Bioinformatics, 37(15), 2021, 2126–2133
doi: 10.1093/bioinformatics/btab066
Advance Access Publication Date: 4 February 2021
Original Paper

Structural bioinformatics

RNAxplorer: harnessing the power of guiding potentials to sample RNA landscapes

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

[1]Department of Theoretical Chemistry, Faculty of Chemistry, University of Vienna, Vienna 1090, Austria, [2]Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna 1090, Austria, , [3]LIX, CNRS UMR 7161, Ecole Polytechnique, Institut Polytechnique de Paris, Paris, France and [4]Department of Cell and Developmental Biology, Center for Anatomy and Cell Biology, Medical University of Vienna, Vienna 1090, Austria

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on July 3, 2020; revised on December 16, 2020; editorial decision on January 22, 2021; accepted on February 2, 2021

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 folding space. Abstractions are therefore needed to simplify downstream analyses, and thus 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 to efficiently explore the structure space of RNAs. `RNAxplorer` uses dynamic programming to perform an efficient Boltzmann sampling in the presence of guiding potentials, which are accumulated into pseudo-energy terms and reflect similarity to already well-sampled structures. This way, we effectively steer sampling toward underrepresented or unexplored regions of the structure space.

Results: 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, yields rare conformations that may be inaccessible to other sampling methods and is better at finding the most relevant kinetic traps in the landscape. Thus, it produces a more representative coarse graining of the landscape, which is well suited to subsequently compute better approximations of RNA folding kinetics.

Availabilityand implementation: https://github.com/ViennaRNA/RNAxplorer/.

Contact: andrea.tanzer@meduniwien.ac.at or 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 (Djebali *et al.*, 2012; ENCODE Project Consortium, 2012). Accordingly, genome-wide bioinformatics studies (Eddy, 1999; Saito *et al.*, 2009) have confirmed the dense population of the intergenic space with transcripts, 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 toward understanding their function.

At the secondary structure level, efficient dynamic programming (DP) algorithms enable the computation of various RNA structural properties at 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, 2010).

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 within its life time. Concrete instances of kinetics, where the thermodynamic ground state is not the final state, notoriously include RNAs whose function is mediated by co-transcriptional folding, e.g. transcriptional riboswitches. Examples of such riboswitches are discussed in Breaker (2012) and most of our benchmark sequences (see Supplementary Table S1) belong to this class. Since experimental methods are limited in the number of alternative structures they can verify, computational models are essential for studying riboswitches in detail. For instance, experimentally derived NMR structures (Helmling *et al.*, 2017) have been used to model concentration-dependent metabolite binding/unbinding kinetics (Wolfinger *et al.*, 2018). Furthermore, kinetic methods are invoked in rational design of artificial riboswitches (Günzel *et al.*, 2020). Folding dynamics can also be modeled on the level of tertiary structures, e.g. for G-Quadruplexes (Stadlbauer *et al.*, 2016). Due to its computational complexity, however, only small substructures can be analyzed on a coarse grained level.

A general framework for studying kinetics relies on an abstraction of the folding process as a Continuous-Time Markov Chain (CTMC) over a discrete 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 critical in order to allow for the omission of large parts of the conformational space, while at the same time maintaining key states in the RNA landscape for subsequent accurate approximation of RNA folding kinetics. Available approaches for coarse-graining include flooding strategies (Entzian and Raden, 2019; Wolfinger *et al.*, 2004), 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 (Kucharík *et al.*, 2014; Li and Zhang, 2011; Lorenz and Clote, 2011; Michálik *et al.*, 2017).

However, the accuracy of virtually all the aforementioned methods is hindered by a strong bias toward low-free energy structures. This situation leaves 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 rates requires approximating the energy barrier between two secondary structures, a NP-hard problem even under simplistic assumptions (Maňuch *et al.*, 2011). 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 are 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 backtracking (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). In the context of RNA secondary structures guiding potentials have previously been used in path finding (Dotu *et al.*, 2010). 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 Materials and methods

Formally, given an RNA sequence σ of length n, a secondary structure $s(\sigma) = \{(i,j)|(\sigma[i], \sigma[j]) \in BP)\}$ is a set of base pairs (i,j) compatible with σ. Interacting nucleotides are usually restricted to the canonical Watson-Crick pairs (A,U) and (G,C) and the Wobble pair (G,U), i.e. $BP = \{(A,U),(U,A),(G,C),(C,G),(G,U),(U,G)\}$. The generally accepted definition of secondary structures also excludes pseudo-knots and assumes a minimum of three unpaired bases between any two pairing bases due to sterical reasons. A detailed definition is given in Supplementary Section S1.1.

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 only use Ω instead of $\Omega(\sigma)$ 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 $E : s \rightarrow \mathcal{R}$ that assigns each structure $s \in \Omega$ a real numbered value, one obtains the notion of the energy landscape

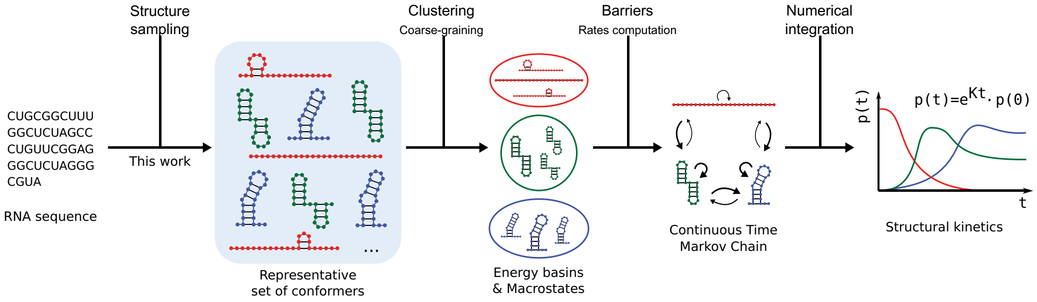

Fig. 1. Typical workflows for RNA folding kinetics start with a sequence, for which a representative subset is sampled. This is followed by a coarse graining, rate computation and the final kinetics computation as a Markov Process

$\ell = \{\Omega, \mathcal{M}, E\}$. 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 neighboring 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), which employs lexicographical order to break ties between structures with equal energy, such that the mapping of a structure to its basin representative structure becomes unique.

Moreover, gradient basins and the minimal saddle points connecting them can be used to conveniently visualize and compare high-dimensional energy landscapes as 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 the length (Waterman, 1978).

Equilibrium ensemble properties. Most RNA secondary structure prediction methods borrow a key concept of statistical mechanics, namely that structures s in thermodynamic 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. $s_{\text{MFE}} = \text{argmin}_{s\in\Omega} E(s)$ 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). 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 Boltzmann sampling, i.e. to draw structures s randomly from the ensemble according to their probability $p(s)$. This can be regarded as (random) backtracing in the DP matrices with worst case time complexity of $\mathcal{O}(n\log(n))$ per sample (Ding and Lawrence, 2003; Ponty, 2007).

2.1 The `RNAxplorer` method

The `RNAxplorer` method approximates RNA energy landscapes using an iterative scheme which samples random structures using 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 away from a (set of) structure(s). Pseudo energy terms are accumulated after each iteration to avoid a concentration of samples within low free-energy basins, thus ensuring maximal coverage of the landscape. This allows a finer level of control over the redistribution of the emission probabilities than previous alternatives, such as the temperature elevation method introduced by Kucharík *et al.* (2014) (see Supplementary Material S1.3).

Sampling with base pairs-associated guiding potentials. Given a pseudo energy $E_\Psi(s)$, our sampling procedure considers a pseudo-energy function $\hat{E}(s) = E(s) + E_\Psi(s)$ where $E(s)$ is the classic Turner free-energy and $E_\Psi(s)$ is a guiding potential defined below. Our goal is then to sample from the distribution

$$\hat{p}(s) = \frac{e^{-\beta \hat{E}(s)}}{\hat{Z}} \quad \text{with} \quad \hat{Z} = \sum_{s\in\Omega} e^{-\beta \hat{E}(s)}. \tag{2}$$

Boltzmann sampling requires the precomputation of the (pseudo) partition function $\hat{Z}$, not through exhaustive summation due to the combinatorial explosion of Ω, but rather by 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 with derivations of the underlying folding grammar. Sampling under such guiding potentials is generically supported by the soft constraints framework introduced by Lorenz *et al.* (2016).

In particular, let us consider simple, base pairs-associated potentials, which can be decomposed into energy terms $E^{i,j} \in \mathbb{R}$, each associated to a base pair (i, j). For any structure s, one has

$$E_\Psi(s) = \sum_{(i,j)\in[1,n]^2} \delta_{i,j}(s) \cdot E^{i,j} \tag{3}$$

where $\delta_{i,j}(s)$ is the indicator function, taking value 1 if $(i, j) \in s$ and 0 otherwise. Despite their simplicity, E_Ψ terms can be used to steer the sampling toward/away from one or several reference structure(s).

For instance, the base pairs of a reference structure s' can be individually penalized/promoted by setting $E^{i,j} = \delta_{i,j}(s') \cdot \alpha$, for some arbitrary real α, leading to $E_\Psi(s) = |s \cap s'| \cdot \alpha$. Setting $\alpha < 0$ will decrease the expected distance between sampled structures to s', while $\alpha > 0$ will increase it. Note that more elaborate guiding potentials can be supported, e.g. through variations of the energy values α and/or a combination of using individual base pairs and structures as targets (see Supplementary Sections S2.2 and S2.3).

Defining guiding potentials to avoid recurrent structures. In order to steer sampling away from a given structure s' that has already been sampled repeatedly, we consider a guiding potential

$$E_c(s', s) := \alpha \cdot \frac{|s \cap s'|}{|s'|} \tag{4}$$

that for each structure $s \in \Omega$ adds individual pseudo-energy penalties, depending on the number of base pairs s shares with s'.

Moderate penalties arise if the weight factor α is chosen close to thermal fluctuations, which is why our method defaults to $\alpha = kT$ unless stated otherwise. Such potentials could also be used to attract subsequent sampling toward a region of interest within the kinetics landscape, by simply changing the sign of α.

Finally, guiding potentials can be modified to capture the base pair distance between s' and s, i.e. the minimum number of base pairs to insert/remove to transform s' structure into s (see Supplementary Equations 9 and 12). This alternative definition yields comparable, yet slightly inferior results as shown in the Supplementary Material S2.1.

Overall iterative strategy. For each structure s, the pseudo-energy $E_\Psi(s)$ is initially set to 0 and incrementally updated to accumulate contributions from the dominant structures encountered over the course of sampling.

At each round m, a multiset $\mathcal{S}_m \subseteq \Omega$ of structures is sampled from a (distorted) Boltzmann distribution. Through gradient descent, each structure $s \in \mathcal{S}_m$ is mapped to its local minimum $\gamma^\infty(s)$, used as a representative $\hat{s}$ for its energy basin. The resulting set of local minima is then analyzed to identify the most over-represented structure, denoted as

$$s'_m = \text{argmax}_{\hat{s}\in\Omega} |\mathcal{B}_{\mathcal{S}_m}(\hat{s})|, \quad \text{with} \quad \mathcal{B}_{\mathcal{S}}(\hat{s}) := \{s \in \mathcal{S} | \gamma^\infty(s) = \hat{s}\}.$$

In other words, s'_m is the local minimum that attracts the most samples in $\mathcal{S}_m$. In the (unlikely) case of ties, one of the most highly represented structure is chosen arbitrarily and returned.

The method then updates the pseudo-energy term E_Ψ for the next iteration, based on the structural features of s'_m, by setting:

$$E_{\Psi_{m+1}}(s) = E_{\Psi_m}(s) + E_c(s'_m, s) = \sum_{\ell=1}^{m} E_c(s'_\ell, s). \tag{5}$$

This can be expressed at level of individual base pairs (i, j) by setting

$$E_{i,j}^{m+1} := \alpha \sum_{\ell=1}^{m} \frac{\delta_{i,j}(s'_\ell)}{|s'_\ell|}$$

where $\delta_{i,j}(s'_\ell)$ denotes the presence (1) or absence (0) of (i, j) in s'_ℓ. Then, any structure s inherits a total pseudo-potential of:

$$\begin{aligned}\sum_{(i,j)\in s} E_{i,j}^{m+1} &= \sum_{(i,j)} \delta_{i,j}(s)\alpha \sum_{\ell=1}^{m} \frac{\delta_{i,j}(s'_\ell)^{ij}}{|s'_\ell|} \\ &= \alpha \sum_{\ell=1}^{m} \frac{|s \cap s'_\ell|}{|s'_\ell|} = \sum_{\ell=1}^{m} E_c(s'_\ell, s) = E_{\Psi_{m+1}}(s)\end{aligned}$$

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

The total number of iterations $m_{\max} = \frac{N}{g}$ is governed by the fraction of the requested sample size N and a user-adjustable granularity g that determines the number of samples drawn at once in each round. Unless stated otherwise, we use default values $N = 10^5$ and $g = 100$.

Additionally, we use a strategy which determines whether E_Ψ needs an update after each iteration. This avoids unnecessary recomputation of the rather costly partition function (see Section 2.1). 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). Therefore we compare the set $\mathcal{M}^1$ of local minima that are observed only once over all iterations, against $\mathcal{M}^{>1}$, those observed multiple times. Our algorithm then only updates E_Ψ if the ratio $|\mathcal{M}^1|/|\mathcal{M}^{>1}|$ does not exceed a saturation threshold μ, set by default to $\mu = 0.1$ by analogy to Kucharík *et al.* (2014).

2.2 Quality assessment

Whether or not a landscape ℓ is adequately approximated by a set of structures strongly depends on the requirements of downstream analysis and thus is difficult to generalize. In the following we discuss the use of general measures for structure diversity, distance class based diversity measures and the coverage of basins and energy barriers in ℓ. The approximated shape of ℓ is important for the overall dynamical behavior of subsequent folding simulations. We, therefore, also analyze sample sets for the presence of certain key structures.

General measures. Typical measures that express the diversity of a set of structures are the number of unique local minima, the Density of States (DOS) (Cupal *et al.*, 1997) and the weighted mean base pair distance. Density of States is simply the number of structures per energy level. The weighted mean pairwise distance is defined as the sum over all base pair distances between structures s and t multiplied by their probabilities $p(s)$ and $p(t)$. For details we refer to Supplementary Material S5.1.

Distance classes. The partitioning of Ω into distance classes with respect to one or many reference structures leads to a projection of the high-dimensional state space. Such a projection still captures the structural diversity of the ensemble, but can now be easily visualized and characterized. 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

$$\text{MFE}^{d_1,d_2} = \min_{s\in\mathcal{C}^{d_1,d_2}} E(s), \tag{6}$$

or the corresponding ensemble free energy

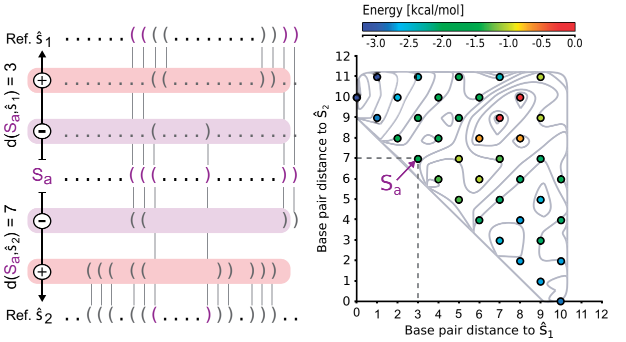

Fig. 2. Toy example of a 2D projection. Each colored circle corresponds to a valid base pair distance to both reference structures. The color corresponds to the energy of the MFE structure representative of this spot. The background consists of isolines which are created from an interpolation of the free energy values that correspond to the MFE structures of each circle. Structure S_a has base pair distance three to the reference $\hat{s}_1$, because one base pair has to be removed and two added in order to change $\hat{s}_1$ into S_a. The distance to $\hat{s}_2$ is seven because two base pairs must be removed and five inserted to change S_a into $\hat{s}_2$. (Color version of this figure is available at *Bioinformatics* online.)

$$G^{d_1,d_2} = -\beta \ln Z^{d_1,d_2}, \quad \text{with} \quad Z^{d_1,d_2} = \sum_{s\in\mathcal{C}^{d_1,d_2}} e^{-\beta E(s)}. \tag{7}$$

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 a class to be covered, 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. Supplementary Section S5.5.

Energy barriers. Comparing folding simulations produced by different tools can be challenging, because the inherently different coarse graining of each sampling method results in different representations of both fast fluctuations at the beginning of the simulation and slow folding components close to thermodynamic equilibrium. To assess the quality of our sampling strategies within the RNA folding kinetics workflow (Fig. 1) we need to evaluate our samples in terms of providing a basis for calculating folding rates. For this, samples must not only cover the lowest energy states of ℓ, but also refolding events with large energy barriers, i.e. those associated with slow rates that effectively determine the long time behavior of folding dynamics (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 ensemble. We 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 Supplementary Section S5.3.

2.3 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 which are then integrated into the prediction algorithms (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 described in Section 2.1 using the programming language C, as part of the executable program `RNAxplorer`. For the iterative sampling method, the user can choose between two guiding potentials, the base pair associated potential, described in Section 2.1, or an alternative based on base pair distance, described in Supplementary Section S2.1. To deviate from the default parameters, the granularity g, the total sample size N, the saturation threshold μ, as well as the weighting factor α are user-adjustable. The `RNAxplorer` program offers additional guiding potential-based structure sampling modes, e.g. one that

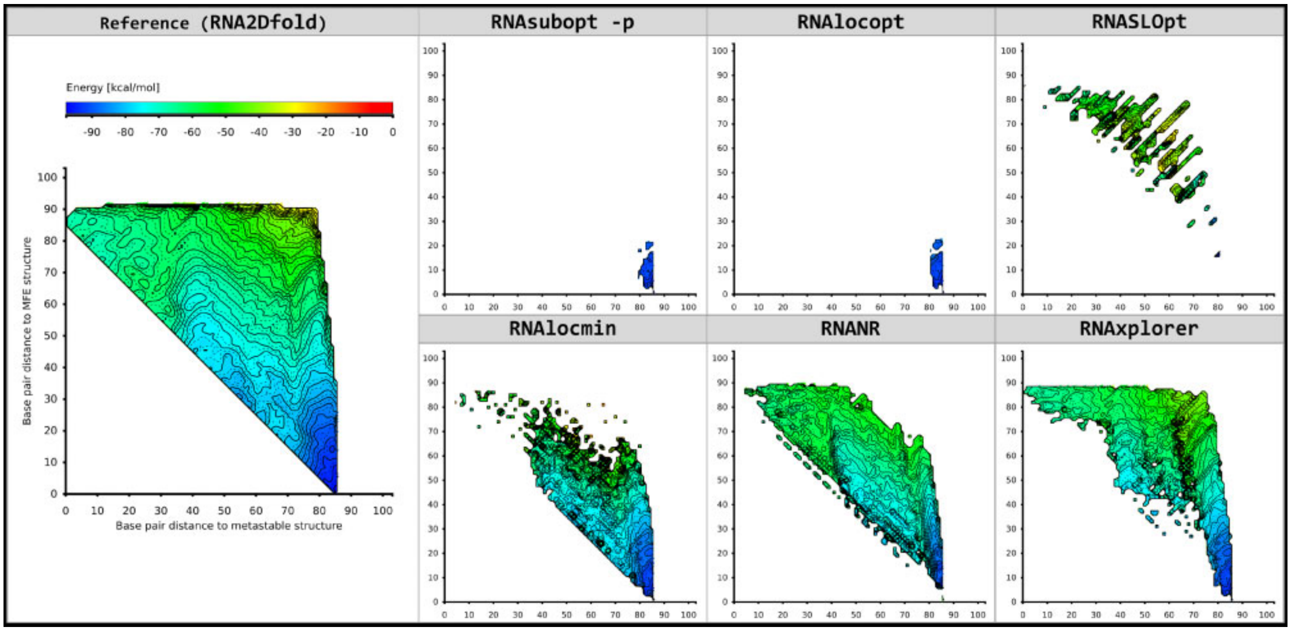

Fig. 5. 2D projections of local minima as obtained from different methods for the riboswitch SV-11 Q beta replicase template (Biebricher and Luce, 1992). Reference structures for the projection are the MFE and metastable structure. The left most column depicts the *ground truth* as computed by RNA2Dfold, chosen here as a reference for comparison. 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 6.75, 21.81, 115.99, 487.73, 4285.81 and 27.87 s to produce the sample sets, respectively. The sample size for each tool is 10^6 (except for RNASLOpt which always yields less structures even with exhaustive enumeration, i.e. artificially high ε)

explicitly penalizes or rewards a pre-defined set of reference structures and structures in its vicinity. Furthermore, it implements different heuristics to compute (optimal) transition paths to eventually determine saddle points required to assess transition rates, and finally, provides gradient walk methods to coarsen the sampled state space. The program also comes with a Python script that enables hierarchical clustering of secondary structures, see Supplementary Section S2.3.

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 were made. They have a minimum length of 110 nt, a maximum length of 233 nt and a median length of 130 nt. The exact sequences can be found in Supplementary Table S1 of Supplementary Material. The methods and tools we compare our approach against are (i) *uniform sampling* with RNAsubopt achieved using Boltzmann sampling at extremely high temperatures ($10^{6\,\circ}$), (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 of RNAlocmin (Kucharík *et al.*, 2014) and (vi) a set of locally stable 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 ε 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 ε 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.

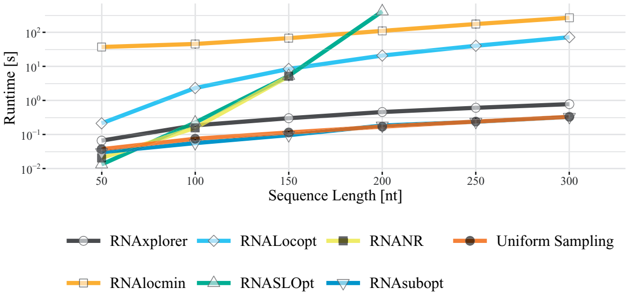

Fig. 3. Runtime comparison. Runtimes observed for a sample of 1000 structures for RNAs with lengths from 50 to 300 nucleotides, averaged over 10 randomly generated sequences. For RNASLOpt, we precomputed a ε value in order to obtain at least 1000 structures. For RNAxplorer and RNAlocmin the number of iterations was set to 100

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. Programs were compiled with GCC 8.2.1 and all computations were performed on a Linux 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

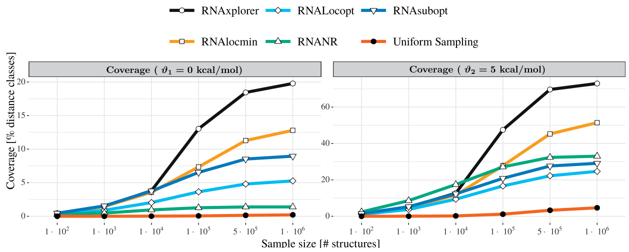

Fig. 4. Distance class coverage as a function of sample size. Shown are the fractions of distance classes $\mathcal{C}^{d_1,d_2}$ covered by at least one local minimum. The local minima are derived from the sample set and have to be energetically close to the respective MFE^{d_1,d_2}. The data averages 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)

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 3 days) for sequence longer than 150 nt (RNANR) and 200 nt (RNASLOpt) due to the limited memory of our testing machine. However, for shorter RNA sequences up to 150 nt, these two programs are still faster than RNAlocmin (Fig. 3). Further runtime and memory benchmarking results are available in Supplementary Section S4.

3.2 Structure sample diversity

For each method we assessed sampling quality in terms of diversity of structures obtained. First we calculated standard measures to provide an overall description of the samples produced. Sampling redundancy in terms of (i) number of unique local minima and (ii) mean base pair distance both turn out favorable for RNAxplorer and RNANR (see Supplementary Section S5.1 and Supplementary Fig. S5). The energy spectrum of the samples expressed as Density of States shows that most methods are prone to over-sample the low free energy regime, while RNAxplorer also captures structures at higher energy levels, e.g. structures around the meta-stable state of SV-11 appear as second peak (Supplementary Fig. S6). However, these results are not sufficient to evaluate the methods regarding their suitability within the folding kinetics workflow. In the following we therefore focus on a newly developed measure to investigate the spatial resolution of the sample sets based on distance classes.

Coverage of distance classes. Our main question was 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), where 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.

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 Supplementary Figure S16. 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 (more than 200GB), 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 Supplementary Figure S16. The analog measure based on partition functions is shown in Supplementary Figure S18 for individual sequences and in Supplementary Figure S19 as average over all sequences. For more details on the coverage measure see Supplementary Section S5.5.

Table 1. Coverage of barrier trees for ten 100 nt long sequences, using sample sizes of 10^3 and 10^5, respectively

	Sample size					
Tool	#Structures = 10^3			#Structures = 10^5		
Length: 100 nt	Coverage [%]			Coverage [%]		
Max. #basins: 63536	Barriers	Basins	$\bar{t}$ [s]	Barriers	Basins	$\bar{t}$ [s]
RNAxplorer	**78.60**	1.67	0.06	**93.50**	**22.25**	3.72
RNAlocmin	74.80	1.46	41.19	90.20	15.48	73.61
RNAlocopt	75.00	1.03	1.20	85.30	6.21	1.92
RNANR	70.40	**2.07**	2.17	70.80	3.32	562.75
RNAsubopt	72.70	1.40	0.02	89.00	9.56	0.70
Uniform Sampling	0.00	0.00	0.00	0.00	0.00	0.00

Note: Values in 'barriers' columns show coverage of the 100 highest saddle points associated to the deepest left and right minima in the barriers tree. For columns 'basin' the percentage of total basins covered is shown. $\bar{t}$ columns report average runtime in seconds. The highest coverage for barriers and basins is shown in bold.

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 could be found based on the samples. 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 Supplementary Section S5.3.

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 main contribution to the asymptotic time complexity 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. For sequences of length n and a total number of structures N to sample, the upper limit on the runtime becomes $\mathcal{O}\left(\frac{N}{g}n^3 + Nn^2\right)$. Choosing $g \approx n$ ensures that the folding and sampling part of the program have approximately equal costs, leading to a worst-case asymptotic complexity in $\mathcal{O}(Nn^2)$. Moreover, most sampling rounds do not satisfy the saturation criterion, so a typical run of `RNAxplorer` requires much less than $\frac{N}{g}$ recomputations of the partition function, further reducing its practical computational demand.

While RNAxplorer has a number of tunable parameters, these parameters have default values that should work well for almost any application. Users can adjust these parameters manually, but are invited to proceed with caution. Indeed, setting the weight factor α to a low value, yields structures that are approximately Boltzmann distributed, and mainly populates the MFE basin. On the other hand, α should not be much larger than kT to ensure that the Boltzmann distributions of consecutive iterations have sufficient overlap.

Coverage of distance classes. The coverage of distance classes is a combined measure which consists of important local minima and structural diversity based on 2D projections. For small sample sizes (less than 10^4), `RNANR`, `RNAlocmin` and `RNAlocopt`, could be alternative methods with comparable quality. For larger sample sizes (10^5–10^6), `RNAxplorer` clearly outperforms the other methods. Although only at most 35% of local minima are uniquely sampled, the samples are diverse and `RNAxplorer` covers more than 70% of the projected landscapes with low energy structures, for 20% of the landscape we even identify the local minimum (Fig. 4).

Studying the topology of a landscape projection can help to choose the most efficient sampling strategy for a given problem. If, for instance, only one local minimum is present, simple Boltzmann sampling might suffice. In cases with additional metastable states, guiding potentials are the method of choice, because they steer the sampling procedure directly to structures far away of the MFE structure. This scenario is exemplified by our data for riboswitch SV-11, (Fig. 5), where only `RNAxplorer` identified the two functional states of the switch. `RNANR` produces a projection with a similar cell coverage, however, it does not find as many local minima as `RNAxplorer`, which is shown by the less intense coloring of cells compared to the reference 2D plot (Fig. 5), the lower coverage of distance classes (Supplementary Fig. S16) and the DOS (Supplementary Fig. S6). `RNAlocmin` could not find the metastable state, because at later stages, i.e. at higher temperatures, it turns into uniform sampling, which results in mostly high free energy structures and thus misses potentially important local minima. In the 2D projection this can be seen as separate spots in the higher energy areas (Fig. 5).

Coverage of energy landscapes. The coverage of distance classes already indicates whether important local minima have been sampled. Using `barriers` we test for support of important transitions over high energy barriers by comparing all samples to a ground truth calculated by exhaustive enumeration. Unfortunately, `barriers` is limited to RNAs smaller than 100 nt due to time and memory constraints. For this reason we cannot use our benchmarking set of natural RNAs, but created random sequences in the range of 50 to 100 nt.

`RNAxplorer` covers the largest energy barriers much better than other tools, even for small sample sizes. Although, `RNAlocmin`, `RNAlocopt` and `RNAsubopt` produce a comparable high number of largest barriers, they cover a much smaller fraction of basins (Table 1). `RNAxplorer` covers much more basins and thus provides in addition to the major transitions more detailed information on fast refolding processes. With growing sequence length, `RNAxplorer` outperforms the other methods in terms of covered barriers and basins, as well as run time. Thus, we show that `RNAxplorer` yields a very good approximation of the actual state space and is better suited for fast and efficient sampling of long sequences.

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.

RNAxplorer in the folding kinetics workflow. To summarize our benchmark study, we have demonstrated that `RNAxplorer` is well suited to serve as sampling tool within the RNA folding kinetics workflow (see Fig. 1). It creates both diverse and low free energy structures much faster than other tools. As a result, it covers most of the basins and largest barriers which are crucial to model the long term folding behavior. Thus, the samples produced by `RNAxplorer` sufficiently capture the relevant areas of the structure space and we think that `RNAxplorer` sets a mile stone in computing the folding kinetics for RNAs of 300 nt and beyond.

Funding

This work was supported by the Austrian/French project 'RNALands' (ANR-14-CE34-0011 & FWF I 1804 N28), the FWF Elise-Richter project 'RNAmod' (FWF V 762) and the Austrian science fund FWF project 'Regulation of the RNA' transcriptome (FWF SFPF43).

Conflict of Interest: none declared.

References

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

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

Breaker,R.R. (2012) Riboswitches and the RNA world. *Cold Spring Harbor Perspect. Biol.*, **4**, a003566–a003566.

Cupal,J. *et al.* (1997) Density of states, metastable states, and saddle points exploring the energy landscape of an RNA molecule. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 88–91.

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

Ding,Y. *et al.* (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.

Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Dotu,I. *et al.* (2010) Computing folding pathways between RNA secondary structures. *Nucleic Acids Res.*, **38**, 1711–1722.

Eddy,S.R. (1999) Noncoding RNA genes. *Curr. Opin. Genet. Dev.*, **9**, 695–699.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

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

Flamm,C. *et al.* (2000) RNA folding at elementary step resolution. *RNA*, **6**, 325–338.

Flamm,C. *et al.* (2002) Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, **216**, 155.

Freyhult,E. *et al.* (2007) Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, **23**, 2054–2062.

Giegerich,R. *et al.* (2004) Abstract shapes of RNA. *Nucleic Acids Res.*, **32**, 4843–4851.

Günzel,C. *et al.* (2020) Beyond plug and pray: context sensitivity and in silico design of artificial neomycin riboswitches. *RNA Biol.*, **25**,1-11.

Helmling,C. *et al.* (2017) NMR structural profiling of transcriptional intermediates reveals riboswitch regulation by metastable RNA conformations. *J. Am. Chem. Soc.*, **139**, 2647–2656.

Huber,T. *et al.* (1994) Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J. Comput. Aided Mol. Des.*, **8**, 695–708.

Kucharík,M. *et al.* (2014) Basin Hopping Graph: a computational framework to characterize RNA folding landscapes. *Bioinformatics*, **30**, 2009–2017.

Laio,A. and Parrinello,M. (2002) Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA*, **99**, 12562–12566.

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

Lorenz,R. *et al.* (2009) 2D projections of RNA folding landscapes. In: Grosse,I. *et al.* (eds) *German Conference on Bioinformatics 2009*. Gesellschaft für Informatik eV, Bonn, Germany.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Lorenz,R. *et al.* (2016) RNA folding with hard and soft constraints. *Algorithms for Mol. Biol.*, **11**, 8.

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

Mañuch,J. *et al.* (2011) NP-completeness of the energy barrier problem without pseudoknots and temporary arcs. *Nat. Comput.*, **10**, 391–405.

Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3-31.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolym. Original Res. Biomol.*, **29**, 1105–1119.

Michálik,J. *et al.* (2017) Efficient approximations of RNA kinetics landscape using non-redundant sampling. *Bioinformatics*, **33**, i283–i292.

Ponty,Y. (2007) Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy. *J. Math. Biol.*, **56**, 107–127.

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

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

Saito,S. *et al.* (2009) Novel small RNA-encoding genes in the intergenic regions of *Bacillus subtilis*. *Gene*, **428**, 2–8.

Stadlbauer,P. *et al.* (2016) Coarse-grained simulations complemented by atomistic molecular dynamics provide new insights into folding and unfolding of human telomeric g-quadruplexes. *J. Chem. Theory Comput.*, **12**, 6077–6097.

Turner,D.H. and Mathews,D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38**, D280–D282.

Waterman,M. (1978) Secondary structure of single-stranded nucleic acids. In: Gian-Carlo,R (ed.) *Studies in Foundations and Combinatorics, Volume 1 of Advances in Mathematics: Supplementary Studies*. Academic Press, New York, NY, pp. 167–212.

Wolfinger,M.T. *et al.* (2004) Efficient computation of RNA folding dynamics. *J. Phys. A Math. Gen.*, **37**, 4731–4741.

Wolfinger,M.T. *et al.* (2018) Efficient computation of co-transcriptional RNA–ligand interaction dynamics. *Methods*, **143**, 70–76.

Xayaphoummine,A. *et al.* (2003) Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. USA*, **100**, 15310–15315.

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