Structural bioinformatics

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

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 \( L \) 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 \( \text{RNAsubopt} \) 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 \( O(n^2 \text{Dlog}_2 n) \), where \( n \) is the sequence length, \( v \) is the number of local minima in the partial landscape \( L \) under consideration and \( D \) is the maximum number of steepest descent steps in attraction basins associated with \( L \).

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 \( dp_u(t)/dt = \sum_{v \neq u}(k_{uv}p_u(t) - k_{vu}p_v(t)) \) are feasible only for short sequences due to the large number \( n \) of folding conformations \( C_n \) 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 \( k_{uv} \) are the transition rates from state \( u \) to state \( v \). A standard setting for values \( k_{uv} \) is given by \( k_{uv} = \exp\left(-\left(G_v - G_u\right)/RT\right) \) for \( G_v > G_u \) and \( k_{uv} = 1 \) for \( G_v < G_u \), where \( G_v \) is the free energy of conformation \( v \). 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, 2009). Xayaphoummine et al. (2003) pay particular attention to structures with pseudoknots and how to overcome the problem of kinetic traps consisting of rapidly changing 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 (2006) recently devised a fast algorithm that runs in \( O(n^2) \) time and utilizes base paring 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 \( O(n^2) \) time and requires \( O(n^3) \) space. The underlying energy model is the Turner nearest neighbour model Xu 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 Lee and Clote (2011). The method employs the Wang–Landau algorithm Wang and Landau (2001) 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 (2003) and is also part of the detailed
The authors associate macrostates with attraction basins of local
minima in energy landscapes, see also Wolfinger et al. (2011) 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 conformational space 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.

Our procedure is designed for the analysis of partial landscapes \(p_L\) that cover only a small subset of the entire conformation space. Such \(p_L\) 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 \(v\) of locally optimal structures within \(p_L\) is approximated by using a stochastic procedure devised by Furnari and Kalin (2003). Secondly, the value of \(v_{\text{app}}\) is taken as an input into an approximation formula established by Reeves and Twigg (2003) for the number \(W(p_L,v_{\text{app}})\) of repeated runs of steepest descent within \(p_L\) out of randomly selected conformations, where \(W(p_L,v_{\text{app}})\) 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 (Wächter et al. 1998) and Vienna package implementations (Bryan et al. 2008) taken from the Vienna package (Bryan et al. 2008).

2 APPROACH

We consider combinatorial landscapes \(L=\{C,N,F\}^n\) defined by the set of conformations \(C\), the neighbourhood relation \(N\) and the fitness function \(f: C \rightarrow R\). 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 \( S \to S' \in N_S \):

(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 \( f: \mathbb{C} \to \mathbb{R} \) is defined as the free energy \( E: \mathbb{C} \to \mathbb{R} \), 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 \#(bp-S'\cdot bp-S) is the number of base pairs common to \( S \) and \( S' \), then \( d(S,S') = \#(bp-S)+\#(bp-S')-2\#(bp-S'\cdot bp-S') \).

The method proposed in the present paper aims at small subsets \( S \), that this holds for energy landscapes induced by RNA secondary structures. For each of the \( S_p \), the steepest descent procedure is deterministic and terminates at the local minimum that satisfies the condition \( \nabla E(p) = 0 \) for the energy \( E(p) \). The steepest descent procedure terminates at the local minimum that this holds for energy landscapes induced by RNA secondary structures. For each of the \( S_p \), the steepest descent procedure is deterministic and terminates at the local minimum that satisfies the condition \( \nabla E(p) = 0 \) for the energy \( E(p) \).

### 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; Garnier and Kallel 2005). Let \( \nu_S = |A_S|/|C_S| \) denote the normalized size of attraction basin \( A_S \). Garnier and Kallel (2002) assume a parameterized random distribution of the normalized sizes \( \nu = \nu_S \) of attraction basins with the density function \( p_\nu \) defined by

\[
p_\nu(z) = \frac{\gamma^z}{\Gamma(\gamma)} \frac{\gamma^{-1}}{z^{-\gamma}}
\]

where \( \gamma > 0 \) and \( \Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx \) is the Euler function. The density function \( p_\nu \), i.e. the parameter \( \gamma \), is approximated by a sampling method over attraction basins.

Let \( M = \{S_1, S_2, \ldots, S_M\} \subseteq \mathbb{C} \) denote a set of randomly selected secondary structures. For each of the \( S_p \), the steepest descent procedure is executed, which leads to \( S_p \to m_p \), if \( S_p \) belongs to the attraction basin \( A_{m_p} \), spawned by \( m_p \). By \( SD(m) = \{A_S \mid \nu_S \geq \nu \} \) we denote the set of conformations from \( M \) where steepest descent terminates at \( m_p, \nu = 1, \ldots, \nu \). The set \( M \) together with steepest descent identifies a set of local minima denoted by

\[
LM(M) := \{m \mid \text{size of } SD(m) > 0\}
\]

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

\[
\beta_j = \{B_j\}, j \geq 0.
\]

By this definition, \( \beta_j \) is the number of local minima ‘visited’ by exactly \( j \) out of \( M = |M| \) conformations after steepest descent. By definition of \( \text{Im}(M) \) we then have

\[
\text{Im}(M) = \sum_{j=1}^{M} \beta_j.
\]

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

\[
M = \sum_{j=0}^{M} \beta_j.
\]

Supplementary Figure S1 illustrates the particular values for the case of \( v \to 4 \) and \( M = 7 \).

Let \( \beta_{j,v} \) denote the expected value of \( \beta_j \) under the assumption that the normalized sizes \( \alpha \) of attraction basins \( A \) are distributed according to \( p_\nu \) defined in (3). Garnier and Kallel (2002) established the relation

\[
\beta_{j,v} = \sum_{m=1}^{M} \left( \frac{\gamma^v}{\Gamma(\gamma)} \frac{\gamma^{-1}}{v^{-\gamma}} \right) \frac{\Gamma(v-1)+M-j}{\Gamma(v)}
\]

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

Given the sequence of pairs \((\beta_j, \beta_{j,v}), j = 1, \ldots, M\), the task is to identify a value for \( \gamma \) that provides a best fit of the \( \beta_{j,v} \) to the...
Thus, MathWorks \[ T \] the minimum and maximum \( r \) the observed data \( \sim \) at the end of procedure \( \min(M) \) \( T \) \( r \) \( \gamma \) for the stochastic model, which in our case are the \( Fs \) for \( \delta \) with \( \gamma \) observed values \( \beta \) \( \gamma \) \( T \) \( r \) \( \beta \), with \( \gamma \) being the minimum \( T \) \( r \) found within the given range (assuming the output is unique, otherwise for the maximum \( r \) for the same \( T \) \( r \) and \( \gamma \) respectively).

The procedure \( \min(M) \) searches for a global minimum within a grid of size \( |r_{\text{max}}-r_{\text{min}}|/\delta \times |\gamma_{\text{max}}-\gamma_{\text{min}}|/\gamma \). It is important to note that the \( \beta_j(M,r) \) are involved in an important of the particular problem. Therefore, for given \( M \), fixed \( j \) and \( |r_{\text{max}}-r_{\text{min}}|/\delta \times |\gamma_{\text{max}}-\gamma_{\text{min}}|/\gamma \), all values can be pre-calculated and stored in a matrix of size \( |r_{\text{max}}-r_{\text{min}}|/\delta \times |\gamma_{\text{max}}-\gamma_{\text{min}}|/\gamma \). Typically, the selection of \( r \)-values will be such that \( r_{\text{max}} < 1 \) and \( r_{\text{max}} > 1 \). To make the calculation (approximation) of the \( \beta_j(M,r) \) values numerically stable, the minimum value of \( r \) will be chosen as \( \gamma_{\text{min}} = 0.1 \). For example, for \( r_{\text{max}} = 0.25 \), \( r_{\text{max}} = 2.5 \), \( \delta_j = 0.1 \), \( \gamma_{\text{max}} = 0.1 \), \( \gamma_{\text{max}} = 2.5 \) and \( \delta_j = 0.01 \), the size of the grid is only 5400, i.e. \( 5.3 \) kb memory locations are required. (For a comparison to Matrix\textsuperscript{WKS}’s ga-optimizer, see Section 2 in Supplementary Material.)

We denote the value of \( r = \sqrt{3} \) returned by \( \min(M) \) at the end of procedure \( \min(M) \) by \( r \). For the set of randomly selected conformations \( M = \{ S_1, S_2, \ldots, S_6 \} \), the approximation of the number of local minima provided by the Garnier–Kallel method can then be chosen by setting \( \rho_{\text{app}} = \frac{M}{r_{\text{app}}} \). (8)

Thus, \( \rho_{\text{app}} = \sqrt{3} \) returned by \( \min(M) \) at the end of procedure \( \min(M) \) that \( \min(M) \) is not directly involved in calculating \( \rho_{\text{app}} \). Minimizing \( T \) \( r \) with respect to \( y \) aims at finding the best stochastic model that fits the observed data \( \beta_j \) and a ‘byproduct’ we gain information about \( r \).

For the calculation of \( \beta_j \) according to \( (3) \), we first calculate \( \beta_j(M,r) = \frac{\sum_{j=1}^{M-r} \sum_{i=1}^{r} (\gamma_{\text{max}}-\gamma_{\text{min}}) \int_{\gamma_{\text{min}}}^{\gamma_{\text{max}}} f_{\gamma} d\gamma} \) for each of the six \( \Gamma_j(M) \) involved, and additionally we use \( \int_{\gamma_{\text{min}}}^{\gamma_{\text{max}}} f_{\gamma} d\gamma = \frac{\sqrt{3}}{2} \pi \rho_{\text{app}} \). Finally, we set \( \beta_j(M,r) = (M/r)^2 \), where \( \rho_{\text{app}} = \sqrt{3} \) \( (M/r)^2 \).

For fixed \( M, j \) and \( \beta_j(M,r) = \frac{\sum_{j=1}^{M-r} \sum_{i=1}^{r} (\gamma_{\text{max}}-\gamma_{\text{min}}) \int_{\gamma_{\text{min}}}^{\gamma_{\text{max}}} f_{\gamma} d\gamma} \), the values of \( \beta_j(M,r) \) are pre-calculated and require only a small memory space \( \text{memory}(\gamma_{\text{max}}-\gamma_{\text{min}}, r_{\text{max}}-r_{\text{min}}, \delta_j) \). Theoretically, \( \text{memory}(\gamma_{\text{max}}-\gamma_{\text{min}}, r_{\text{max}}-r_{\text{min}}, \delta_j) \) has to be multiplied by \( M \), since \( j = 1, \ldots, M \). However, the \( \chi^2 \)-test usually takes into account only a fraction of the values computed for the stochastic model, which in our case are the \( \beta_j \), see also Section 6 in Garnier and Kallel (2002). This is due to the fact that small values in the denominator in Equation \( (3) \) can distort the quality of approximations obtained by the first major terms. There are different rules for adapting the \( \chi^2 \)-test to the individual problem under consideration. Considered on the basis of preliminary computational experiments, we have decided to include into the actual minimization of \( T \) \( r \) only summands with \( \beta_j \geq 1.0 \), which is at the lower end of recommended values. Consequently, only a relatively small number of \( \beta_j \) has to be taken into account, which is denoted by \( J(M) \) (for the sequences considered in the present paper, \( J(M) \) is in the range of small double-digit numbers). (For a detailed analysis of \( \beta_j \) 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 \( W_k \): \( p(k) = \frac{v-k}{v} \) \((k+1)\). (9)

The overall number \( W \) of trials can then be estimated by \( W = 1 + \sum_{k=1}^{n-1} W_k \), (10) which implies for the expected value and the variance \( E(W) = v\ln(v+g) \), (11) \( V(W) = \left( v \right)^2 / 6 \), (12) where \( g \approx 0.58 \) is the Euler–Mascheroni constant. Both approximations provide the lower bound \( W > v\ln(v+g) + v\sqrt{\left( v \right)^2 / 6 - v\ln(v+g)} \). (13)

The factor \( v \) 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 \( v = 1.83 \) 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 AL95260 of length 79nt reported by Reupert 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 60nt up to 401nt. The upper bound on the length of 3'UTR (and 3'UTR = 50nt 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 (and 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
e above the minimum free energy conformation. The offset e 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 e =
denoted by $S_{\text{EG}}$. The selection of $M := \{S_1, S_2, \ldots, S_{\ell}\} \subset S_{\text{EG}}$ is then executed in the following way:

(0) Initialize $M = M_0 = \emptyset$. The selection of $M_0$ is a crucial task. A
potential guideline for selecting $M_0$ is presented in more detail in
Section 3.6.

(1) $S_{\text{EG}}(S)$ is partitioned into $S'$ and $S''$ with $|S'| = \frac{1}{3}|S_{\text{EG}}(S)|$ (and therefore $|S''| = \frac{2}{3}|S_{\text{EG}}(S)|$) and $E(S') < E(S'')$ for $S' \in S'$ and $S'' \in S''$.

(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, $M \approx 3M$
neighbourhood 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) $d(S_0, S) \leq k$ of the starting conformation $S_0$.

Furthermore, in order to support a relatively fast run-time, for a
given $M$ only one set $S$ 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 $S_j \in M$ deepest descent is executed according to (i),
(ii) and (iii), where for each intermediate $S \delta$ the neighbourhood $N_S$ is computed by applying Op-I, Op-II and Op-III.

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

(5) The values $\beta_j$ from Equation 1 are calculated for $j = 1, \ldots, |M|$.

(6) If $R := \beta_1/\beta_3 > 3$, then $M := M + M_0$ and go to Step (2).
Alternatively, if the initial $M_0$ is small and $\beta_1/\beta_3 > 3$, the
increase of $M$ can be chosen as $M := 2M$. Step (6) indicates (in the general case of unknown $v$) whether $M$
is too small for exploring the (partial) landscape under consideration.

Moreover, the step contributes to the numeric stability by avoiding
small $y$ in out[2] of min($T_j$). Finally, $v_{\text{app}}$ is calculated:

(7) We set $M' := \sum_{j=1}^{|M|} \beta_j$. $M'$ is the number of initial
conformations $S_j$ actually involved in the $\chi^2$-test; cf. Equation 4.

(8) min($T_j$) is executed on $M$ and Equation 3 is applied to $M$
and $M'$ (both with $v_{\text{app}} = \text{out}[3]$).

(9) The final output $v_{\text{app}}$ is determined in the following way:
Case A: If $M' > M/2$, then $v_{\text{app}} = v_{\text{app}}'$; Case B: If $M' \leq M/2$, then $v_{\text{app}} = 2(M/2 + 1)$.

The two cases A and B are the result of a large number of test-
runs for different (3′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 $\tilde{\beta}_j (M, r < 1.0)$, i.e. the random selection $M$ is
a ‘good choice’ within $\beta$. If $M' \leq M/2$, only a relatively small part of
$M$ actually participates in min($T_j$). Since we want to avoid repeated
runs on different random selections $M$, we establish a relation between
$M$ and by setting $v_{\text{app}} := (w^\delta v_{\text{app}} + w w_{\text{app}})/(w + w')$. 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 $v$ 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 $T_J$ and $\text{out}[2] > 0.1$. However, none of the runs
presented in Table 1 resulted in $R < 2$ and all $M' > M/2$ (Case A)
with all $\text{out}[2] < 0.1$. 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 $R = \beta_1/\beta_3 \leq 3$ for the smaller $M$, the second run was executed in
order to demonstrate the observation that a further increase of
$M$ if $R < 3$ may produce improved approximations, especially if a
particular run results in $R < 2$, which then corresponds to relatively
small values of $T_J$ and $\text{out}[2] > 0.1$. However, none of the runs
presented in Table 1 resulted in $R < 2$ and all $M' > M/2$ (Case A)
with all $\text{out}[2] < 0.1$. 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.)

The analysis of this effect, which could be attributed, for example,
explained in Section 3.6. For the cases where the seed is involved in the
selection and for the two ‘vertical’ selection, could potentially improve the approximation.

The value of $\Delta$ from Table 1 is defined by $\Delta := \delta(\text{app} - v)/v \times 100$. When taking the first run with $R \leq 3$ for each of the sequences
from Table 1, the average value of $\Delta$ is equal to 38.5%. The average
value of $\Delta$ for $M$-runs with lowest value of $T_J$ is 36.2%. As expected,
the ratio $|S_{\text{EG}}|/v$ has an impact on the quality of approximations: for
the two sequences with $S(3′ \text{UTR}) = 98$ 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 $|S_{\text{EG}}|/v$ produced the worst $\Delta$ value.
3.2 Approximating sets of local minima

The value of $v_{\text{appr}}$ is used for the approximation of $W$ from Equation (3). 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 (3) and (4). Therefore, we select only $W_{\text{sub}}=W-M$ pairwise different conformations $S_i$ from the corresponding set $S_{\text{fin}}$, where $S_i \neq S_j$ for each $S_i \in M$ is required. The resulting set of $W_{\text{sub}}$ conformations is denoted by $\hat{W}$, and as before, after executing steepest descent, we set $LM(\hat{W})=\{m|\text{size of } S_0(m)\geq0\}$ and $LM(\hat{W})=\{LM(\hat{W})\}$. Although $W=\hat{W}+M$, $LM(\hat{W})$ and $LM(M)$ are not necessarily disjoint. Hence, we set $LM(\hat{W})=\{LM(\hat{W})\}\cup LM(M)$, and the total size $|LM(\hat{W})|=LM(W)\cup LM(M)$ is then given by $LM(\hat{W})=LM(\hat{W})+LM(M)$. The corresponding results are shown in Table 2. The value of the approximation rate $A_{\text{fin}}$ is defined by $A_{\text{fin}}=(v_{\text{appr}}/v)\times100$. The sequences NM_000363 and NM_172109 are left out due to the large value of $W_{\text{sub}}$. The ratio $W_{\text{sub}}/|S_{\text{fin}}|$ is quite large for four sequences (between 0.85 and 0.998) and $\leq 0.34$ 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 $|S_{\text{fin}}|/v=7.85$ and $W_{\text{sub}}/|S_{\text{fin}}|=0.998$. However, the three sequences with $0.21 < W_{\text{sub}}/|S_{\text{fin}}| \leq 0.34$ (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 $pL$

For the 3’UTR of NM_002410 with $\ell(3’UTR)=50$, we executed both heuristics for a bounded-depth neighbourhood induced by the secondary structure $S=\ldots(\ldots(\ldots(\ldots)\ldots)\ldots)\ldots)$ with seven base pairs and $E(S)=0.77\text{ kcal/mol}$. The minimum energy is $E_{\text{min}}(3’UTR)=-17.6\text{ kcal/mol}$. The bounded-depth neighbourhood is defined by $N_{K}=\{S’|d(S,S’)\leq k\& E(S’)\leq 0.77\text{ kcal/mol}\}$. The minimum energy is $-12.9\text{ kcal/mol}$. The operations OP-I, OP-II and OP-III were executed only within $N_{K}$ and the number of $M’$ 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 $R \leq 3$. For $M=1500$, the value of $v$ 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 $\Delta$ for $R \leq 3$ of 36%, and with respect to minimum values of $T_r$, the average value of 33.9%. The average approximation rate is then slightly $\geq 92\%$, which deviates from the target by 3%.
A potential way of specifying the complexity of the proposed heuristic for a partial landscape 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 $d(S_0, S) \leq k$, and further research is needed for establishing secure rules of selecting initial $M_0$.

As already mentioned, selecting the initial $M = M_0$ is the most time consuming part of finding $\nu_{\text{opt}}$. A potential way of specifying $M_0$ could be based on the invariance of the number of local optima with respect 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 $d(S_0, S) \leq k$, 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 $d(S_0, S) \leq k$ within an iso energetic cone is $2k^n$. This number can be easily established based upon the order of magnitude of $(\text{max}(2n - 2p; p)/k)^{2k}$. The maximum pathway length (depth of attraction basins) for both parts of the procedure, we have $O(Dn^2\nu^2)$ basic steps for the second part. Making sure that the initial conformations are pairwise different requires $O(M^2)$ and $O(\nu^2)$ steps, respectively. Based on our computational experiments, assuming $M = O(\nu^2n^2)$ and $\nu = O(Dn^2)$ seems to be justified. Therefore, taking both parts together leads to a total time complexity of $O(Dn^2\nu^2)^2$.

In the present paper, the sets $M$ are randomly drawn from an RNAopt subopt output. RNAsubopt produces an exponentially increasing output with increasing $M$. Therefore, in the general case, the sets $M$ are sampled from random trajectories of length $K$ within $\mathcal{L}$. The aim is to reach out to different areas of the partial landscape. Consequently, the transitions are not governed by constants $k_{\text{opt}}$ 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 3 and the condition of at least two consecutive bonds must be maintained. Overall, for given $K$ and $M$, $3M$ random transitions are executed.
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 as the true number of local minima (calculated by RNAfold:opt: 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 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 $2^2 M_0$ 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 $2^{-1} M_0$ and $2^0 M_0$ additive steps for better values of $T_f$ where the current $M$-run still complies with $\beta_1/\beta_2 < 3$). 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 $\text{out}[3]$ 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


