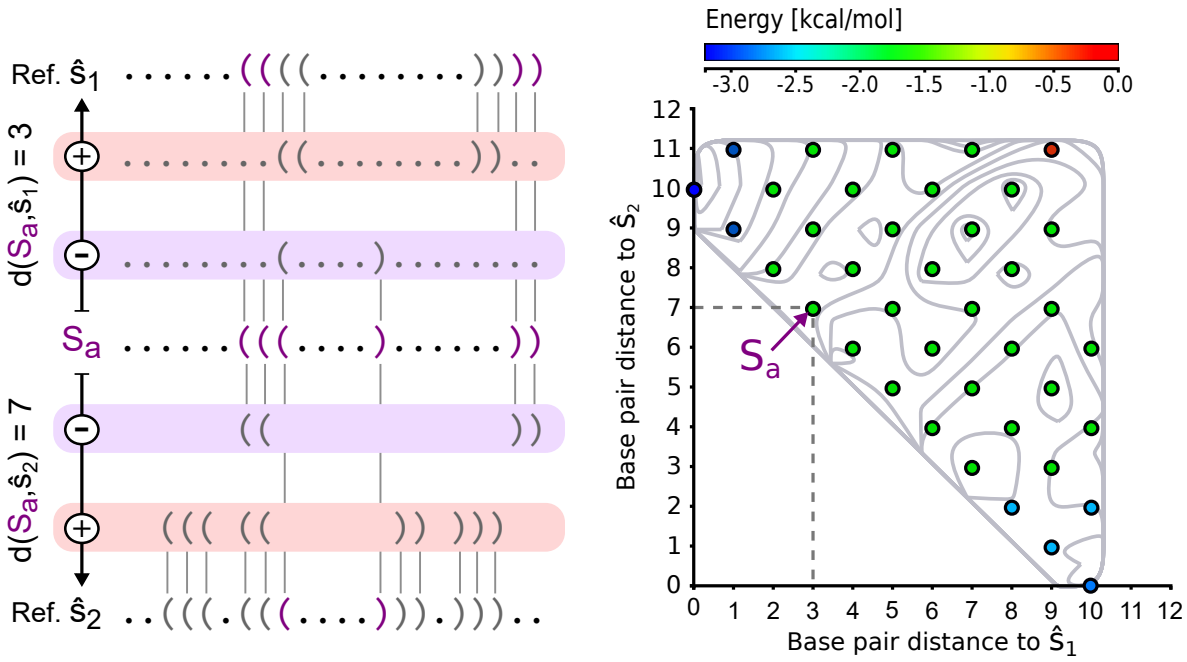

Figure 2: Toy example of a 2D projection. Each colored circle corresponds to a valid base pair distance to both reference structures. The color corresponds to the energy of the MFE structure representative of this spot. The background consists of isolines which are created from an interpolation of the free energy values that correspond to the MFE structures of each circle. Structure S_a has base pair distance 3 to the reference $\hat{s}_1$, because one base pair have to be removed, and two added, in order to change $\hat{s}_1$ into S_a. The distance to $\hat{s}_2$ is 7 because two base pairs must be removed, and 5 inserted, to change S_a into $\hat{s}_2$.

ensemble. One can then compute the fraction of leaves covered by, and the highest energy barriers associated with the structures within each sample set. For details, we refer to Supp. Sec. 5.3.

2.4 Implementation

From the implementation perspective, we use the constraint framework of the `ViennaRNA Package` that allows us to specify guiding potentials E_Ψ as separate functions that are then integrated into the prediction algorithms in a plugin-like manner (Lorenz *et al.*, 2016). This allows us to dynamically adapt E_Ψ without the need to re-implement the computation of $\hat{Z}$ and the subsequent Boltzmann sampling. We implemented the novel guiding potential based sampling approach as described in sections 2.2 and 2.1 in 'C' as part of the executable program `RNAxplorer`. For the iterative sampling method, the user can choose between two penalties. Either it is proportional to the number base pairs a structure shares with an overrepresented local minimum (Eqn. (2)) or it is proportional to the base pair distance (Eqn. (3)). The total number of iterations $i_{\max} = \frac{N}{g}$ is automatically computed from the requested sample size N and a user-defined granularity g. Additionally, `RNAxplorer` offers access to produce individual sets of structure samples, implements different heuristics to compute (optimal) transition paths to eventually determine saddle points required to assess transition rates, and finally, provides gradient walk methods for coarse grain the sampled state space. The program also comes with a `python` script that enables hierarchical clustering of secondary structures, see Supp. Sec. 2.3.

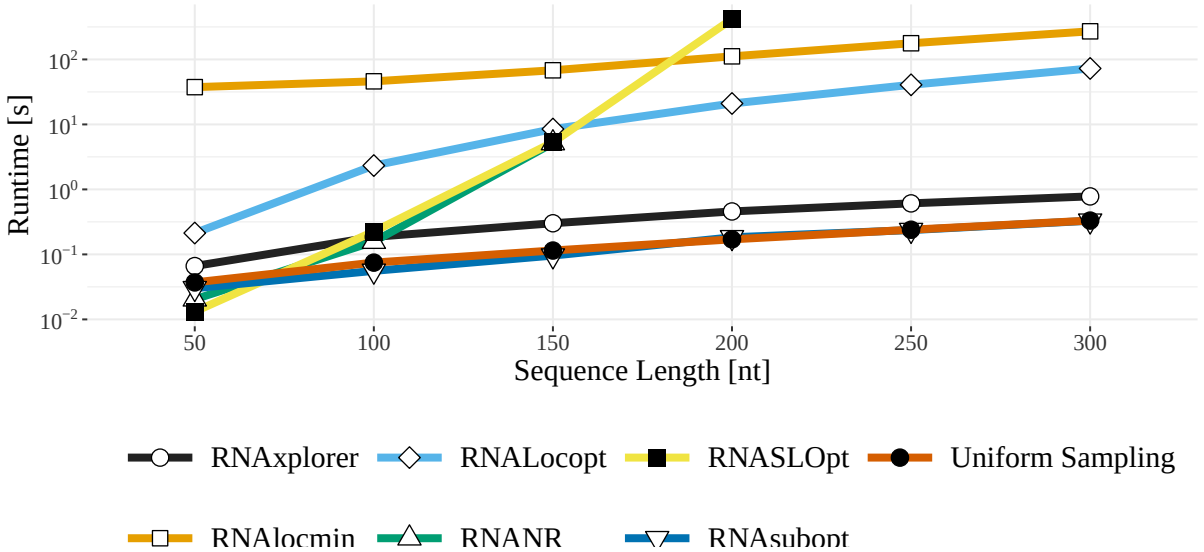

Figure 3: Runtime comparison. Runtimes observed for a sample of 1,000 structures for RNA lengths from 50 to 300 nucleotides, averaged over 10 randomly generated sequences. For `RNASLOpt`, we precomputed a δ value to obtain at least 1,000 structures. For `RNAxplorer` and `RNAlocmin` the number of iterations was set 100.

3 Results

In the following we assess the quality and applicability of our novel sampling method by comparing its results against other, widely-used, RNA secondary structure sampling methods. For that purpose, we first collected a set of 9 benchmark sequences for which landscape approximations will be made. The exact sequences can be found in Table S1 of the supplementary material. The methods and tools we compare our approach against are (i) *uniform sampling* with `RNAsubopt` (Lorenz *et al.*, 2011) realized through Boltzmann sampling at extremely high temperatures, (ii) regular Boltzmann sampling with `RNAsubopt` (*-p* command line option), (iii) Boltzmann sampling of locally optimal structures with `RNAlocopt` (Lorenz and Clote, 2011), (iv) non-redundant sampling of saturated structures with `RNANR` (Michálik *et al.*, 2017), (v) the temperature elevation scheme Boltzmann sampling of `RNAlocmin` (Kucharík *et al.*, 2014), and (vi) a set of stable local optimal structures generated by `RNASLOpt` (Li and Zhang, 2011). Note, that `RNASLOpt` differs from all the others in that it is deterministic and always exhaustively enumerates locally optimal structures (LOpts) in a pre-defined energy band above the MFE. The width δ of this band can be specified in discrete steps of kcal/mol or percentages. This, unfortunately, prohibits one to explicitly set the number of output structures in advance. Therefore, in some of the analysis below, we either determined the minimal width δ that results in at least the number of required samples in a pre-processing step, or we simply omit its use altogether. All programs were used with default parameters unless stated otherwise.

3.1 Time and Memory Consumption

First, we prepared a set of artificially generated random sequences with equal probabilities for each of the four RNA nucleotides to assess the runtime and memory requirements for all programs in our comparison. To that end, we generated 10 sequences with lengths of 50–300 nt in steps of 50 nt. For each of the resulting 60 sequences the 6 different tools were instructed to (randomly) draw $1,000$ structures from the respective ensembles. For the iterative methods implemented in `RNAlocmin` and `RNAxplorer`, the number of iterations was set to 100. All computations were performed on a workstation with Intel® Core™ i7-7700K CPU running at 4.20 GHz and 32 GB of RAM.

As expected, the standard Boltzmann sampling strategies of `RNAsubopt` with default parameters as well as

uniform sampling were the fastest methods tested (Fig. 3) and required the least amount of memory. The next best tool in terms of both, runtime and memory requirements, is our new heuristic `RNAxplorer`, followed by `RNAlocopt` and `RNAlocmin`. While runtimes of `RNAxplorer` and `RNAsubopt` are within the same order of magnitude, `RNAlocopt` and `RNAlocmin` are by two orders of magnitude slower. The exponential runtime asymptotics of `RNANR` and `RNASLOpt` render them the slowest for longer sequences. Note, that we were not able to produce results (within reasonable time) for sequence longer than 150 nt (`RNANR`) and 200 nt (`RNASLOpt`) due to the restrictions imposed by our testing machine (limited memory). However, for shorter RNA sequences up to about 150 nt, these two programs are still faster than `RNAlocmin` (Fig. 3). Further runtime and memory benchmarking results are available in Supp. Sec. 4.

3.2 Structure Sample Diversity

For each method we assessed sampling quality in terms of diversity of structures obtained. However, since the concepts of sampling vary greatly between methods it seemed difficult to impossible to express structural diversity of the samples drawn as a single common value. We therefore calculated and compared (i) structural redundancy with and without thermodynamic equilibrium assumptions in both spatial (base pair distance) and energy dimensions. Furthermore, we compared both (ii) coverage of distance classes and (iii) density of states. In this analysis, we explicitly omitted `RNASLOpt` not only because it always produces unique structures in a deterministic manner, but also due to the excessive amount of time required to precompute the correct δ for each sample size.

Structural Redundancy. We express the redundancy of the generated sample sets by two different measures, (i) the fraction of unique local minima reachable from the structures within each set and (ii) the mean pairwise base pair distance within the unique part of these local minima (for definitions see Supp. Sec. 5.1). The redundancy measures were then averaged over 10 independent rounds of sampling with 10^6 structures for each benchmark sequence (for details see Supp. Sec. 5.1 and Fig. S5).

The largest fraction of unique local minima was obtained from `RNANR` with an average of 99%, followed by *uniform sampling* with 84%. Among the remaining tools, `RNAxplorer` achieved the highest fraction with an average of 14%, followed by `RNAlocmin` (9%), `RNAlocopt` (3%), and `RNAsubopt -p` (2%). In terms of normalized mean base pair distance within the unique fraction of local minima, `RNANR` achieves the highest diversity with an average of 0.45 bp/nt, closely followed by `RNAxplorer` with 0.44 bp/nt. All other tools yield much lower diversity with just 0.31 (`RNAlocmin`), 0.26 (`RNAlocopt`), 0.24 (Regular Boltzmann sampling), and 0.23 (*uniform sampling*) base pairs per nucleotide. However, when weighting the individual structures by their equilibrium probabilities, the individual differences vanish for the majority of tools and benchmark sequences. In this analysis, only *uniform sampling* produces, on average, a rather low normalized weighted mean base pair distance of 0.07 bp/nt while the other tools yield 0.11 bp/nt.

Coverage of Distance Classes. Next, we took a closer look at the spatial resolution of the sample sets. In particular, we were interested whether (i) the samples spread over a large number of representative structures with fundamentally different base pair patterns, or (ii) the samples mainly reflect representatives of structurally similar clusters. For that purpose, we use distance classes $\mathcal{C}^{d_1,d_2}$ (cf. Fig. 2), were we partitioned the sample sets according to their distance to (i) the MFE structure and (ii) the most stable structure that does not share any base pair with the MFE structure. Note, that the latter can be obtained from a constrained MFE prediction where all base pairs of the actual MFE structure are prohibited. For each class we computed the MFE and ensemble free energy to compare them against exact values as computed with `RNA2Dfold` (Lorenz *et al.*, 2009).

Such projections into lower dimensions provide easy to assess visual impressions of the sample diversity, as shown in Figure 5. However, here, we use them to count how many $\mathcal{C}^{d_1,d_2}$ were covered by the different sampling methods. To alleviate the impact of randomness during the sample generating process, we averaged the results for each experiment over 10 independent runs. Figure 4 summarizes the results over all benchmark sequences as a function of sample size and two thresholds $\vartheta_1 = 0$ kcal/mol and $\vartheta_2 = 5$ kcal/mol.

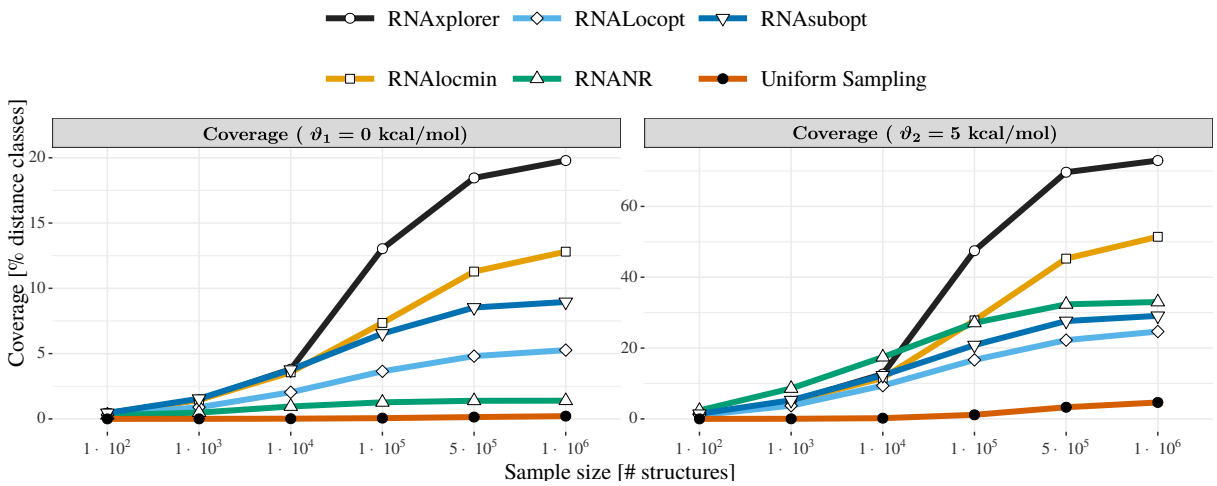

Figure 4: Distance class coverage as a function of sample size. Shown are the fractions of distance classes $\mathcal{C}^{d_1,d_2}$ covered by at least one local minimum derived from the sample set that is energetically close to the respective MFE^{d_1,d_2}. The data averages over all 9 benchmark sequences, 10 independent runs per tool and margins $\vartheta_1 = 0$ kcal/mol (left plot), and $\vartheta_2 = 5$ kcal/mol (right plot).

`RNAxplorer` clearly outperforms the other methods even for small sample sizes. With increasing sample size the coverage quickly rises and is always higher compared to the other methods. Only for `RNAlocmin` the coverage rises similarly fast with increasing sample size. The next best tools are `RNAsubopt` and `RNAlocopt` (ϑ_1) and `RNANR` (ϑ_2). As expected, *uniform sampling* covers just a tiny, almost constant fraction even for very large sample sizes of 10^6 structures. For `RNANR` the diversity is very sequence dependent which is depicted in Figure S3. Since `RNANR` could not be applied to 3 of the 9 benchmark sequences (SAM riboswitch of `metE`, lysine riboswitch of `lysC` and TPP riboswitch of `thiamine` gene) due to its demanding memory requirements, the average for this tool as shown in Figure 4 only consists of the remaining 6 sequences. Results for the individual benchmark sequences can be found in Figure S22. The analog measure based on partition functions is shown in Figure S24 for individual sequences and in Figure S25 as average over all sequences. For more details on the coverage measure see Supp. Sec. 5.5.

Density of States. Our last diversity measure provides a focus on the free energy distribution of structures within each sample set. We first computed the energy spectra as density of states (DOS) of the full structure ensemble for each of the benchmark sequences (Cupal *et al.*, 1997) utilizing the program `RNAdos` of the `ViennaRNA Package`. Then, for each sequence, we determined the corresponding energy spectrum for the subset of unique structures within a sample of 10^6 structures as obtained from each tool. The results are available for visual comparison in Figure S6 of the supplementary material.

`RNAxplorer`, `RNAlocmin`, `RNAsubopt` and `RNAlocopt` show their highest densities in the low free energy regime where they cover almost all available structures. At higher energies, coverage flattens out for `RNAxplorer` and `RNAlocmin` with an extent close to 0 kcal/mol and above. `RNAlocopt`, `RNAsubopt`, and `RNANR`, on the other hand, exhibit a steep decrease in coverage towards higher energies with prominent peak in most of the cases. However, the latter shows a much larger variance than the former two, thus also expanding far into the high energy regime. For `RNAlocopt` and `RNANR` the spectrum most often starts slightly above the MFE, highlighting the effect of their implicit coarse graining of the structure ensemble. Interestingly, `RNAxplorer` produces a bump in the higher energy spectrum for *sv11* which coincides with the energy of the corresponding metastable state.

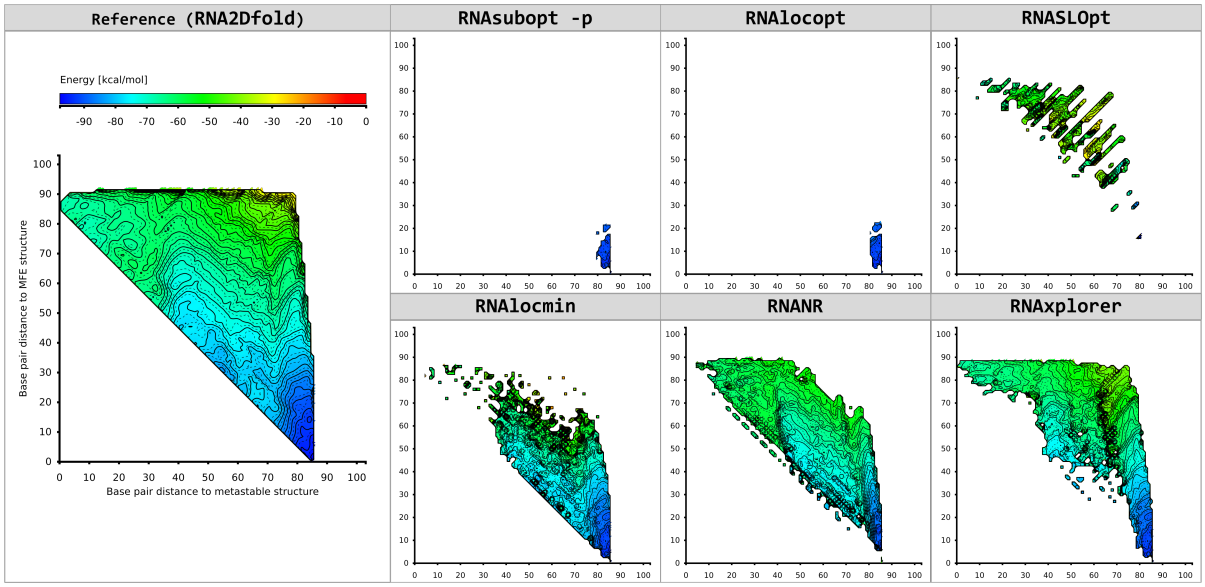

Figure 5: 2D projections of local minima as obtained from different methods for the SV-11 Q beta replicase template (Biebricher and Luce, 1992). Reference structures for the projection are the MFE and metastable structure. The left most column depicts the *ground truth* as computed by `RNA2Dfold`, chosen here as a reference for comparability reasons. The remaining panels show the results for Boltzmann sampling (`RNAsubopt -p`), local optima sampling (`RNAlocopt`), `RNASLOpt`, variable temperature sampling (`RNAlocmin`), non-redundant sampling (`RNANR`), and repellant sampling (`RNAxplorer`), which required 6.75s, 21.81s, 115.99s, 487.73s, 4285.81s, and 27.87s to produce the sample sets, respectively. The sample size before local minima determination for each tool was 10^6 (except for `RNASLOpt` that always yields less structures even with exhaustive enumeration, i.e. artificially high δ).

3.3 Suitability for RNA folding kinetics

Using the `barriers` program Flamm *et al.* (2002), we generated barrier trees for our benchmark set of random sequences using exhaustive structure enumeration up to 15 kcal/mol above the MFE with `RNAsubopt`. Coarse graining of the barrier tree was set to a minimal energy barrier of 3 kcal/mol between neighboring basins. We then mapped the local minima generated by each sampling method into the respective barrier trees to determine how many of the 100 largest energy barriers have been found. The results were further averaged over 10 rounds of sampling to alleviate the impact of randomness in the sample sets.

As shown in Table 1, for 100 nt long sequences all tools already find a large amount of the highest energy barriers even for small sample sizes such as 10^3. At the same time, the number of recovered basins is as low as $1-2\%$. `RNANR` in general recovers more basins than the other tools for sequence lengths of 70 nt or longer. For sample sizes of 10^5 structures, the tools `RNAxplorer`, `RNAlocmin`, `RNAlocopt`, and `RNAsubopt` perform equally good in finding the highest energy barriers. In contrast, both, `RNAxplorer` and `RNAlocmin` stand out in the number of recovered basins with 22.25% and 15.48%, respectively, compared to less than 10% achieved by the other methods. In terms of run time, `RNAxplorer` is much faster than `RNAlocmin` with an average of just 3.72 s compared to 73.61 s. Details and remaining results for this analysis are available in Supp. Sec. 5.3.

4 Conclusion and discussion

In this paper we have introduced RNAxplorer, a tool based on an RNA secondary structure sampling method with guiding potentials to approximate the underlying energy landscape. Its very small foot print in terms of memory and computation time requirements enables it to be applied to RNAs with sequence lengths beyond those that can be handled with other, comparable approaches. Our tool creates diverse structure samples with low as well as high free energy, that seem to nicely encompass those relevant for subsequent folding kinetics simulations. This has been shown in a benchmark analysis for biologically relevant and randomly generated RNAs using various quality measures. Thus, our novel sampling method may enable the investigation of the folding dynamics of longer RNAs than possible with state-of-the-art tools.

Efficient implementation, simple strategy and utilization of features of the ViennaRNA Package in general and soft constraints in particular make RNAxplorer one of the fastest structure sampling methods available. Memory consumption is minimal and mostly attributed to storing the list of structures obtained and the DP matrices of the partition function computations. As a consequence, unlike other tools in our benchmark, RNAxplorer yields representative samples within reasonable time frames even for RNAs with lengths of 300 nt or beyond.

The limiting factor on the runtime of our new approach is the number of times new guiding potentials are added, as they each require additional $\mathcal{O}(n^3)$ time to re-compute the partition function. Thus, we hand over control of setting the granularity g, i.e. the number of samples drawn at once, to the user. For sequences of length n and a total number of structures N to sample, the upper limit on the total asymptotic runtime then becomes $\mathcal{O}(\frac{N}{g}n^3 + N\,n^2)$. Depending on the shape of the underlying energy landscape the algorithm, however, usually performs much less partition function re-computations than suggested by this crude upper bound.

A close investigation of the sample quality also reveals favorable for RNAxplorer. In terms of structural redundancy, the samples obtained generally show a high degree of uniqueness. In contrast to RNANR and *uniform sampling*, the algorithm still spends some time to redundantly draw the same structures. But this redundancy is typically smaller than 50%, and therefore comparable to that observed with RNAlocmin. RNAsubopt and RNAlocopt, on the other hand, are pure Boltzmann sampling approaches without any provisions to prevent oversampling. Thus, they were expected to perform worse than RNAxplorer in this benchmark. While samples obtained from *uniform sampling* almost only consisted of unique structures, their rather good performance in terms of weighted mean base pair distance is quite misleading. The samples almost exclusively consist of high free energy structures, as shown in the corresponding DOS computations, thus missing potentially important local minima. Therefore, they have to be regarded as a bad approximation of the actual energy landscape.

Within the subset of unique structures RNAxplorer clearly outperforms RNAlocmin in terms of structural and stability diversity. In general, structures obtained with our new sampling approach show a very high degree of dissimilarity in their base pairing patterns, comparable to those obtained with RNANR or *uniform sampling* and, on

	10^3			10^5		
Tool — 100 nt	coverage [%]			coverage [%]		
max. basins: 63536	barriers	basins	$\bar{t}$ [s]	barriers	basins	$\bar{t}$ [s]
RNAxplorer (rs)	**79**	1.67	0.06	**94**	**22.25**	3.72
RNAxplorer (rsps)	74	1.48	0.38	**88**	9.85	10.80
RNAlocmin	75	1.46	41.19	90	15.48	73.61
RNAlocopt	75	1.03	1.20	85	6.21	1.92
RNANR	70	**2.07**	2.17	71	3.32	562.75
RNAsubopt	73	1.40	0.02	89	9.56	0.70
Uniform Sampling	0	0.00	0.00	0	0.00	0.00

Table 1: Coverage of the 100 highest saddle points associated to the deepest left and right minima in the barriers tree for ten 100 nt long sequences. Values in 'barriers' columns (resp. 'basins') represent the proportions of best barriers (resp. basins) covered by the tools, while $\bar{t}$ columns report average runtime.

average, at least twice as high as the remaining methods we tested. At the same time, and in contrast to *uniform sampling*, the samples still mostly consist of low free energy representatives. This can not only be observed from the weighted mean base pair distance computations, but even more prominently from the results in our distance class partition function coverage analysis. Note, that the rather high weighted mean base pair distances observed for RNANR and RNAlocopt mostly originate from their respective coarse graining of the solution space. Their intrinsic Boltzmann sampling then to the largest extent generates low free energy representatives in this reduced structure space. This is also clearly visible in the DOS computations, where, despite their access to high free energies, RNANR, RNAlocopt, and RNASLOpt seem to overlook major parts of the low free energy spectrum of the actual state space. Our new method shows a free energy range comparable to that of RNANR. In contrast, however, it almost exhaustively covers the very low free energy spectrum while still showing very good coverage over the entire energy range.

From the above analysis, we conclude that RNAxplorer yields a very good approximation of the actual state space, even for small sample sizes. Although methods like RNANR or RNASLOpt show better performance in many parts of our benchmarks, their application is typically limited to small RNAs due to their exponentially growing demands on computation time and/or memory. For longer RNAs, where the former two approaches can not be applied, RNAxplorer still yields highly diverse sample sets consisting of energetically representative structures. On average, RNAxplorer achieves sample qualities that are at least as good as those of RNAlocmin, while having a much smaller computational footprint.

Still, the results of RNAxplorer very much depend on the penalty values and the granularity at which they are set. For too high penalties only a few unique but very diverse structures are obtained. Too low penalties, on the other hand, yields structures that are more or less Boltzmann distributed. The former is similar to the 'Zuker suboptimals effect' (described by Wuchty *et al.* (1999)), where potentially important structures are neglected due to partially shared substructures. In RNAxplorer we aim to avoid this effect by increasing the number of samples drawn in each round, as well as through a very moderate default energy penalty of $\alpha = kT$. Furthermore, different user settings of the proportion factor μ allows for good control of whether a sampling round was sufficient and a penalty should be applied.

Relationship to continuous energy landscapes

It should be noted, that the application of penalizing pseudo-energy potentials is similar to the concept of meta dynamics simulations on continuous energy landscapes (Laio and Parrinello, 2002), in particular the Local Elevation (LE) method, as used for Monte Carlo protein folding simulations (Huber *et al.*, 1994). However, for the discrete energy landscapes of RNA secondary structures, we can use efficient methods to compute the partition function and to sample from the entire Boltzmann distributed ensemble. Thus, approximations of the landscape can be directly obtained from the samples rather than from time-consuming Monte Carlo simulations. Furthermore, the RNA folding grammar does not allow for the application of Gaussian potentials as required for the LE method, but is rather limited to potentials that linearly depend on particular structural features.

Funding

This work has been supported by the Austrian/French project RNALands (ANR-14-CE34-0011 & FWF-I-1804-N28), and by the Austrian science fund FWF project SFP F43 Regulation of the RNA transcriptome.

References

Becker, O. M. and Karplus, M. (1997). The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of chemical physics*, **106**(4), 1495–1517.

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

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 *ISMB*, pages 88–91.

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

Ding, Y., Chan, C. Y., and Lawrence, C. E. (2005). RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna*, **11**(8), 1157–1166.

Eddy, S. R. (1999). Noncoding RNA genes. *Current Opinion in Genetics & Development*, **9**(6), 695–699.

Entzian, G. and Raden, M. (2019). pourRNA–a time- and memory-efficient approach for the guided exploration of RNA energy landscapes. *Bioinformatics*, **36**(2), 462–469.

Flamm, C., Fontana, W., Hofacker, I. L., and Schuster, P. (2000). RNA folding at elementary step resolution. *RNA*, **6**(3), 325–338.

Flamm, C., Hofacker, I. L., Stadler, P. F., and Wolfinger, M. T. (2002). Barrier trees of degenerate landscapes. *Zeitschrift für physikalische chemie*, **216**(2), 155.

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

Giegerich, R., Voß, B., and Rehmsmeier, M. (2004). Abstract shapes of RNA. *Nucleic Acids Research*, **32**(16), 4843–4851.

Hofacker, I. L., Schuster, P., and Stadler, P. F. (1998). Combinatorics of rna secondary structures. *Discrete Applied Mathematics*, **88**(1-3), 207–237.

Huber, T., Torda, A. E., and Van Gunsteren, W. F. (1994). Local elevation: a method for improving the searching properties of molecular dynamics simulation. *Journal of computer-aided molecular design*, **8**(6), 695–708.

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.

Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, **99**(20), 12562–12566.

Li, Y. and Zhang, S. (2011). Finding stable local optimal RNA secondary structures. *Bioinformatics*, **27**(21), 2994–3001.

Lorenz, R., Flamm, C., and Hofacker, I. L. (2009). 2D projections of RNA folding landscapes. In *German conference on bioinformatics 2009*. Gesellschaft für Informatik eV.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, **6**(1), 26.

Lorenz, R., Hofacker, I. L., and Stadler, P. F. (2016). RNA folding with hard and soft constraints. *Algorithms Mol. Biol.*, **11**(1), 8.

Lorenz, W. A. and Clote, P. (2011). Computing the partition function for kinetically trapped RNA secondary structures. *PLoS One*, **6**(1), e16178.

Maňuch, J., Thachuk, C., Stacho, L., and Condon, A. (2009). Np-completeness of the direct energy barrier problem without pseudoknots. In *International Workshop on DNA-Based Computers*, pages 106–115. Springer.

Markham, N. R. and Zuker, M. (2008). *UNAFold*, pages 3–31. Humana Press, Totowa, NJ.

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

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

Ponty, Y. (2008). Efficient sampling of RNA secondary structures from the boltzmann ensemble of low-energy. *Journal of mathematical biology*, **56**(1-2), 107–127.

Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, **11**(1), 129.

Sahoo, S. and Albrecht, A. A. (2012). Approximating the set of local minima in partial RNA folding landscapes. *Bioinformatics*, **28**(4), 523–530.

Saito, S., Kakeshita, H., and Nakamura, K. (2009). Novel small RNA-encoding genes in the intergenic regions of Bacillus subtilis. *Gene*, **428**(1), 2–8.

Turner, D. H. and Mathews, D. H. (2009). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, **38**(suppl_1), D280–D282.

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.

Wuchty, S., Fontana, W., Hofacker, I. L., and Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers: Original Research on Biomolecules*, **49**(2), 145–165.

Xayaphoummine, A., Bucher, T., Thalmann, F., and Isambert, H. (2003). Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **100**(26), 15310–15315.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, **9**(1), 133–148.