BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 4 2012, pages 523–530
doi:10.1093/bioinformatics/btr715

Structural bioinformatics

Advance Access publication December 30, 2011

Approximating the set of local minima in partial RNA folding landscapes

S. Sahoo and A.A. Albrecht*

Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast BT9 7BL, UK

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: We study a stochastic method for approximating the set of local minima in partial RNA folding landscapes associated with a bounded-distance neighbourhood of folding conformations. The conformations are limited to RNA secondary structures without pseudoknots. The method aims at exploring partial energy landscapes pL induced by folding simulations and their underlying neighbourhood relations. It combines an approximation of the number of local optima devised by Garnier and Kallel (2002) with a run-time estimation for identifying sets of local optima established by Reeves and Eremeev (2004).

Results: The method is tested on nine sequences of length between 50nt and 400nt, which allows us to compare the results with data generated by RNAsubopt and subsequent barrier tree calculations. On the nine sequences, the method captures on average 92% of local minima with settings designed for a target of 95%. The run-time of the heuristic can be estimated by $O(n^2 D\nu \ln\nu)$, where n is the sequence length, ν is the number of local minima in the partial landscape pL under consideration and D is the maximum number of steepest descent steps in attraction basins associated with pL.

Contact: a.albrecht@qub.ac.uk

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

Received on May 11, 2011; revised on November 17, 2011; accepted on December 23, 2011

1 INTRODUCTION

The method presented in the article targets sampling algorithms designed for RNA secondary structure prediction and pathway analysis that employs in one way or another features of the underlying free energy landscape. For this type of algorithms, knowledge of metastable states (local minima) in the vicinity of intermediate conformations can be of advantage in speeding up the folding simulation. Here, the term sampling algorithm covers any type of random walks, Markov processes and kinetic or co-transcriptional folding simulations applied to RNA secondary structure prediction. Simulations of kinetic folding based upon the master equation $dp_u(t)/dt = \sum_{v=1}^{m}\left(k_{vu}p_v(t) - k_{uv}p_u(t)\right)$ are feasible only for short sequences due to the large number m of folding conformations (Chen, 2008), even when executed on parallel computing systems ($p_u(t)$ is the probability that the folding process is in conformational state u at time t and the k_{uv} are the transition rates from state u to state v). A standard setting for values k_{uv} is given by $k_{uv} = \exp\left(-(G_v - G_u)/RT\right)$ for $G_v > G_u$ and $k_{uv} = 1$ for $G_v \leq G_u$, where G_w is the free energy of conformation w. The results from numerical simulations of short sequences provide valuable information about transition rates at different folding stages and the stability of helices, hairpins, various types of loops and pseudoknots. The data can then be incorporated into coarse-grain models of folding.

In an attempt of reducing the ensemble of conformations, Zhao *et al.* (2010) combine the master equation and the analysis of free-energy landscapes where kinetic moves are based on the addition and deletion of entire helix stems. The method further develops the analysis of kinetic moves based upon the exchange of entire helices as presented by Xayaphoummine *et al.* (2003, 2005). Xayaphoummine *et al.* (2003) pay particular attention to structures with pseudoknots and how to overcome the problem of kinetic traps consisting of rapidly exchanging states. Zhao *et al.* (2010), instead of completely unwinding two overlapping helices, devise a strategy (arm-by-arm exchange) that requires only a partial melting of the initial helix during the transition to the new metastable state. Prior knowledge of metastable states (local minima) in the vicinity of the current structure could be helpful in executing such transitions and in identifying pathways between such states. The exchange of entire helices is also the basic neighbourhood move analysed by Danilova *et al.* (2006). For structure prediction with pseudoknots, Bellaousov and Mathews (2010) recently devised a fast algorithm that runs in $O(n^2)$ time and utilizes base pairing probabilities instead of changes in free energy values.

Lorenz and Clote (2011) comprehensively study the computation of the partition function over the set of local minima in energy landscapes induced by addition and deletion of single base pairs. The authors devise an algorithm that runs in $O(n^3)$ time and requires $O(n^2)$ space. The underlying energy model is the Turner nearest neighbour model (Xia *et al.*, 1998) without dangles, and the algorithm is an extension of McCaskill's algorithm (McCaskill, 1990) that accounts for locally optimal structures by additional terms in the recursion scheme. Based upon computational experiments over randomly generated RNA sequences, the number of locally optimal structures is estimated to be roughly the square root of the total number of feasible structure of a given length.

A new method for approximating the partition function over all secondary structures has been recently presented by Lou and Clote (2010). The method employs the Wang–Landau algorithm (Wang and Landau, 2010a, b) for approximations of the density of energy states and is less restrictive in regard to local interactions than dynamic programming applied to secondary structure prediction.

An overview of methods for simulations of kinetic folding is given by Flamm and Hofacker (2008) and is also part of the detailed

*To whom correspondence should be addressed.

summary on progress in RNA secondary structure prediction by Schuster (2006). A variety of methods has been proposed for simulations of kinetic folding based upon the addition, deletion and shift of single base pairs. Flamm *et al.* (2000) introduced the shift move, which is a combination of a base pair removal and a base pair addition where one position remains invariant. The shift move aims at the simulation of 'defect diffusion' reported by Pörschke (1974), which denotes a process where, for instance, the position of a bulge in a helix may move along a helix as the result of rapid base pair formation and dissociation. The transition rates k'_{uv} between neighbouring conformations are calculated according to the symmetric rule $k'_{uv} = \exp\bigl(-(G_v - G_u)/2RT\bigr)$.

Geis *et al.* (2008) present details of the Kinwalker tool along with simulation results on 12 RNA sequences with a length ranging from 115nt to 1492nt. The underlying method generates secondary structures by combining building blocks that correspond to thermodynamically optimal substructures. The optimal substructures are calculated by standard dynamic programming. An important step in the execution of Kinwalker is to identify the energy barrier between two locally optimal conformations. The tool utilizes a modification of the Morgan–Higgs heuristic (Morgan and Higgs, 1998), which is designed to find a direct folding pathway of minimum height between two secondary structures. Most of the total run-time of Kinwalker is spent on this particular task. The authors report a good agreement of folding simulations with experimentally verified folding pathways, and the same applies to calculated folding times and corresponding values predicted in the literature.

Hofacker *et al.* (2010) propose a new framework for the analysis of time-variable landscapes. The underlying idea is to model the impact of external triggers (ligand binding, nucleolytic cleavage, RNA elongation) and environmental changes (temperature) at folding dynamics as small but discrete changes in energy landscapes. The authors associate macrostates with attraction basins of local minima in energy landscapes, see also Wolfinger *et al.* (2004). For any two local minima, a 'transition state energy' is defined that equals the energy level at which their associated macrostates (attraction basins) merge in the barrier tree of the landscape induced by individual conformations (microstates). The transition rate between macrostates is then determined similar to the above-mentioned k_{uv}, where the energy values involved are the transition state energy (which is the saddle height between u and v) and the free energy of the locally optimal structure v. The time-variable landscapes are represented by sequences of barrier trees and associated sets of macrostates. The transitions (changes) between successive barrier trees and sets of macrostates are calculated by solving a master equation for population densities of macrostates. The master equation is determined by the aforementioned transition rates between macrostates. Since the transition rates between macrostates are time independent for a given barrier tree, the complexity of solving the master equation is significantly reduced within this specific framework. Applications to an RNA thermometer, to co-transcriptional folding and a specific type of refolding are briefly discussed. The concept of macrostates in the context of RNA folding is also studied in a recent paper by Mann and Klemm (2011) where new sampling techniques for macrostates are proposed.

Tang *et al.* (2005, 2008) utilize methods from motion planning developed for robotics for computing folding pathways. The conformational space is approximated by a graph (roadmap) that connects conformations according to a distance metric defined by the difference in base pairs. The analysis of the roadmap and subsequent computation of folding pathways invokes the master equation.

Dotu *et al.* (2010) present a tabu search algorithm for finding folding pathways of minimum height between two secondary structures. The algorithm employs a weighted fitness function that tries to balance between low free energy and the distance to the target structure. At each step, a base pair that minimizes the fitness function in the 1-distance neighbourhood is added or removed, and this base pair is kept in a tabu list for a certain number of steps. The weight attached to the distance measure is decreased, if the current step leads to a decrease in the distance to the target structure. The distance weight is increased, if the distance to the target has not been improved for a number of iterations, and with the increased weight the procedure is restarted from the structure found so far to be closest to the target. The evaluation on 18 structures (riboswitches) demonstrates the competitiveness of the approach in finding near-optimum pathways.

We study a procedure for approximating sets of locally optimal structures that can be integrated into simulations of kinetic folding or computations of folding pathways between secondary structures. A major motivation is the analysis of metastable RNA secondary structures in the context of microRNA target predictions. Recent prediction tools incorporate information about the strength of secondary structures and subsequent duplex bindings, see Ragan *et al.* (2011) and the literature therein. The likelihood of bindings is usually related to minimum free energy secondary structures. Knowledge about metastable states (local minima) could provide a more comprehensive analysis of potential duplex bindings, in particular, if co-transcriptional folding and the differences in concentrations of mRNA and microRNA are taken into account (Ragan *et al.*, 2011).

Our procedure is designed for the analysis of partial landscapes $p\mathsf{L}$ that cover only a small subset of the entire conformation space. Such $p\mathsf{L}$ can be defined, for example, by neighbourhood relations where the number of neighbourhood transitions (addition or deletion of base pairs) is bounded by a constant. The proposed method consists of two major steps: first, the number ν of locally optimal structures within $p\mathsf{L}$ is approximated by using a stochastic procedure devised by Garnier and Kallel (2002). Secondly, the value of ν_{appr} is taken as an input into an approximation formula established by Reeves and Eremeev (2004) for the number $W(p\mathsf{L}, \nu_{\text{appr}})$ of repeated runs of steepest descent within $p\mathsf{L}$ out of randomly selected conformations, where $W(p\mathsf{L}, \nu_{\text{appr}})$ ensures with a certain high probability that all locally optimal structures are covered. The procedure is tested on sequences of length up to 400nt by using data generated by RNAsubopt (Wuchty *et al.*, 1999) and barriers implementations (Gruber *et al.*, 2008) taken from the Vienna package (Hofacker, 2003).

2 APPROACH

We consider combinatorial landscapes $\mathsf{L} = [\mathsf{C}, \mathsf{N}, f]$ defined by the set of conformations C, the neighbourhood relation N and the fitness function $f : \mathsf{C} \rightarrow \mathbb{R}$. The conformation space C is attached to an individual RNA sequence (usually identified by an accession number) and consists of secondary structures produced by RNAsubopt with standard settings, i.e. no isolated base pairs and at least three nucleotides in loops. The neighbourhood relation N is defined by three operations for single-step transitions $S \rightarrow S' \in \mathsf{N}_S$:

(i) Addition of one or two base pairs: a single base pair is added, if an existing helix is extended; two base pairs are added, if an unpaired position admits such and extension and does not lead to an extension of a helix by two base pairs; the addition must ensure that the condition for the minimum loop size is not violated.

(ii) Deletion of one or two base pairs: a single base pair is deleted as part of a helix, if at least two adjoined base pairs remain; otherwise, two base pairs are deleted.

(iii) Shift of base pair: for positions $u < v < w$, an existing base pair $[u, v]$ is substituted by $[u, w]$, if the resulting structure belongs to C.

The neighbourhood N_S covers all conformations that can be generated by a single application of Op-I, Op-II or Op-III from S. We note that by definition the secondary structure S itself belongs to N_S. For a graphical presentation of the three operations, we refer the reader to Flamm *et al.* (2000), Fig. 2(A–E) and Fig. 3 on p. 327/328. Finally, the fitness function $f : \mathsf{C} \rightarrow \mathbb{R}$ is defined as the free energy $E : \mathsf{C} \rightarrow \mathbb{R}$ calculated by the program `RNAeval` (Hofacker *et al.*, 1994).

Within a given L, the distance $d(S, S')$ between two secondary structures (conformations) S and S' is defined by the Hamming-distance, i.e. if $\#(\text{bp-}S \cap \text{bp-}S')$ is the number of base pairs common to S and S', then $d(S, S') = \#(\text{bp-}S) + \#(\text{bp-}S') - 2\#(\text{bp-}S \cap \text{bp-}S')$. The method proposed in the present paper aims at small subsets of C, or partial landscapes $p\mathsf{L}$, defined by an intermediate $S \in \mathsf{C}$ and all $S' \in \mathsf{C}$ with $d(S, S') \leq k$ that are generated from S by subsequent executions of single-step transitions without violating the distance bound, where k is a constant such that the number of conformations in this vicinity of S is of 'manageable' size.

A simple but key procedure in our approach is the steepest descent algorithm. For $S \in \mathsf{C}$, the following steps are executed within L:

(i) Initialize $S_0 = S$;

(ii) For $u \geq 0$, set $S_{u+1} = \mathtt{argmin}_{S \in \mathsf{N}_{S_u}} f(S)$;

(iii) If $E(S_{u+1}) < E(S_u)$, then $u := u + 1$ and return to (ii), otherwise terminate.

We follow Garnier and Kallel (2002) and Reeves and Eremeev (2004) by assuming that only a single $S = S_{u+1}$ from N_{S_u} minimizes the fitness function $f(S) = E(S)$. In general, there is no guarantee that this holds for energy landscapes induced by RNA secondary structures.

If the partial function $\hat{f} : \mathsf{N}_S \rightarrow \mathbb{R}$ always produces a single minimum, then the steepest descent procedure is deterministic and leads to a partition of L into attraction basins A_u of local/global minima, $u = 1, ..., \nu$, where $A_u \cap A_v = \emptyset$, $\bigcup_{u=1}^{\nu} A_u = \mathsf{L}$, and ν is the total number of local minima, including the global minima. The steepest descent procedure terminates at the local minimum that defines the attractions basin, i.e. in terms of C, A_u consists of all S for which (ii) and (iii) terminate at the local minimum m_u, $u = 1, ..., \nu$.

2.1 Garnier–Kallel method

We briefly describe the approach devised by Garnier and Kallel for the problem setting called 'inverse problem' (Garnier and Kallel, 2002). Let $\alpha_u := |A_u| / |\mathsf{C}|$ denote the normalized size of attraction basin A_u. Garnier and Kallel (2002) assume a parameterized random distribution of the normalized sizes $z = \alpha_u$ of attraction basins with the density function p_γ defined by

$$p_\gamma(z) = \frac{\gamma^\gamma}{\Gamma(\gamma)} \frac{z^{\gamma-1}}{e^{\gamma z}}, \tag{1}$$

where $\gamma > 0$ and $\Gamma(s) := \int_0^\infty e^{-t} t^{s-1} dt$ is the Euler function. The density function p_γ, i.e. the parameter γ, is approximated by a sampling method over attraction basins.

Let $\mathsf{M} := \{S_1, S_2, ..., S_M\} \subset \mathsf{C}$ denote a set of randomly selected secondary structures. For each of the S_u, the steepest descent procedure is executed, which leads to $S_u \rightarrow m_v$, if S_u belongs to the attraction basin A_v spawned by m_v. By $\mathrm{SD}(m_v) = \mathsf{M} \cap A_v$ we denote the set of conformations from M where steepest descent terminates at m_v, $v = 1, ..., \nu$. The set M together with steepest descent identifies a set of local minima denoted by

$$\mathrm{LM}(\mathsf{M}) := \{m | \text{ size of } \mathrm{SD}(m) > 0\}, \tag{2}$$

and we set $\mathrm{lm}(\mathsf{M}) := |\mathrm{LM}(\mathsf{M})|$. Furthermore, let $B_j := \{\mathrm{SD}(m) | \text{ size of } \mathrm{SD}(m) = j\}$ denote the set of sets $\mathrm{SD}(m)$ that have the same size j. As in Garnier and Kallel (2002) we set

$$\beta_j := |B_j|, j \geq 0. \tag{3}$$

By this definition, β_j is the number of local minima 'visited' by exactly j out of $M = |\mathsf{M}|$ conformations after steepest descent. By definition of $\mathrm{lm}(\mathsf{M})$ we then have

$$\mathrm{lm}(\mathsf{M}) = \sum_{j=1}^{M} \beta_j. \tag{4}$$

On the other hand, the number M of initial conformations can be expressed by

$$M = \sum_{j=0}^{M} j \beta_j. \tag{5}$$

Supplementary Figure S1 illustrates the particular values for the case of $\nu = 4$ and $M = 7$.

Let $\beta_{j,\gamma}$ denote the expected value of β_j under the assumption that the normalized sizes α of attraction basins A are distributed according to p_γ defined in (1). Garnier and Kallel (2002) established the relation

$$\beta_{j,\gamma} = \nu \binom{M}{j} \frac{\Gamma(\gamma+j)}{\Gamma(\gamma)} \frac{\Gamma(\nu\gamma)}{\Gamma((\nu-1)\gamma)} \frac{\Gamma((\nu-1)\gamma+M-j)}{\Gamma(\nu\gamma+M)} \tag{6}$$

for $j = 1, ..., M$. In Equation (6), the value of ν is unknown, but we note that for $\nu = M/r$, $r > 0$, a fixed value of M, and appropriate approximations of the Γ-function, the $\beta_{j,\gamma}$ can be approximated according to Equation (6) as functions of (M, j, γ, r). Such a parameterized representation enables us to connect the computed values $\beta_{j,\gamma} = \beta_{j,\gamma}(M, r)$ to the observed values β_j.

Given the sequence of pairs $(\beta_{j,\gamma}; \beta_j)$, $j = 1, ..., M$, the task is to identify a value for γ that provides a best fit of the $\beta_{j,\gamma}$ to the observed values β_j. Garnier and Kallel (2002) propose the χ^2-test for approximating γ, which leads to minimizing

$$T_\gamma = \sum_{j=1}^{M} \frac{(\beta_j - \beta_{j,\gamma})^2}{\beta_{j,\gamma}} \tag{7}$$

with respect to γ. Since in our approach we have chosen a parameterized representation for the $\beta_{j,\gamma}$ from Equation (6), we have $T_\gamma = T_\gamma(r)$ and for minimizing $T_\gamma(r)$ we execute a simple search procedure $\min(T_\gamma)$ with an outer-loop running for $r \in [r_{\min}; r_{\max}$ with stepsize δ_r and an inner-loop running for $\gamma \in [\gamma_{\min}\gamma_{\max}$ with stepsize δ_γ. The output is denoted by $\texttt{out} = [T_\gamma(r); \gamma; r]$, with $\texttt{out}[1]$ being the minimum $T_\gamma(r)$ found within the given range (assuming the output is unique, otherwise for the maximum γ for the same $T_\gamma(r)$ and maximum r, respectively).

The procedure $\min(T_\gamma)$ searches for a global minimum within a grid of size $[(r_{\max} - r_{\min})/\delta_r] \times [(\gamma_{\max} - \gamma_{\min})/\delta_\gamma]$. It is important to note that the $\beta_{j,\gamma}(M,r)$ involved are independent of the particular problem. Therefore, for given M, fixed j and $[r_{\min}, r_{\max}, \delta_r; \gamma_{\min}, \gamma_{\max}, \delta_\gamma]$, all values can be pre-calculated and stored in a matrix of size $[(r_{\max} - r_{\min})/\delta_r] \times [(\gamma_{\max} - \gamma_{\min})/\delta_\gamma]$. Typically, the selection of r-values will be such that $r_{\min} < 1$ and $r_{\max} > 1$. To make the calculation (approximation) of the $\beta_{j,\gamma}(M,r)$-values numerically stable, the minimum value of γ will be chosen as $\gamma_{\min} = 0.1$. For example, for $r_{\min} = 0.25$, $r_{\max} = 2.5$, $\delta_r = 0.1$, $\gamma_{\min} = 0.1$, $\gamma_{\max} = 2.5$ and $\delta_\gamma = 0.01$, the size of the grid is only 5400, i.e. ~5.3 kb memory locations are required. (For a comparison to MathWorks' ga-optimizer, see Section 2 in Supplementary Material.)

We denote the value of $r = \texttt{out}[3]$ returned by $\texttt{out} = [T_\gamma(r);\ \gamma; r]$ at the end of procedure $\min(T_\gamma)$ by r_{appr}. For the set of randomly selected conformations $\mathsf{M} := \{S_1, S_2, ..., S_M\}$, the approximation of the number of local minima provided by the Garnier–Kallel method can then be chosen by setting

$$\nu_{\text{appr}} = \frac{M}{r_{\text{appr}}}. \tag{8}$$

Thus, $\gamma_{\text{appr}} = \texttt{out}[2]$ from $\texttt{out} = [T_\gamma(r); \gamma; r]$ that minimizes $T_\gamma(r)$ is not directly involved in calculating ν_{appr}. Minimizing $T_\gamma(r)$ with respect to γ aims at finding the best stochastic model that fits the observed data β_j, and as a 'byproduct' we gain the information about r_{appr}.

For the calculation of $\beta_{j,\gamma}$ according to (6), we first calculate $F_s := \ln\Gamma(z_s)$ for each of the six $\Gamma(z_s)$ involved, and additionally we use $\ln\binom{M}{j} = \sum_{s=M-j+1}^{M} \ln s - \sum_{t=1}^{j} \ln t$. Finally, we set $\beta_{j,\gamma}(M,r) = (M/r)e^Z$, where $Z := \ln\binom{M}{j} + \sum_{s=1}^{3} F_s - \sum_{t=4}^{6} F_t$.

For fixed M, j and $[r_{\min}, r_{\max}, \delta_r; \gamma_{\min}, \gamma_{\max}, \delta_\gamma]$, the values of $\beta_{j,\gamma}(M,r)$ are pre-calculated and require only a small memory space $\texttt{mem-sp}[r_{\min}, r_{\max}, \delta_r; \gamma_{\min}, \gamma_{\max}, \delta_\gamma]$. Theoretically, $\texttt{mem-sp}$ has to be multiplied by M, since $j = 1, ..., M$. However, the χ^2-test usually takes into account only a fraction of the values computed for the stochastic model, which in our case are the $\beta_{j,\gamma}$, see also Section 6 in Garnier and Kallel (2002). This is due to the fact that small values in the denominator in Equation (7) can distort the quality of approximations obtained by the first major terms. There are different rules for adapting the χ^2-test to the individual problem under consideration. Based on preliminary computational experiments, we have decided to include into the actual minimization of $T_\gamma(r)$ only summands with $\beta_{j,\gamma} \geq 1.0$, which is at the lower end of recommended values. Consequently, only a relatively small number of $\beta_{j,\gamma}$ has to be taken into account, which is denoted by $\mathsf{J}(M)$ (for the sequences considered in the present paper, $\mathsf{J}(M)$ is in the range of small double-digit numbers). (For a detailed analysis of β_j and basin sizes, see Section 3 in Supplementary Material.)

2.2 Reeves–Eremeev estimation

Reeves and Eremeev (2004) study the problem of how many trials of steepest descent are required for detecting with high probability all local minima in a given landscape. In other terms, the authors provide a lower bound for the size of M that ensures (with a certain confidence) the detection of all local minima. The Reeves–Eremeev estimation involves the number of local minima. That is why we first try to estimate the number of local minima by the Granier–Kallel method, and then we apply the lower bound provided by Reeves and Eremeev (2004).

The Reeves–Eremeev estimation is derived from a step-by-step analysis of the increase of the number of detected local minima. We assume that already k local minima have been detected. Let W_k denote the expected number of randomly chosen structures with subsequent execution of steepest that need to be processed for an increase to $(k+1)$ detected local minima. Reeves and Eremeev (2004) propose the following geometric distribution for W_k:

$$p(z|k) = \frac{\nu - k}{\nu}\left(\frac{k}{\nu}\right)^{z-1}. \tag{9}$$

The overall number W of trials can then be estimated by

$$W = 1 + \sum_{k=1}^{\nu-1} W_k, \tag{10}$$

which implies for the expected value and the variance

$$E(W) \approx \nu\ln(\nu+g), \tag{11}$$

$$V(W) \approx \frac{(\nu\pi)^2}{6} + 1 - \nu\ln(\nu+g), \tag{12}$$

where $g \approx 0.58$ is the Euler–Mascheroni constant. Both approximations provide the lower bound

$$W > \nu(\ln\nu+g) + c_\rho\sqrt{\frac{(\nu\pi)^2}{6} + 1 - \nu(\ln\nu+g)}. \tag{13}$$

The factor c_ρ is a coefficient associated with the assumption that Equation (13) is valid with confidence $\rho \in [0,1]$ (or in terms of percentages). Based on numerical simulations, Reeves and Eremeev (2004) suggest $c_\rho = 1.83$ for $\rho = 95\%$.

3 RESULTS

We applied the methods described in the preceding section to ten 3′UTRs of human RNAs and to the riboswitch AL935260 of length 79nt reported by Bengert and Dandekar (2004). We analysed one 3′UTR of length 50nt (for a partial landscape defined by a bounded-depth neighbourhood relation) and nine examples of 3′UTRs with a length ranging from 66nt up to 401nt. The upper bound on the length of 3′UTR (and $\ell(3'\text{UTR}) = 50$nt for the bounded-depth neighbourhood) is caused by the chosen evaluation strategy: the approximation of the number of local minima is verified against the values calculated by the `barriers` implementation (Gruber *et al.*, 2008), which requires to invoke the `RNAsubopt` program (Wuchty *et al.*, 1999) with an exponentially increasing number of structures for an increasing distance to the global minimum.

The sequences were chosen rather randomly, with some of them being highly ranked (but not yet validated) in microRNA target predictions by MicroCosm (miRanda) (Griffiths-Jones *et al.*, 2008). However, we tried to design a test set that contains sequences with varying ratios of the size of the subspace versus the number of local minima within this subspace, including one sequence with a very 'rugged' partial landscape close to the minimum free energy conformation.

3.1 Approximating the number of local minima

For a given $3'$UTR S (= linear sequence without base pairs), we generated the `RNAsubopt` output with a certain energy difference δE above the minimum free energy conformation. The offset δE was chosen in such a way that the subsequent `barriers` application produced a number of local minima roughly in the region of 1000. The set of conformations produced by `RNAsubopt` with δE is denoted by $\mathsf{S}_{\delta E}$. The selection of $\mathsf{M} := \{S_1, S_2, ..., S_M\} \subset \mathsf{S}_{\delta E}$ is then executed in the following way:

(0) Initialize $M = M_0 >> 0$. The selection of M_0 is a crucial task. A potential guideline for selecting M_0 is presented in more detail in Section 3.3.

(1) $\mathsf{S}_{\delta E}(S)$ is partitioned into S' and S'' with $|\mathsf{S}'| \approx |\mathsf{S}_{\delta E}(S)|/3$ (and therefore $|\mathsf{S}''| \approx 2|\mathsf{S}_{\delta E}(S)|/3$) and $E(S') > E(S'')$ for $S' \in \mathsf{S}'$ and $S'' \in \mathsf{S}''$.

(2) Out of S', M secondary structures are randomly selected.

The Steps (1) and (2) are specific to our evaluation procedure. In general, during a search-based landscape analysis, $\widehat{M} \approx 3M$ independent runs are executed for a fixed number K of neighbourhood transitions with a subsequent selection of M terminal conformations with higher energy values. The K transitions per run can be performed, for example, within a neighbourhood (Hamming-distance sphere) $d(S_0, \widetilde{S}) \leq k$ of the starting conformation S_0.

Furthermore, in order to support a relatively fast run-time, for a given M only one set M is selected. To counter-balance deviations caused by random selections, the number of local minima is calculated in two ways, which is further explained in Steps (7) and (9).

The next steps relate to the application of the Garnier–Kallel method and are independent of our particular evaluation method:

(3) For each $S_u \in \mathsf{M}$ steepest descent is executed according to (i), (ii) and (iii), where for each intermediate $\widehat{S}$ the neighbourhood $\mathsf{N}_{\widehat{S}}$ is computed by applying Op-I, Op-II and Op-III.

(4) The set LM(M) from Equation (2) is identified.

(5) The values β_j from Equation (3) are calculated for $j = 1, ..., \mathsf{J}(M)$.

(6) If $R := \beta_1/\beta_2 > 3$, then $M := M + M_0$ and go to Step (2). Alternatively, if the initial M_0 is small and $\beta_1/\beta_2 >> 3$, the increase of M can be chosen as $M := 2M$.

Step (6) indicates (in the general case of unknown ν) whether M is too small for exploring the (partial) landscape under consideration. Moreover, the step contributes to the numeric stability by avoiding small γ in `out`[2] of min(T_γ). Finally, ν_{appr} is calculated:

(7) We set $M' := \sum_{j=1}^{\mathsf{J}(M)} j\beta_j$. M' is the number of initial conformations S_u actually involved in the χ^2-test; cf. Equation (7).

(8) min(T_γ) is executed on M and Equation (8) is applied to M and M' (both with r_{appr} = `out`[3]).

(9) The final output ν_{appr} is determined in the following way:
Case A: If $M' > M/2$, then $\nu_{\mathrm{appr}} = M'/r_{\mathrm{appr}}$;
Case B: If $M' \leq M/2$, then $\nu_{\mathrm{appr}} = 2\big(M/r_{\mathrm{appr}}\big)/\big(M/M' + 1\big)$.

The two cases A and B are the result of a large number of test-runs for different $\ell(3'\mathrm{UTR})$ and different settings of M. If $M' > M/2$, there are fewer local minima that attract many elements of M (and are represented by $\beta_{j,\gamma}(M, r) < 1.0$), i.e. the random selection M is a 'good choice' within $p\mathsf{L}$. If $M' \leq M/2$, only a relatively small part of M actually participates in min(T_γ). Since we want to avoid repeated runs on different random selections M, we establish a relation between M' and M by setting $\nu_{\mathrm{appr}} = \big(w'\nu_{M'} + w\nu_M\big)/\big(w' + w\big)$. A straightforward setting of weights is $w' = M/M'$ and $w = 1$, which leads to the expression in Case B.

The results are summarized in Table 1. We emphasize that the data were generated by arbitrary runs for each of the sequences. The setting corresponds to the case that the heuristic is used as a subroutine in search-based folding simulations. In such simulations, it is not desirable to optimize approximations of ν by a statistical analysis over a larger number of runs for the same value of M.

On the other hand, we noticed that repeated runs for the same M with $R < 3$ can produce improved approximations, especially if a particular run results in $R < 2$, which then corresponds to relatively small values of T_γ and `out`[2] $>> 0.1$. However, none of the runs presented in Table 1 resulted in $R < 2$ and all $M' > M/2$ (Case A) with all `out`[2] close or equal to 0.1. In this sense, the data from Table 1 seem to represent an ordinary scenario and do not display specific positive features. (Large values of M are analysed in Section 5 Supplementary Material.)

If $R = \beta_1/\beta_2 \leq 3$ for the smaller M, the second run was executed in order to demonstrate the observation that a further increase of M not necessarily improves the quality of the approximation. The observation complies with the data reported by Garnier and Kallel (2002). The analysis of this effect, which could be attributed, for example, to increased β_j with $\beta_{j,\gamma} \geq 1.0$ and $j > 2$, will be a subject of further research. The same applies to the selection of M from the upper energy level, where a combination of two runs for the same M, one for a 'horizontal' selection and the second for a 'vertical' selection, could potentially improve the approximation.

The value of Δ from Table 1 is defined by $\Delta := (|\nu_{\mathrm{appr}} - \nu|/\nu) \times 100\%$. When taking the first run with $R \leq 3$ for each of the sequences from Table 1, the average value of Δ is equal to 38.5%. The average value of Δ for M-runs with lowest value of T_γ is 36.2%. As expected, the ratio $|\mathsf{S}_{\delta E}|/\nu$ has an impact on the quality of approximations: for the two sequences with $\ell(3'\mathrm{UTR}) = 98$ and the sequence of length 79, the ratio is particularly large, and the corresponding M-runs with $R \leq 3$ all produce $\Delta < 10\%$. The sequence NM_024482 with the smallest value of $|\mathsf{S}_{\delta E}|/\nu$ produced the worst Δ value.

Table 1. Approximation of the number ν of local minima

Ref. No. mRNA	$3'$UTR length	$\lvert\mathsf{S}_{\delta E}\rvert$	δE	ν	M	lm(M)	M'	T_γ	r	$R \le 3$	ν_{appr} $= \nu_{M'}$	Δ %
NM_000363	66	4888	6.50	553	750	182	451	16.1	0.72	+	626	13.2
					1500	320	1022	22.1	0.91	+	1123	103.1
NM_172109	66	5589	6.50	491	750	185	412	31.9	0.63	−	654	33.2
					1500	270	775	61.8	0.98	+	791	61.1
NM_018217	78	5514	5.50	455	750	179	428	25.6	0.68	−	629	38.2
					1000	251	830	25.8	1.21	+	686	50.8
NM_000915	98	24 744	9.50	906	1500	326	840	39.5	0.85	+	988	9.1
					3000	439	1596	58.9	1.37	+	1165	28.6
NM_002584	98	37 521	8.50	998	1500	306	814	49.6	0.85	+	958	4.0
					3000	433	1549	65.5	1.37	+	1131	13.3
NM_024482	100	10 214	5.00	1301	1500	485	1170	23.7	0.50	+	2340	79.9
					3000	687	2140	76.7	0.74	+	2892	122.3
NM_170726	400	20 428	1.85	1100	1500	396	1052	30.6	0.60	−	1753	59.4
					3000	532	1821	55.2	1.12	+	1626	47.8
NM_001043229	400	22 482	2.75	1185	1500	451	1133	35.2	0.52	+	2179	83.9
					3000	647	2089	42.9	0.88	+	2374	100.3
NM_031420	401	15 596	2.75	1017	1500	349	915	54.1	0.70	+	1307	28.5
					3000	515	1742	91.8	1.12	+	1555	52.9
AL935260	79	19 139	8.50	782	750	186	408	30.3	0.60	−	680	13.0
					1500	285	801	42.8	0.96	+	834	6.6
Average Δ for first M-runs with $R \le 3$												38.5
Average Δ for M-runs with lowest value of T_γ												36.2

3.2 Approximating sets of local minima

The value of ν_{appr} is used for the approximation of W from Equation (13). The preceding procedure on M already identified a subset of local minima LM(M), and lm(M) is the size of the subset of local minima, cf. Equations (2) and (4). Therefore, we select only $W_{\text{sub}} := W - M$ pairwise different conformations S_v from the corresponding set $\mathsf{S}_{\delta E}$, where $S_v \neq S_u$ for each $S_u \in \mathsf{M}$ is required. The resulting set of W_{sub} conformations is denoted by $\widehat{\mathsf{W}}$, and as before, after executing steepest descent, we set LM($\widehat{\mathsf{W}}$) $:= \{m \mid$ size of SD$(m) > 0\}$ and lm($\widehat{\mathsf{W}}$) $:= |$LM($\widehat{\mathsf{W}}$)$|$. Although $\widehat{\mathsf{W}} \cap \mathsf{M} = \emptyset$, LM($\mathsf{M}$) and LM($\widehat{\mathsf{W}}$) are not necessarily disjoint. Hence, we set $\text{lm}_\cap := |\text{LM}(\widehat{\mathsf{W}}) \cap \text{LM}(\mathsf{M})|$, and the total size lm_{fin} of $\text{LM}_{\text{fin}} = \text{LM}(\widehat{\mathsf{W}}) \cup \text{LM}(\mathsf{M})$ is then given by $\text{lm}_{\text{fin}} = \text{lm}(\widehat{\mathsf{W}}) + \text{lm}(\mathsf{M}) - \text{lm}_\cap$.

The corresponding results are shown in Table 2. The value of the approximation rate A_{fin} is defined by $A_{\text{fin}} := (\nu_{\text{appr}}/\nu) \times 100\%$. The sequences NM_000363 and NM_172109 are left out due to the large value of W_{sub}. The ratio $W_{\text{sub}}/|\mathsf{S}_{\delta E}|$ is quite large for four sequences (between 0.85 and 0.998) and ≤ 0.34 for three sequences (between 0.21 and 0.34). As mentioned at the beginning of the current section, the test set was designed to contain sequences with a ‘rugged’ energy landscape within the selected subspace. For example, for NM_024482 we have for both ratios the maximum values $|\mathsf{S}_{\delta E}|/\nu = 7.85$ and $W_{\text{sub}}/|\mathsf{S}_{\delta E}| = 0.998$. However, the three sequences with $0.21 \le W_{\text{sub}}/|\mathsf{S}_{\delta E}| \le 0.34$ (as well as the sequence considered in Section 3.3 with a much smaller ratio) produce only slightly worse results with an average of 89%.

The maximum deviation below the targeted 95% approximation rate is for NM_002584 with 7.8%, and the average deviation below 95% is 4.2%, with one sequence above 95%.

3.3 Approximations within a partial landscape $p\mathsf{L}$

For the $3'$UTR of NM_002410 with $\ell(3'\text{UTR}) = 50$, we executed both heuristics for a bounded-depth neighbourhood induced by the secondary structure S = ...(((..((......((..........))......)))))... with seven base pairs and $E(S) = -0.7$kcal/mol. The minimum energy is $E_{\min}(3'\text{UTR}) = -17.6$ kcal/mol. The bounded-depth neighbourhood is defined by

$$\widehat{\mathsf{N}}^k_S := \{S' \mid (d(S,S') \le k) \& (E(S') \le 10 \frac{\text{kcal}}{\text{mol}})\} \cap \mathsf{S}_{40}. \qquad (14)$$

We selected $k = 5$, and the setting $\delta E = 40$ is the default value for `RNAsubopt`. The size of $\widehat{\mathsf{N}}^5_S$ is 296 331, and the number of local minima within this set is 1000. The total number of local minima within $\{S' \mid S' \in \mathsf{S}_{40} \& E(S') \le 10.00\text{kcal/mol}\}$ is 1247.

The maximum energy of a local minimum that belongs to $\widehat{\mathsf{N}}^5_S$ is $+7.10$kcal/mol, whereas the minimum energy is -12.9kcal/mol. The operations OP-I, OP-II and OP-III were executed only within $\widehat{\mathsf{N}}^5_S$ and the number M' of conformations relates only to minima within this Hamming-sphere.

The results for $M = 1500$ and $M = 3000$ are displayed in Table 3. In both cases, we have $R \le 3$. For $M = 1500$, the value of ν is overestimated by 10.8%. For the subsequent run with $W_{\text{sub}} = 9494$ initial conformations, the approximation rate is equal to $A_{\text{fin}} = 96.2\%$.

Adding the results for NM_002410 to the results from Tables 1 and 2, we obtain an average value of Δ for $R \le 3$ of 36%, and with respect to minimum values of T_γ the average value of 33.9%. The average approximation rate is then slightly $> 92\%$, which deviates from the target by 3%.

Table 2. Approximation of the set of local minima

Ref No	ℓ	ν	M	$\mathrm{lm}(\mathsf{M})$	ν_{appr}	W	W_{sub}	$\mathrm{lm}(\widehat{\mathsf{W}})$	$\mathrm{lm}_{\cap}$	$\mathrm{lm}_{\mathrm{fin}}$	A_{fin}
NM_018217	78	455	1500	251	685	6473	4973	392	230	413	90.8
NM_000915	98	906	1500	326	988	9700	8200	681	197	810	89.4
NM_002584	98	998	1500	306	958	9375	7875	739	175	870	87.2
NM_024482	100	1301	1500	461	2340	11 696	10 196	1196	461	1196	91.9
NM_170726	400	1100	3000	532	1626	16 776	13 776	987	454	1065	96.8
NM_001043229	400	1185	1500	451	2179	23 121	21 621	1091	439	1103	93.1
NM_031420	401	1017	1500	349	1307	13 198	11 698	912	318	943	92.7
AL935260	79	782	1500	285	834	8046	6546	629	208	706	90.3
							Average approximation rate A_{fin}				91.5

Table 3. Approximations within $p\mathsf{L}$ for NM_002410 with $\ell(3'\mathrm{UTR})=50$ and $\nu=1000$

M	M'	r	T_{γ}	ν_{appr}	Δ	$\mathrm{lm}(\mathsf{M})$	W	W_{sub}	$\mathrm{lm}(\widehat{\mathsf{W}})$	$\mathrm{lm}_{\cap}$	$\mathrm{lm}_{\mathrm{fin}}$	A_{fin}
1500	720	0.65	76.4	1108	10.8	442	10 994	9494	901	381	962	96.2
3000	1069	0.88	137.3	1215	21.5	525	12 180	9180	891	486	930	93.0

As already mentioned, selecting the initial $M=M_0$ is an important step in the general case of partial landscapes defined by $d(S_0,\widetilde{S})\leq k$. A potential way of specifying M_0 could be based on the observation made by Lorenz and Clote (2011) that the number of local optima relates to the square root of the size of the entire conformation space. Adopting this relation to partial landscapes could then be used for identifying the order of magnitude of an initial M_0. However, this would require adequate estimations of the number of canonical structures (Clote *et al.*, 2009) in Hamming-spheres defined by $d(S_0,\widetilde{S})\leq k$, and further research is needed for establishing secure rules of selecting initial M_0.

For example, if p with $2k\leq 2p\leq n-2k$ is the number of base pairs within S_0 of length n, then a coarse estimation of the number of conformations that obey $d(S_0,\widetilde{S})\leq k$ with no isolated base pairs (canonical structures) can be easily established based upon the sum of products of binomial coefficients. Assuming $k<<n$ and using $\binom{a}{b}\sim(ea/b)^b$, the estimation has the order of magnitude of $\left(e\max\{2(n-2p);p\}/k\right)^k$. For the square root, we set $(a_{m-1},\ldots,a_0)_{10}:=\lfloor\left(e\max\{2(n-2p);p\}/k\right)^{k/2}\rfloor$ with $0<a_{m-1}\leq 9$, and the suggested initial M_0 would then be defined by $M_0=10^{m-1}$. For the sequence from Table 3, we have $2(n-2p)=72>p=7$, and $\lfloor(e\times 72/5)^{2.5}\rfloor=9601$ then defines $M_0=1000$.

3.4 Complexity analysis

We analyse the complexity of the proposed heuristic for a partial landscape $p\mathsf{L}$ defined by a fixed secondary structure S and the bounded-depth neighbourhood $\mathsf{N}_S^k:=\{S'|d(S,S')\leq k\}$.

The most time consuming part of finding ν_{appr} consists of the execution of steepest descent according to (i), (ii) and (iii), in particular, step (ii). The second step is based on Op-I, Op-II and Op-III, where the time complexity of each of the operations can be upper bounded by $O(n^2)$, which includes the update of the energy value assigned to each element of the neighbourhood. The update can be carried out in a similar way as described by Lorenz and Clote (2011) for the nearest neighbour model by local changes to the energy value. The operations OP-I, OP-II and OP-III perform a complete search for feasible conformations, and therefore the information about the affected loop (as well as neighbouring loops) in the loop representation of free energy (Lorenz and Clote, 2011) can be maintained and locally updated.

If $D=D(\mathsf{M})$ denotes the maximum length of a steepest descent pathway from an $S\in\mathsf{M}$ to the corresponding element of LM(M), the total number of steps can be estimated by $O(MDn^2)$. For the sequences and partial landscapes we analysed, the value of D is relatively small, typically in the region of low double digit numbers. In the second part, namely the approximation of sets of local minima, the same operations are executed over $\widetilde{M}\leq O(\nu\ln\nu)$ randomly selected initial conformations, cf. Equation (13). If D is the maximum pathway length (depth of attraction basins) for both parts of the procedure, we have $O(Dn^2\nu\ln\nu)$ basic steps for the second part. Making sure that the initial conformations are pairwise different requires $O(M^2)$ and $O((\nu\ln\nu)^2)$ steps, respectively. Based on our computational experiments, assuming $M=O(\nu\ln\nu)$ and $\nu\ln\nu\leq O(Dn^2)$ seems to be justified. Therefore, taking both parts together leads to a time complexity of $O(Dn^2\nu\ln\nu)$.

In the present paper, the sets M are randomly drawn from an `RNAsubopt` output. `RNAsubopt` produces an exponentially increasing output with increasing δE. Therefore, in the general case, the sets M are sampled from random trajectories of length K within $p\mathsf{L}$. The aim is to reach out to different areas of the partial landscape. Consequently, the transitions are not governed by constants k_{uv} as mentioned in Section 1, but changes to conformations are executed randomly, where processing each transition requires $O(n)$ steps. The worst case bound $O(n)$ results from checking the validity of the randomly generated neighbour: the minimum loop size is 3nt, and the condition of at least two consecutive bonds must be maintained. Overall, for given K and M, $3KM$ random transitions are executed. Out of the $3M$ final conformations, M conformations with the highest energy values are then selected. If we assume $K\approx O(Dn)$, the total complexity $O(3KMn)$ does not exceed $O(Dn^2\nu\ln\nu)$.

4 DISCUSSION

For the 11 sequences we analysed in the present paper, the Garnier–Kallel method not only produced estimations within the same order of magnitude of the true number of local minima (calculated by `RNAsubopt` and `barriers`), but also with an average deviation of 36% (in most cases overestimation); the estimations of the number of local minima seem to be a useful input to the subsequent approximation of the entire set of local minima. On the nine sequences with a sufficiently large size of the partial landscape, the estimations and parameter settings provided by Reeves and Eremeev (2004) have led to an average approximation rate of 92% for a target rate of 95%, which we see as a confirmation of the Reeves– Eremeev method and a justification for the entire approach. The results were achieved for sequences with varying ratios of the size of the partial landscape versus the number of local minima within the subspaces.

An open question is the secure selection of the initial sample size M_0. A potential way for solving this problem could be based on good approximations of the number of secondary structures that constitute the Hamming-sphere around a particular structure under consideration. The increase of the sample size in case of $\beta_1/\beta_2 >> 3$ can be organized in different ways. For example, when starting with a relatively small M_0, a sample size of $2^t M_0$ in step t provides a rapid increase that could lead within a few steps to at least $\beta_1/\beta_2 \leq 3$ (additionally, one could then search backwards between $2^{t-1}M_0$ and $2^t M_0$ additive steps for better values of T_γ where the current M-run still complies with $\beta_1/\beta_2 \leq 3$). On the other hand, for a given sequence length n and a fixed radius it might be possible to devise standard rules for initial settings of M_0, which could result in a very small number of M-runs.

Another line of further improvements could be the combination of the data from two runs for the same value of M, with one run for a 'horizontal' selection of samples (high free energy within the partial landscape) and the second run for a random selection over all conformations obtained by M random walks of length K within the partial landscape. Here, one would need to establish rules of how to determine the approximation from two different sets of values for M' and `out`[3] taken together with the current value of M.

ACKNOWLEDGEMENTS

The authors are grateful to the anonymous referees for their careful reading of the manuscript and valuable suggestions that resulted in an improved and more detailed presentation of our results.

Funding: Department for Employment and Learning through its 'Strengthening the all-Island Research Base' initiative (in part).

Conflict of Interest: none declared.

REFERENCES

Bellaousov,S. and Mathews,D.H. (2010) ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA*, **16**, 1870–1880.

Bengert,P. and Dandekar,T. (2004) Riboswitch finder–a tool for identification of riboswitch RNAs. *Nucleic Acids Res.*, **32**, W154–W159.

Chen,S.J. (2008) RNA folding: Conformational statistics, folding kinetics, and ion electrostatics. *Annu. Rev. Biophys.*, **37**, 197–214.

Clote,P. *et al.* (2009) Asymptotics of canonical and saturated RNA secondary structures. *J. Bioinform. Comput. Biol.*, **7**, 869–893.

Danilova,L.V. *et al.* (2006) RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. *J. Bioinform. Comput. Biol.*, **4**, 589–596.

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

Flamm,C. and Hofacker,I.L. (2008) Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatshefte f. Chemie*, **139**, 447–457.

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. *Z. Phys. Chem.*, **216**, 155–173.

Garnier,J. and Kallel,L. (2002) Efficiency of local search with multiple local optima. *SIAM J. Discr. Math.*, **15**, 122–141.

Geis,M. *et al.* (2008) Folding kinetics of large RNAs. *J. Mol. Biol.*, **379**, 160–173.

Griffiths-Jones,S. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.

Gruber,A.R. *et al.* (2008) The Vienna RNA Web suite. *Nucleic Acids Res.*, **36**, W70–W74.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, **125**, 167–188.

Hofacker,I.L. *et al.* (2010) BarMap: RNA folding on dynamic energy landscapes. *RNA*, **16**, 1308–1316.

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

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

Mann,M. and Klemm,K. (2011) Efficient exploration of discrete energy landscapes. *Phys. Rev. E*, **83**, 011113.

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

Morgan,S.R. and Higgs,P.G. (1998) Barrier heights between ground states in a model of RNA secondary structure. *J. Phys. A: Math. Gen.*, **31**, 3153–3170.

Pörschke,D. (1974) Model calculations on the kinetics of oligonucleotide double helix coil transitions. Evidence for a fast chain sliding reaction. *Biophys. Chem.*, **2**, 83–96.

Ragan,C. *et al.* (2011) Quantitative prediction of miRNA–mRNA interaction based on equilibrium concentrations. *PLoS Comput. Biol.*, **7**, e1001090.

Reeves,C.R. and Eremeev,A.V. (2004) Statistical analysis of local search landscapes. *J. Oper. Res. Soc.*, **55** 687–693.

Schuster,P. (2006) Prediction of RNA secondary structures: from theory to models and real molecules. *Rep. Prog. Phys.*, **69**, 1419–1477.

Tang,X. *et al.* (2005) Using motion planning to study RNA folding kinetics. *J. Comput. Biol.*, **12**, 862–881.

Tang,X. *et al.* (2008) Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.*, **381**, 1055–1067.

Wang,F. and Landau,D.P. (2001a) Determining the density of states for classical statistical models: a random walk algorithm to produce a flat histogram. *Phys. Rev. E*, **64**, 056101(1)–056101(16).

Wang,F. and Landau,D.P. (2001b) Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, **86**, 2050–2053.

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

Wuchty,S. *et al.* (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–164.

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.

Xayaphoummine,A. *et al.* (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.*, **33**, 605–610.

Xia,T.B. *et al.* (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.

Zhao,P. *et al.* (2010) Predicting secondary structural folding kinetics for nucleic acids. *Biophys. J.*, **98**, 1617–1625.