Monatsh Chem 139, 447–457 (2008)
DOI 10.1007/s00706-008-0895-3
Printed in The Netherlands

Monatshefte für Chemie
Chemical Monthly

Review

Beyond energy minimization: approaches to the kinetic folding of *RNA*

Christoph Flamm, Ivo L. Hofacker

Institute of Theoretical Chemistry, University of Vienna, Wien, Austria

Received 21 December 2007; Accepted 12 January 2008; Published online 10 March 2008

Abstract The term *RNA* folding is often used synonymously with the prediction of equilibrium structures. Yet many *RNA*s function thanks to their ability to undergo structural changes. In this contribution we present a systematic overview of existing approaches to the prediction of *RNA* folding kinetics, and in particular discuss the strengths and limitations of each method.

Keywords RNA folding kinetics; Co-transcriptional folding; Folding pathway; Metastable structures.

Introduction

Most functional *RNA* molecules depend on their structure to perform their respective function. *RNA* secondary structures have established themselves as the most convenient level of description, mostly because of the availability of efficient algorithms to predict the structure of minimum free energy [1–3]. In fact, on the level of secondary structures, any equilibrium property of an *RNA* molecule can be computed either directly *via* the partition function over all structures [4] or by sampling structures from the *Boltzmann* ensemble [5]. Nevertheless, the equilibrium view of *RNA* folding can be misleading: the time needed to reach equilibrium can become very long and, since *RNA*s in the cell have high turnover, may easily exceed the lifetime of the *RNA* molecule.

Correspondence: Ivo L. Hofacker, Institute of Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria. E-mail: ivo@tbi.univie.ac.at

The tendency of *RNA* molecules to form long-lived folding intermediates is a direct consequence of the high stability of *RNA* helices. It is therefore not surprising that Nature makes use of this feature to produce *RNA*s that can switch between conformational states with different function.

It is still an open question to what extent the functional structures of natural *RNA*s are determined by folding kinetics rather than by equilibrium thermodynamics. Nevertheless, there is a growing number of well-studied examples where *RNA* function is clearly mediated by structural changes, and thus the static view of *RNA* structure is insufficient.

The renewed interest in *RNA* as a versatile biomolecule has also inspired diverse experimental approaches to measure folding kinetics in detail, ranging from classical temperature jump experiments [6] to time-resolved NMR spectroscopy [7, 8] and single molecular methods [9]. In this contribution we aim to provide an overview of the different computational strategies for modeling *RNA* folding kinetics and discuss strengths and limitations of the respective approaches.

Evidence for kinetic folding in natural *RNA*s

In a cellular context the nascent *RNA* molecule starts folding before the transcription process is completed and the folded structure may therefore depend on the speed of elongation, site-specific pausing of the *RNA* polymerase, and interactions of the nascent

RNA molecule with proteins or small-molecule metabolites [10, 11].

In naturally occurring (m)*RNA*s two broad classes of structural elements, capable of toggling between alternative conformations, can be observed which differ mainly in their switching mechanism. The first class comprises structure elements whose functions are triggered by a local or global external signal such as temperature, *pH*, or binding of small metabolites. These types of elements therefore function as a sensor, and are often located in the 5′-UTRs of m*RNA*s, in particular of fundamental metabolic genes [12]. Typical examples include *RNA* thermometers [13] or riboswitches [14].

Riboswitches, for example, are common regulatory elements in bacteria. In general, they can be divided into an aptamer part that binds a small metabolite, and an ''expression platform'' whose structure modulates gene expression. Conformational changes in the aptamer part are relayed to the expression platform where translation can be modulated by changing the accessibility of the ribosome entry site. Similarly, transcription can be effected by the formation (or destruction) of a terminator hairpin. Riboswitches control gene expression directly without any intermediates, and thus allow an extremely rapid response to environmental changes. They can be found in all kingdoms and are presumably one of the oldest regulatory mechanisms [15].

The second class are ''self-induced'' *RNA* switches [16], that are initially present in a long-lived metastable state, that eventually re-folds spontaneously without an outside trigger. Self-induced switches allow to limit biologically functional properties of *RNA* structures to certain time windows. The most prominent examples are the attenuation regulation of bacterial amino acid bio-synthetic operons [17] or the hok/soc *RNA* antitoxin system [18] for the maintenance of R1 plasmid in *E. coli*.

A deeper understanding of how this additional layer of *RNA* regulation integrates into cell-wide regulatory circuits requires the study of the folding kinetics of *RNA*. Furthermore, several computational studies suggest that the folding pathways of naturally occurring *RNA*s are encoded within their primary sequences [19–21]. In other words, evolution has optimized the co-transcriptional folding pathway of these sequences by employing strategies like transient structural elements guiding the folding or the suppression of the formation of alternative helices that would compete with the functional structure.

Modeling the folding process

Most approaches to kinetic *RNA* folding aim to directly model the physical folding process. All these approaches are based on a straightforward description of folding in terms of a stochastic process. In general any such model is defined by three key ingredients: (i) The state space, comprising the set of structures or conformations a given *RNA* sequence may assume, (ii) a *move-set* defining the elementary transitions that can occur between such conformations, and (iii) transition rates for each of these allowed transitions.

The folding process can now be described as a continuous time *Markov* process, governed by a master equation for the state probabilities $P_x(t)$ of observing state x at time t.

$$\frac{dP_x(t)}{dt} = \sum_{y \neq x} [P_y(t)k_{xy} - P_x(t)k_{yx}] \qquad (1)$$

k_{xy} is the transition rate from state y to state x, with $k_{xy} > 0$ for all transitions allowed by the move set. Conservation of probability, *i.e.*, the fact that $\sum_x P_x(t) = 1$ for all t, implies that the diagonal elements $k_{xx} = -\sum_{x \neq y} k_{yx}$.

Since we aim to describe a physical process that converges towards thermodynamic equilibrium in the limit of long time, the move-set and rates have to meet additional ergodicity requirements. Firstly, the *Markov* chain should be irreducible, *i.e.*, it should be possible to reach every conformation y from any starting conformation x using a finite number of moves. Secondly, the transition rates should fulfill the *detailed balance* condition.

$$\pi_y k_{xy} = \pi_x k_{yx} \qquad (2)$$

Here π_x is the stationary distribution of the process, which in our case should be the *Boltzmann* distribution $\pi_x = \exp(-\Delta G(x)/RT)/Z$, with $\Delta G(x)$ the free energy of state x and Z the partition function $Z = \sum_x \exp(-\Delta G(x)/RT)$. If the above conditions are fulfilled, *Markov* chain theory guarantees that the stationary state π_x is unique and $\lim_{t\to\infty} P_x(t) = \pi_x$ for any initial condition $P_x(0)$.

As an example, let us consider the simple model folding kinetics in the space of secondary structures,

as used *e.g.*, in the kinfold program [22]. Given an *RNA* molecule with sequence s, the state space is given by the set of secondary structures X that are *compatible* with s, *i.e.*, structures that can be formed by sequence s while considering only *Watson-Crick* (**GC**, **AU**) and wobble (**GU**) pairs and avoiding pseudoknots.

The simplest move-set considers only addition and removal of single base pairs. In other words a transition between conformation x and y is allowed only if the two structures differ by a single base pair. It is easy to see that this move-set is ergodic, since any structure x can be converted into the ''open chain'' structure containing no base pairs, by successively removing all base pairs. Note also, that move-sets introduce a notion of distance between conformations as the minimum number of moves needed to move from x to y. In the case of single base pair addition and removal this is known as the ''base-pair distance''.

For pseudoknot-free secondary structures there is a well established energy model that assigns a free energy to every structure based on the *Turner* energy rules [23–25]. Based on these energies the *Metropolis* rule [26] is the simplest and most widely used rule to obtain transition rates that satisfy detailed balance:

$$k_{xy} = \Gamma \cdot \max(1,\, e^{(\Delta G(x) - \Delta G(y))/RT}) \qquad (3)$$

The constant Γ sets the time-scale of the process and should be chosen by comparison with experiment.

Other possibilities for the choice of conformation space, move-set, and rate models will be described below. Note that the master Eq. (1) can be written in vector form with $\mathbf{K} = (k_{xy})$ the transition rate matrix

$$\frac{d}{dt}P(t) = \mathbf{K}P(t) \qquad (4)$$

This equation gives rise to the formal solution

$$P(t) = e^{t\cdot\mathbf{K}}P(0), \qquad (5)$$

where $P(0)$ is the initial distribution vector.

Simulation techniques

If the dimension of $\mathbf{K}$, *i.e.*, the total number of conformations, is small enough to allow diagonalization, then the complete folding behavior can be computed with ease for arbitrarily long times. In most cases, however, the size of the conformation space makes this approach infeasible, and the only practical recourse is stochastic simulation of Eq. (4) using *Monte-Carlo* techniques.

It may be worth noting that special care has to be taken in *Monte-Carlo* simulation to conserve detailed balance, since the number of neighbors for different conformations is not constant. In a basic rejection based *Monte-Carlo* implementation the transition rate for the move $x \to y$ is the product of two parts, the *a priori* probability to attempt a certain move $\mathcal{A}(x \to y)$ times the acceptance probability $\mathcal{P}(x \to y)$. Normally, one would simply choose a neighbor at random from the neighborhood of x, $\mathcal{N}(x)$, and thus $\mathcal{A}$ is inversely proportional to the number of neighbors, $\mathcal{A}(x \to y) = 1/|\mathcal{N}(x)|$. Since this is not constant, using the *Metropolis* rule for the acceptance probability does not guarantee detailed balance.

One way to circumvent this problem is to use a rejectionless *Monte-Carlo* approach. In this *Monte-Carlo* variant, all possible moves from the start conformation x are evaluated and the new conformation y is chosen from this list with probability $P(x \to y) = k(x \to y)/\sum_z k(x \to z)$. The clock is then advanced by a value Δt chosen from a *Poisson* distribution with mean $1/\sum_z k(x \to z)$. This algorithm is known in physics as the ''n-fold way'' or ''*Bortz-Kalos-Liebowitz* (*BKL*)'' method [27], while in chemistry it is usually referred to as the *Gillespie* algorithm [28].

Rather than perform *Monte-Carlo* at constant temperature one may use simulated annealing techniques in order to accelerate folding [29]. Here the simulation starts at a high temperature which is gradually lowered to physiological temperature. It should be noted that the folding pathway obtained along such a cooling schedule need not coincide with the folding pathway at constant temperature. Finally, some authors use optimization techniques such as genetic algorithms rather than *Monte-Carlo* simulation. Note that the cross-over operation employed in genetic algorithms has no equivalent in the physical folding process. Nevertheless, the technique has been used to predict likely folding pathways [30, 31].

As we will see below most existing approaches are based on the model described here and mainly differ in the set of allowed conformations (*e.g.*, with or without pseudoknots), the move-set, as well as the energy rules and resulting rate model.

Move sets and coarse grained configuration spaces

As mentioned above, the most elementary move-set consists only of base-pair insertion and deletion and corresponds to the smallest possible steps in conformation space. While this allows the most detailed description of folding pathways, it necessarily leads to extremely long simulation runs. Many approaches therefore choose to allow larger structural changes by using the formation or destruction of an entire helix as the basic step. This allows to explore the conformation space in a much smaller number of steps and consequently enables the simulation of larger *RNA*s up to the size of large ribosomal *RNA*s. An intermediate between base pair moves and helix moves that allows changing several uncorrelated base pairs in a single time step is introduced in Ref. [32].

Helix based methods usually start by compiling a list of all allowed helices, where typically only saturated helices are used, that cannot be extended on either side. In order to keep the list small a minimum helix length of typically 3 or 4 is required.

In the simplest version, a helix is either present in its entirety or completely absent. Thus, an allowed conformation is uniquely determined by the set of helices that are present in the structure, and may be represented by a binary vector that specifies which helices are present and which are absent. Clearly, only non-overlapping helices can be present simultaneously in any given structure. Folding simulations in this scenario are no more complicated than in the case of single base pair moves. The space of allowed conformations is, however, severely restricted. The problem is that two long helices are mutually exclusive, if they overlap even by a single base. Any conformation where one helix is shortened in order to accommodate another is thus excluded.

It is therefore common to include conformations with partial helices [33–36]. An insertion move may now insert a partial helix or even shorten existing helices in order to make room. This is followed by a local optimization where the extents of conflicting helices are optimized to obtain a structure that is a local energy minimum. While this has been used quite successfully in practice, from the theoretical point of view there are some caveats.

Given a set of helices, the concrete structure that is produced by inserting these helices will in general depend on the order in which they were inserted.

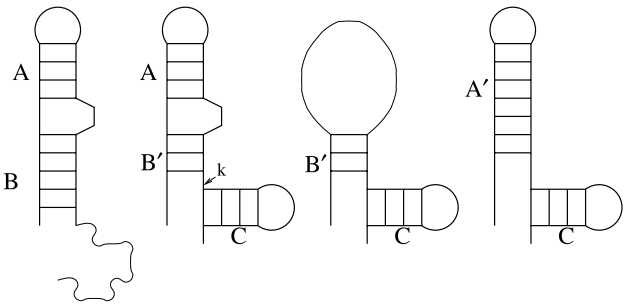

Fig. 1 Conflict resolution for helix kinetics and a scenario for possible violation of reversibility. In the first step helix C is inserted. Since helices B and C partially overlap, the end point k is optimized leading to a shortened helix B′. In the next two steps helix A is destroyed and re-formed. However, helices A and B partially overlap, and since helix B has already been shortened after insertion of C, optimization of the cut-point between A and B results in an elongated A with B being eliminated. Thus, removal and re-insertion of helix A did not restore the original conformation, the *Markov* chain is not reversible

This makes it difficult to even judge how many different conformations can be formed by the algorithm. Moreover, it is in general not possible to ensure that the resulting *Markov* process is reversible. The exact way this conflict resolution is done varies between implementations, but in all cases it is a local optimization procedure that affects only helices adjacent to the newly inserted or destroyed helix. As a consequence, it cannot be guaranteed that a series of moves, followed by their corresponding reverse moves, recovers the original structure. An example for such a scenario is shown in Fig. 1. In practice, one assumes that such cases are rare and should therefore introduce no noticeable artifacts.

Kinetic rate models

The detailed balance requirement (2) leaves much freedom in the choice of kinetic rates, as it fixes only the ratio between forward and backward rates. The usual Ansatz is to define a transition state and set the rate using the *Arrhenius* equation

$$k_{xy} = \Gamma \exp(-(\Delta G^{\dagger}_{xy} - \Delta G(x))/RT) \qquad (6)$$

where $\Delta G^{\dagger}_{xy}$ is the free energy of the transition state. The *Metropolis* rule (3) thus identifies the transition state with the energetically higher of the two states x and y. In general the exact rate model matters less when moves are small such as in the case of single base pair moves. Such simulations therefore often

simply use the *Metropolis* rule. Simulations using the symmetric "*Kawasaki*" rule $k = \Gamma \exp(-\Delta G/(2RT))$, where ΔG is the free energy difference between the two states, showed qualitatively the same behavior. Nevertheless, *Schmitz* and *Steger* [29] suggest to split ΔG into two parts, the change in free energy from stacking interactions and the change in loop entropies. The change in stacking energy is then used for the barrier when opening a base pair, while the change in loop energy is used when closing a pair. Thus the transition state corresponds to a conformation where loop penalties for bringing the bases together has been paid, but the energetically favorable stacking interactions have not yet been established.

For helix based moves the quality of the rate model is much more important. *Tacker et al.* [37] propose a rate model for helix moves similar to that described for single base pairs above: the (mostly entropic) change in loop energies is used as activation energy when forming a helix, while the change in stacking free energies is used when opening a helix. The same approach was adopted *e.g.*, in Refs. [38, 39]. Similarly, *Zhang* and *Chen* [40] use the total change in entropy when forming helices and the total change in enthalpy when destroying them.

In *Isambert's* Kinefold program [35] the insertion of a helix is initiated by inserting a nucleus of usually length 3, choosing the energetically best nucleation point. The rate of helix formation is then given by an *Arrhenius* law using the free energy barrier for nucleation. This barrier is given by the entropic penalty incurred by inserting the nucleus. The nucleation site may overlap an existing helix, in which case that helix has to be shrunk to make room for the nucleus. In this case the free energy necessary to shrink the helix is added to the barrier as well.

In all cases the prefactor Γ can be chosen to fit experimentally measured re-folding times.

Energy rules

Free energies for *RNA* secondary structures are normally modeled using the so-called nearest neighbor model, and parameters for this model have been derived in the *Turner* group based on a large number of oligo-nucleotide melting experiments [23–25]. This model, however, does not include energies for pseudoknotted structures. Pseudoknots are often neglected because it is algorithmically difficult to include them in the dynamic programming algorithms used for predicting optimal structures. Kinetic folding algorithms do not share this problem, and consequently many kinetic folding programs explicitly allow pseudoknots.

The free energy of a pseudoknotted structure is primarily composed of two contributions, a stabilizing one arising from the base stacking in the helices and a destabilizing one stemming from the loss in entropy of the looped regions. While the former contribution is accurately described by the nearest-neighbor energy model, the challenge lies in obtaining a realistic estimate of the loop entropies. Modeling these pseudoknot energies is more difficult than for regular secondary structures in several respects: (i) there are almost no thermodynamic measurements for pseudoknotted structures. (ii) While a Pseudobase [41] lists a moderate number of pseudoknotted structures, most of these are small H-type knots. (iii) While any pseudoknot free secondary structure is at least sterically feasible, most hypothetical pseudoknotted structures are not [42]. With the exception of some models for H-type pseudoknots [43, 44], existing approaches are therefore based on statistical mechanics models of simple polymer chains, rather than thermodynamic measurements, *e.g.*, Refs. [35, 45–48].

Chen and *Dill* [45, 49], apparently inspired by successful work on lattice protein models, developed one of the first statistical mechanics polymer models of *RNA*. In their model, a secondary structure is represented as a self-avoiding walk on a 2D square lattice, where each nucleotide occupies one lattice point. Hence, excluded volume effects in the loop regions and between substructures are taken into account in coarse grained manner. The reason for this rather drastic coarse-graining of *RNA* molecules is that the conformational partition function of the lattice *RNA* representation can be calculated quite efficiently up to chain lengths of about 200 nts. Starting from the conformational partition function any desired property of the *RNA* chain such as melting curves can be calculated. In an attempt to improve the lattice model predictions of thermodynamic properties of *RNA* conformational change, *Zhang* and *Chen* [50] extended the two-dimensional lattice *RNA* model to a three-dimensional version on a cubic-square lattice.

In order to tighten the correspondence between polymer model and the "real" *RNA* structure, *Cao*

and *Chen* [51] developed a lattice based "atomic" *RNA* conformation model. Following *Olson*'s virtual bond model [52], the *RNA* backbone is modeled as chain using two atoms per residue. For the lattice, *Cao* and *Chen* chose the diamond lattice since angles and torsion angles correspond well to typical values of the virtual bond model. For helical regions an off-lattice 3D structure is produced initially, again using the virtual bond model and setting all torsion angles to a standard helical value. This tertiary structure model is then fitted onto the diamond lattice. Finally, loop regions are modeled as self-avoiding walks on the diamond lattice such that the end-points of the loops are constrained to the corresponding lattice points of the embedded helices.

For loop types in normal secondary structures, such as hairpins, bulge and internal loops the results agree well with loop entropies from the *Turner* model, at least for the longer loops. Similarly, simple H-type pseudoknots can be modeled well using this approach. Complex pseudoknots are less amenable since possible loop configurations have to be sampled separately for each possible location and orientation of all helices. Consequently, the approach has so far been used only for relatively small examples, including models of *RNA–RNA* interaction complexes [53] and H-type pseudoknots [54].

Isambert and *Siggia* [35] attack the problem of assigning a conformational entropy to a knotted structure by decomposing the structure into so-called local nets (single stranded closed circuits, that enclose up to two internal helices) and global constrains between the local nets. For the local nets, single stranded regions are modeled as springs and helices as stiff rods. In this approximation, entropy contribution of the local nets can be calculated analytically [55]. The constraints between the local nets are modeled as a cross-linked "*Gaussian* gel" obtained by contracting the local nets to single vertices connected by *Gaussian* springs, see Fig. 2. The entropy of this cross-linked gel is then calculated numerically by algebraic iteration. The approach does not explicitly treat excluded volume except through the persistence length of the *RNA* chain. However, among the existing approaches this is the only one that allows for arbitrarily complex structures. For consistency their Kinefold program uses the above approach even for loops that are not involved in pseudoknots and uses *Turner* energies only for stacked pairs.

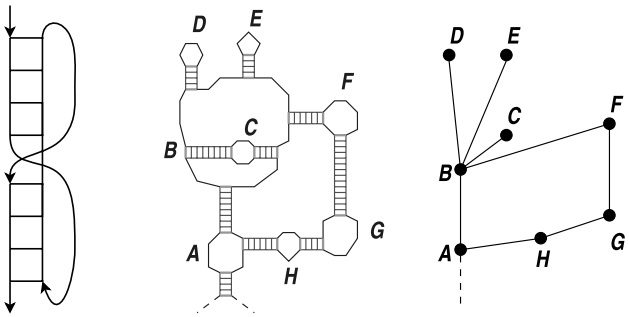

Fig. 2 *Left* the H-type pseudoknot is the simplest and by far most common type of pseudoknot. *Middle* More complicated pseudoknots, such as this one are neglected in most approaches. *Right* the corresponding cross-linked *Gaussian* gel used in *Isambert*'s Kinefold to estimate the global part of the conformation entropy

Heuristic approaches to kinetic folding

All of the above approaches are computationally expensive at least for somewhat longer *RNA*s. In particular, since for stochastic simulation a fairly large number of trajectories has to be sampled. It is therefore tempting to devise simpler heuristics to obtain a single or a small number of plausible folding pathways.

The simplest folding heuristic is based on a stepwise addition of single helices to a structure in a greedy manner. The algorithm starts with an empty structure and a list of potential helices. In each step the energetically most favorable helix, *i.e.*, the one leading to the largest decrease in free energy is inserted, and subsequently all helices that conflict with the selected helix (since they would form base triples or pseudoknots) are deleted from the list. The algorithm stops when the list is empty or all remaining helices would increase the free energy of the structure. In its simplest form this algorithm was introduced already in 1984 [56] as an attempt to obtain an algorithm that is faster than the prediction of minimum free energy structures *via* dynamic programming. Various variants of this "greedy" heuristics have been implemented, that differ mostly in the way which compatible helix is chosen for addition. *Li* and *Wu* [57], for example, pick a helix at random, provided that the free energy of the resulting structure is lowered. *Abrahams et al.* [58] extend the original method by allowing pseudoknotted configurations as well as folding during transcription.

Geis et al. [59] recently implemented in the program *KinWalker* a heuristic approach which com-

putes a co-transcriptional folding pathway for long *RNA* sequences (1500 nts). The algorithm constructs a series of metastable structures by a stepwise combination of thermodynamically optimal structural fragments, which can be calculated efficiently for all substructures by the standard dynamic programming approach for *RNA* folding. In each extension step, the energy barrier for potential structural rearrangements is estimated and only re-arrangements with an activation barrier below some threshold are accepted. Estimation of energy barriers is done by explicitly constructing re-folding paths, where only shortest paths, with a minimal number of base pairs openings and closings, are allowed.

Energy landscapes

Not only are stochastic simulations time consuming, it can also be tedious to extract from them the local minima that act as meta-stable states and kinetic traps in the folding process. It is even more difficult to identify transition states for re-folding between two such meta-stable states, for the simple reason that two different trajectories, even with the same start and end points, will in general share only few exactly identical intermediate structures.

It is therefore of interest to compute local minima, and energy barriers between them, directly *via* an analysis of the energy landscape. In Refs. [22, 60] we developed a *flooding algorithm* that decomposes the landscape into basins surrounding local minima connected by saddle points. Briefly, the program barriers works by processing the conformations of a landscape in energy sorted order, starting at the global minimum. For each conformation x the set of neighboring conformations $\mathcal{N}(x)$ (*e.g.*, in the case of *RNA*, those that can be reached by opening or closing a single base pair) is constructed. If none of the neighbors has been observed before, x is a local minimum and thus the first member of a new basin. If the neighborhood $\mathcal{N}(x)$ contains previously observed conformations from at least two basins m_1, m_2, then x is a saddle point connecting m_1 and m_2. Finally we assign x to the lowest basin in its neighborhood. The saddle points and local minima thus identified form a hierarchy that can be visualized conveniently in the form of a so-called barrier tree, see Fig. 3 for an example.

The flooding algorithm is not specific to *RNA* landscapes and has in fact been used to study the landscapes resulting from various optimization problems [61–63]. However, in the case of *RNA* the analysis is aided by the availability of an efficient algorithm that produces the low energy part of the conformation space [3]. This makes the landscape approach effective for *RNA* the size of, say a t*RNA*, where the complete landscape may contain over 10^{17} structures, while the relevant part, containing low energy conformations with $E<0$, may consist of only a few million structures. While analyzing 10^{17} structures is clearly infeasible, the barriers program can handle 10 million structures with ease. Of course even the number of low energy conformations grows exponentially with sequence length, and as a consequence, the barriers approach is rarely successful for sequences of more than 80–100 nts.

In general the approach is best suited to analyze refolding processes, since the re-folding time between two local minima can be estimated directly from the energy barrier separating them. The barrier tree is less helpful in predicting which of several meta-stable states will be preferentially populated when the folding process starts from an un-folded state.

Moreover, local minima can be used as a starting point for a coarse graining of the conformation space. *Wolfinger et al.* [64] use a partitioning of the landscape into macrostates, where a macrostate is defined as the set of all starting conformations for which a gradient walk ends in the same local minimum m. While constructing the tree, the barriers program identifies these ''gradient basins'' and calculates effective transition rates between any two macrostates α and β as

$$\begin{aligned} k(\alpha \to \beta) &= \sum_{x \in \alpha} \sum_{y \in \beta} k(x \to y) \mathrm{Prob}[x|\alpha] \\ &= \sum_{x \in \alpha} \sum_{y \in \beta} k(x \to y) e^{-E(x)/RT} / Z_\alpha, \end{aligned} \tag{7}$$

where we have assumed local equilibrium within each macrostate and Z_α is the partition function over all conformations in macrostate α and the *Metropolis* rule Eq. (3) is used to model the microstate transition probabilities $k(x \to y)$.

Tang et al. [65, 66] adopt a computational technique that is used for motion planning in robotics, known as probabilistic roadmaps, to build an approximated representation of the *RNA* folding landscape. A probabilistic roadmap is a graph where the vertex

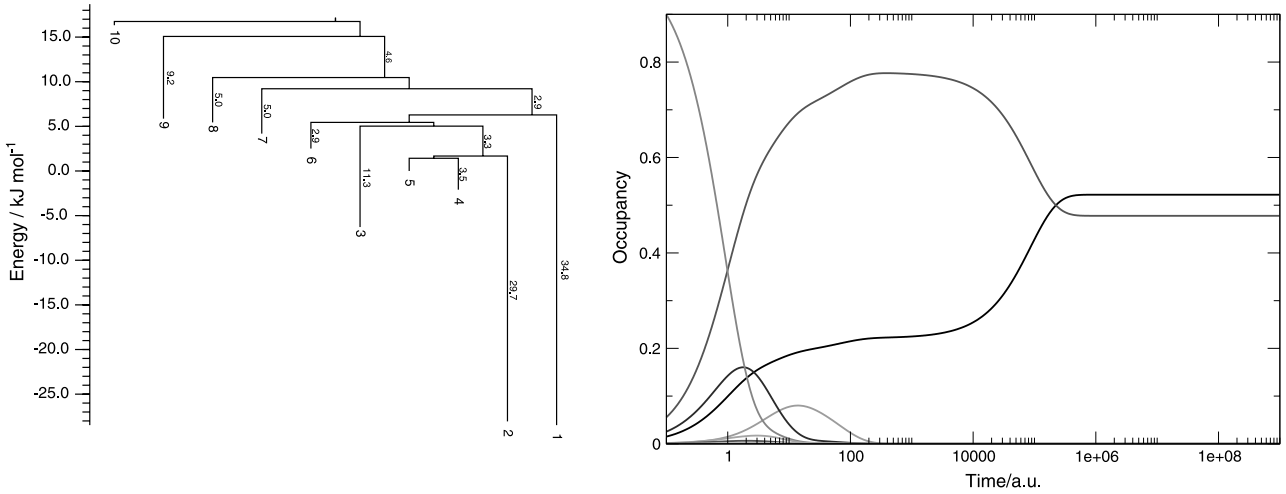

Fig. 3 Barrier tree (*left*) and folding kinetics (*right*) for the artificial sequence **UCCACGGCUGUUAGUGGAUAACGGC**. The right panel shows the occupancy of macro-states as a function of time with the open chain as initial state. The two lowest lying local minima 1 and 2 have almost equal energy and thus equilibrium occupancy. Local minimum 2 however is kinetically preferred achieving almost 80% occupancy around $t = 1000$

set represents valid sampled conformations of the folding landscape and edges are introduced into the graph if a feasible transition exists between the two conformations. A structure based distance criterion is used to avoid the construction of all N^2 (re-folding) paths between the N nodes of the graph. The probabilistic roadmap is used as basis to calculate the time evolution of the population of different conformations providing information on folding rates, transition states, and the equilibrium distribution. In contrast to the analysis of folding landscapes based on exact enumeration using the barriers program, the motion roadmap approach has been applied to *RNA*s up to a size of 200 nts.

Folding on variable landscapes

In many cases one is interested how *RNA* structure changes in response to changes in external parameters like ionic conditions. This type of folding on a variable landscape occurs in several special applications, such as when modelling ''pulling experiments'' [67] where an external force is applied to the *RNA*, or when modeling the transport of an *RNA* molecule through a pore [68]. More importantly, all naturally produced *RNA*s undergo folding during transcription: since transcription is slow compared to local folding processes, the partially synthesized *RNA* will start folding while the molecule is still being synthesized.

While folding of full length *RNA*s from the unfolded state usually results in a variety of primary folding products, co-transcriptional folding can channel folding trajectories such that almost all molecules fold into the same (possibly meta-stable) structure, see Fig. 4 for an example. Thus, for *RNA*s with several long-lived states it is essential to consider co-transcriptional folding in order to predict which of these will be preferentially populated initially.

For all methods that simulate trajectories it is relatively straightforward to include co-transcriptional folding [69]. In the simplest case, the total simulation is simply divided into slices of length τ, corresponding to the mean time for extending the *RNA* by one nucleotide. At the end of these time slices the *RNA* is extended by adding an unpaired nucleotide at the $3'$ end. Instead of using fixed time slices, *RNA* extension itself can also be treated as a stochastic process. Most of the tools mentioned above allow to perform co-transcriptional folding in this way. For the energy landscape approach discussed above, it is possible to analyse the folding landscape for all partially synthesized *RNA*s separately, and then construct a mapping that establishes the correspondence between the local minima in different size landscapes [70].

In Nature, however the situation is still more complicated, since transcription speed is far from constant. Many genes contain specific pause sites [71]

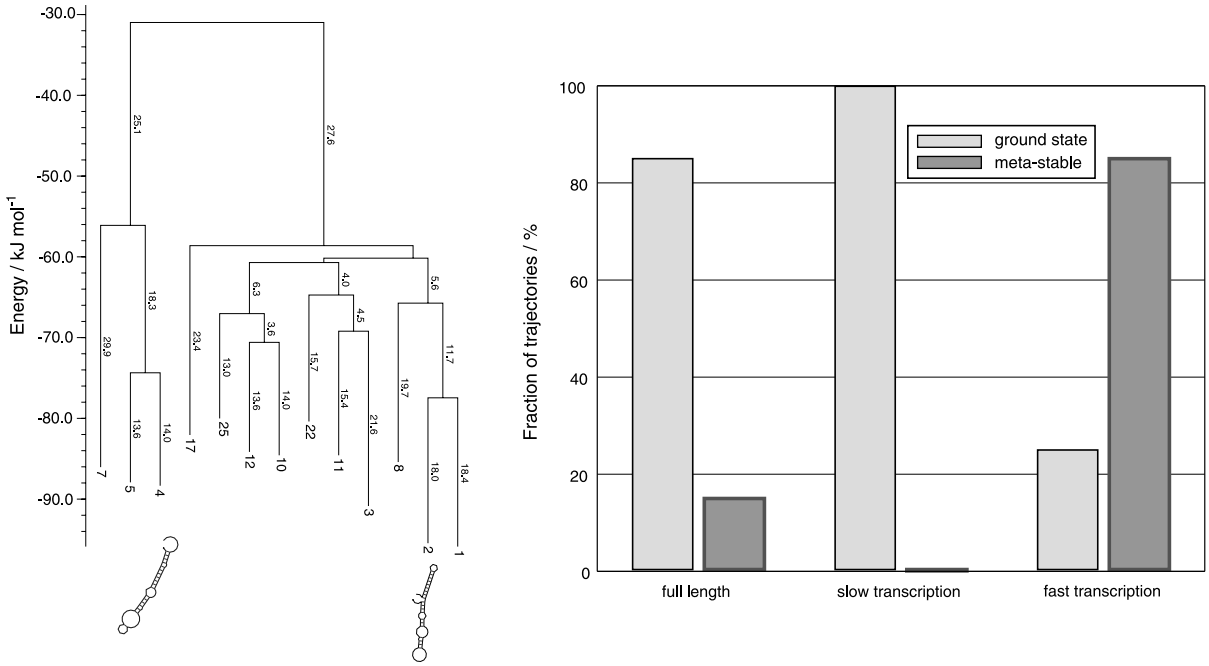

Fig. 4 *Left* Barrier tree for the *E. coli* attenuator sequence, and structures of the ground state 1 and deepest local minimum 4. *Right* Fraction of trajectories ending in either the ground state or meta-stable state. Simulations were performed using the kinfold program either on the full length sequence or with co-transcriptional folding and two different transcription speeds

where transcription is temporarily stalled. Recent evidence suggests that these pause sites are indeed functional, guiding the folding process in order to avoid the formation of severely misfolded intermediates [11].

RNA polymerase II is a highly conserved enzyme that has been studied extensively and shows little variation even between bacteria and eukaryotes. Over the past decade, a fairly large amount of structural, biochemical, and kinetic information about the polymerase [72] and the fundamental biological process of transcription [73] has been collected, and single molecule methods have been established to detect transcription pause sites [74]. Nevertheless, no detailed mechanistic model has been put together that is valid across species, and allows the prediction of transcription speed or pause sites [75, 76]. The above studies imply that pause sites are not just determined by sequence signals but also by the *RNA* structure. A truly faithful simulation of folding during transcription would therefore have to include the interplay between the structure currently formed by the nascent *RNA* strand with transcription speed of the polymerase.

Concluding remarks

In contrast to the prediction of ground state structures and equilibrium properties, modelling of *RNA* folding dynamics remains a challenging problem. Since most approaches are computationally expensive, it is important to choose a method that is suitable for the size of the *RNA* in question.

For short *RNA*s of up to around a hundred nucleotides, methods operating at the resolution of single base pairs are most suitable and will presumably provide the highest accuracy. Helix based approaches are much faster, and therefore represent the method of choice for medium size *RNA*s. The web based Kinefold and RNAkinetics servers fall in this category and allow the simulation of sequences up to 400 and 300 nts.

An interesting alternative is the analysis of energy landscapes. The resulting barrier trees provide a convenient summary of possible folding scenaria without the need to sample trajectories from different initial states. In addition, barrier trees form the basis for a coarse graining such that the folding dynamics can be solved exactly in the reduced conformation space.

Unfortunately, a systematic benchmark comparing the accuracy of different methods is difficult due to the small number and limited resolution of experimental measurements. Nevertheless, there are several examples where computational results were shown to be in good qualitative agreement with experiment. Occasionally, kinetic folding is used as means to include pseudoknots in *RNA* structure prediction. However, current attempts to derive free energies for pseudoknotted structures are still quite rough. Given the poor accuracy of pseudoknot energies compared to regular secondary structure elements, it is not certain that pseudoknot inclusion leads to an overall improvement in prediction accuracy.

In the future, we expect that the design of *RNA*/*DNA* molecules with particular dynamic properties will become an important application for the methods discussed here, especially in the emerging fields of synthetic biology and nucleic acid based nanotechnology.

Acknowledgements

This work was supported by the European Union as part of the FP-6 *EMBIO* project, as well as the Austrian GEN-AU project "Bioinformatics Integration Network II".

References

1. Zuker M, Stiegler P (1981) Nucl Acids Res 9:133
2. Zuker M (1989) Science 244:48
3. Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Biopolymers 49:145
4. McCaskill JS (1990) Biopolymers 29:1105
5. Ding Y, Lawrence CE (2003) Nucl Acids Res 31:7180
6. Nagel JH, Flamm C, Hofacker IL, Franke K, de Smit MH, Schuster P, Pleij CW (2006) Nucl Acids Res 34:3568
7. Micura R, Höbartner C (2003) Chem Biochem 4:984
8. Fürtig B, Buck J, Manoharan V, Bermel W, Jaschke A, Wenter P, Pitsch S, Schwalbe H (2007) Biopolymers 86:360
9. Harlepp S, Marchal T, Robert J, Léger J, Xayaphoummine A, Isambert H, Chatenay D (2003) Eur Phys J E-Soft Matter 12:605
10. Pan T, Sosnick T (2006) Annu Rev Biophys Biomol Struc 35:161
11. Wong TN, Sosnick TR, Pan T (2007) Proc Natl Acad Sci USA 104:17995
12. Nudler E, Mironov AS (2004) Trends Biochem Sci 29:11
13. Narberhaus F, Waldminghaus T, Chowdhury S (2006) FEBS Microbiol Rev 30:3
14. Winkler WC, Breaker RR (2005) Annu Rev Microbiol 59:487
15. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS (2004) Trends Gen 20:44
16. Nagel JHA, Pleij CWA (2002) Biochimie 84:913
17. Yanofsky C (2007) RNA 13:1141
18. Gerdes K, Wagner EGH (2007) Curr Opin Microbiol 10:117
19. Morgan SR, Higgs PG (1996) J Chem Phys 105:7152
20. Meyer IM, Miklós, I (2004) BMC Mol Biol 5:10
21. Xayaphoummine A, Viasnoff V, Harlepp S, Isambert H (2007) Nucl Acids Res 35:614
22. Flamm C, Fontana W, Hofacker IL, Schuster P (2000) RNA 6:325
23. Xia T, SanatLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Biochemistry 37:14719
24. Mathews DH, Sabina J, Zuker M, Turner H (1999) J Mol Biol 288:911
25. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH (2004) Proc Natl Acad Sci USA 101:7287
26. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) J Chem Phys 21:1087
27. Bortz AB, Kalos MH, Lebowitz JL (1975) J Comput Phys 17:10
28. Gillespie DT (1976) J Comput Phys 22:403
29. Schmitz M, Steger G (1996) J Mol Biol 225:254
30. Gultyaev AP, van Batenburg FHD, Pleij CWA (1995) J Mol Biol 250:37
31. Shapiro B, Bengali D, Kasprzak W, Wu J (2001) J Mol Biol 312:27
32. Ndifon W (2005) Biosystems 82:257
33. Mironov AA, Kister AE (1985) J Biomol Struct Dyn 2:953
34. Mironov AA, Lebedev VF (1993) Biosystems 30:49
35. Isambert H, Siggia ED (2000) Proc Natl Acad Sci USA 97:6515
36. Danilova LV, Pervoud DD, Favorov AA, Mironov AA (2006) J Bioinform Comput Biol 4:589
37. Tacker M, Fontana W, Stadler PF, Schuster P (1994) Eur Biophys J 23:29
38. Suvernev AA, Frantsuzov PA (1995) J Biomol Struct Dyn 13:135
39. Jacob C, Breton N, Daegelen P, Peccoud J (1997) J Chem Phys 107:2913
40. Zhang W, Chen SJ (2006) Biophys J 90:765
41. van Batenburg EFHD, Gultyaev AP, Pleij CWA, Ng J, Oliehoek J (2000) Nucl Acids Res 28:201
42. Bois JS, Venkataraman S, Choi HMT, Spakowitz AJ, Wang ZG, Pierce NA (2005) Nucl Acids Res 33:4090
43. Gultyaev AP, van Batenburg EFHD, Pleij CWA (1999) RNA 5:609
44. Aalberts DP, Hodas NO (2005) Nucl Acids Res 33:2210
45. Chen SJ, Dill KA (1998) J Chem Phys 109:4602
46. Bundschuh R, Hwa T (2002) Phys Rev E65:032903
47. Lucas A, Dill KA (2003) J Chem Phys 119:2414
48. Sheng YJ, Mou YC, Tsao HK (2006) J Chem Phys 124:124904

49. Chen SJ, Dill KA (1995) J Chem Phys 103:5802
50. Zhang W, Chen SJ (2001) J Chem Phys 114:7669
51. Cao S, Chen SJ (2005) RNA 11:1884
52. Olson WK (1975) Macromolecules 8:272
53. Cao S, Chen SJ (2006) Nucl Acids Res 34:2634
54. Cao S, Chen SJ (2007) J Mol Biol 367:909
55. Xayaphoummine A, Bucher T, Isambert H (2005) Nucl Acids Res 33:W605
56. Martinez HM (1984) Nucl Acid Res 12:323
57. Li W, Wu J (1998) Bioinformatics 14:700
58. Abrahams JP, van den Berg M, van Batenburg E, Pleij C (1990) Nucl Acids Res 18:3035
59. Geis M, Flamm C, Wolfinger MT, Hofacker IL, Middendorf M, Mandl C, Stadler PF, Thurner C (2007) J Mol Biol submitted
60. Flamm C, Hofacker IL, Stadler PF, Wolfinger MT (2007) Z Phys Chem 216:155
61. Ferreira FF, Fontanari JF, Stadler PF (2000) J Phys A: Math Gen 33:8635
62. Hallam J, Prügel-Bennett A (2005) IEEE Trans Evol Comp 9:385
63. Flamm C, Hofacker IL, Stadler BMR, Stadler PF (2007) Saddles and barrier in landscapes of generalized search operators. In Foundations of Genetic Algorithms volume 4436/2007 of LNCS. Springer-Verlag, p 194
64. Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF (2004) J Phys A: Math Gen 37:4731
65. Tang X, Kirkpatrick B, Thomas S, Song G, Amato NM (2005) J Comp Biol 12:862
66. Tang X, Thomas S, Tapia L, Amato NM (2007) Tools for simulating and analyzing RNA folding kinetics. In Research in Computational Molecular Biology volume 4453/2007 of LNCS. Springer-Verlag, p 268
67. Tinoco I Jr, Li PT, Carlos B (2006) Q Rev Biophys 39:325
68. Gerland U, Bundschuh R, Hwa T (2004) Phys Biol 1:19
69. Mironov AA, Kister AE (1986) J Biomol Struct Dyn 4:1
70. Heine C, Scheuermann G, Flamm C, Hofacker IL, Stadler PF (2006) IEEE Trans Vis Comp Graphics 12:781
71. Artsimovitch I, Landick R (2000) Proc Nat Acad Sci USA 97:7090
72. Greive SJ, von Hippel PH (2005) Natuer Rev Mol Cell Biol 6:221
73. Bai L, Santangelo TJ, Wang MD (2006) Annu Rev Biophys Biomol Struct 35:343
74. Herbert KM, La Porta A, Wong BJ, Mooney RA, Neuman KC, Landick R, Block SM (2006) Cell 125:1083
75. Bai L, Shundrovsky A, Wang MD (2004) J Mol Biol 344:335
76. Tadigotla VR, Maoiléidigh DÓ, Sengupta AM, Epshtein V, Ebright RH, Nudler E, Ruckenstein AE (2006) Proc Nat Acad Sci USA 103:4430

Verleger: Springer-Verlag GmbH, Sachsenplatz 4–6, 1201 Wien, Austria. – Herausgeber: Österreichische Akademie der Wissenschaften, Dr.-Ignaz-Seipel-Platz 2, 1010 Wien, und Gesellschaft Österreichischer Chemiker, Eschenbachgasse 9, 1010 Wien, Austria. – Redaktion: Getreidemarkt 9/163-OC, 1060 Wien, Austria. – Satz und Umbruch: Thomson Press Ltd., Chennai, India. – Offsetdruck: Krips bv, Kaapweg 6, 7944 HV Meppel, The Netherlands. – Verlagsort: Wien. – Herstellungsort: Meppel. – Printed in The Netherlands.