Folding Kinetics of Large RNAs

Michael Geis [a], Christoph Flamm [b], Michael T. Wolfinger [b], Ivo L. Hofacker [b], Martin Middendorf [c], Christian Mandl [g], Peter F. Stadler [d,a,e,b,f], Caroline Thurner [g,b,*]

[a] *Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany*

[b] *Institute of Theoretical Chemistry University of Vienna, Währingerstraße 17, 1090 Wien, Austria*

[c] *Parallel Computing and Complex Systems, Department of Computer Science, University of Leipzig, Johannisgasse 26, 04103 Leipzig, Germany*

[d] *Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany*

[e] *Fraunhofer Institut für Zelltherapie und Immunologie – IZI Deutscher Platz 5e, 04103 Leipzig, Germany*

[f] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[g] *Clinical Institute of Virology, Medical University of Vienna Kindespitalgasse 15, 1095 Wien, Austria*

Abstract

We introduce here a heuristic approach to kinetic RNA folding that constructs secondary structures by stepwise combination of building blocks. These blocks correspond to sub-sequences and their thermodynamically optimal structures. These are determined by the standard dynamic programming approach to RNA folding. Folding trajectories are modeled at base pair resolution using the Morgan-Higgs heuristic and a barrier tree based heuristic to connect combinations of the local building blocks. Implemented in the program `Kinwalker`, the algorithm allows co-transcriptional folding and can be used to fold sequences of up to about 1500 nucleotides in length. A detailed comparison with several well-studied examples from the literature, including the delayed folding of bacteriophage cloverleaf structures, the ASR riboswitch, and the Hok RNA, shows an excellent agreement of predicted trajectories and experimental evidence. The software is available as part of the `Vienna RNA Package`.

Key words: RNA folding kinetics, Co-transcriptional folding, Folding pathway, Metastable structure, Folding Trajectories

Manuscript *10 November 2007*

1 Introduction

Naturally evolved RNA sequences typically have been optimized by natural selection to adopt their correct functional structure efficiently on a biologically relevant timescale. Given the large number of possible conformations, this implies that the natural RNAs should differ substantially from random sequences in their folding kinetics. In a cellular context, furthermore, the nascent RNA molecule starts to fold before the transcription process is completed. Co-transcriptional folding is strongly affected by the speed of elongation, site-specific pausing of the RNA polymerase and interactions of the nascent RNA molecule with proteins or small-molecule metabolites (1),(2). Since transcription is a sequential process, the 5' region of a native helix is synthesized before its 3' portion. Variations in the rate of transcription thus may give nearby segments of the nascent RNA molecule time to form alternative — non-native — structures through folding mechanisms such as strand displacement and branch migration.

Several detailed case studies demonstrated that nature exploits the potential of RNA sequences to form multiple alternative metastable structures to implement highly sensitive molecular switches capable of controlling gene expression at the level of the mRNA. One widespread mechanism is the attenuation of transcription found in many bacterial operons related to the bio-synthesis of amino acids (3; 4). Another impressive example is the control of plasmid R1 maintenance in *E. coli*, reviewed in (5). Furthermore, it has been shown repeatedly, that alternative conformations of the same RNA sequence can perform completely different functions, e.g. (6; 7; 8). A thorough analysis of the dynamics of RNA folding and re-folding is thus a necessary prerequisite for a detailed understanding of the functionality of many RNA molecules.

In contrast to protein folding, the secondary structures of nucleic acids provide a level of description that is sufficient to understand the thermodynamics and kinetics of RNA folding (9) — at least in a useful approximation. Kinetic folding algorithms have a long history in RNA bioinformatics. Initially, kinetic folding was used as an attempt to improve RNA structure prediction. Early approaches were based on using stems as building blocks (10; 11; 12; 13; 14). These algorithms generally operate on a list of all possible helices and consequently use move-sets that destroy or form entire helices in a single move. Such operations can introduce large structural changes in a single move and furthermore, *ad hoc* assumptions have to be made about the rates of helix formation and disruption. More recently, however, interest has shifted towards understanding the folding pathways themselves (15; 16; 17). In this context, a more local move-set is preferable. The extreme case are algorithms that consider opening and closing of a single base pair as basic unit of change. This approach allows the calculation of transition rates, in good approxima-

tion, from the free energies of the involved secondary structures. This idea underlies the program `Kinfold` (18), which allows the simulation of folding trajectories of moderately sized (< 100 nucleotides (nts)) RNA molecules for macroscopic time scales at single base pair resolution.

Several more recent computer programs take new regularities of RNA structure and tertiary interactions into account, e.g. pseudoknots (19; 20) or base triplets (21). For a proper description of this broader class of structural motifs a statistical mechanics polymer model (e.g. virtual bond model (22), Gaussian chain model (19), or lattice-based models (23)) is indispensable for a rigorous treatment of excluded volume effects and the conformational entropy of the non-local interactions.

A viable alternative (24) to folding simulations is the explicit analysis of the folding energy landscape via a decomposition into basins of attraction and connecting ensembles of transition states. This approach first constructs a compact representation of the energy landscape in the form of a hierarchical structure termed barrier tree. Such tree structures have been developed independently for different classes of disordered systems, including spin glasses (25), potential energy surfaces in protein folding (26; 27), molecular clusters (28; 29) and RNA secondary structures (18). Assuming that the basins of individual local minima are in quasi-equilibrium, the rates between all local minima can be calculated during barrier tree construction. The resulting rate matrix is used to solve the approximated master equation explicitly and the folding kinetics can be computed for arbitrarily long folding times (30).

On the one hand, kinetic folding via Monte Carlo simulation (18; 19; 31; 32) is very fast for single trajectories, but a meaningful analysis of the folding path of an RNA molecule requires statistics over a fairly large sample (> 2000, say) of individual trajectories. With the size of the configuration space, furthermore, the number of trajectories necessary to obtain meaningful averages increases due to exponential increase in the number of local minima in the energy landscape. This requirement effectively limits applicability of these methods to short sequences and moderate barrier heights. On the other hand, approaches that use the explicit solution of the master equation (20; 30) are based on the enumeration of the low energy conformations of the structure space of a given RNA sequence. Since the number of low-energy conformations also grows exponentially with sequence length, these methods also cannot be applied to long sequences.

Here we describe an alternative approach that is based on the empirical observation that known metastable states appear to consist of locally optimal substructures (33; 34). In the RNA context, locally optimal substructures can efficiently be calculated by dynamic programming. The restriction to a comparatively small subset of thermodynamically determined intermediates allows

us to exploit thermodynamic-based RNA folding in a kinetic folding context. Operating on the set of all these substructures, the main idea of our approach is to find a refolding path that consists of a sequence of combinations of locally optimal substructures. This approach allows us to study kinetic effects in RNAs of currently up to about 1500 nts, i.e., the size of mitochondrial SSU mRNAs.

A range of 1500 nts covers most of the important regulatory RNA elements that are dependent on refolding effects such as naturally occurring riboswitches (35), self-induced RNA switches such as the *hok* family mRNAs (36) or the fine-tuned system of retarded cloverleaf formation in the case of Bacteriophage MS2 (37) and makes them accessible to computational prediction.

2 Theory

2.1 RNA Secondary Structures

`Kinwalker` is a heuristic that calculates a folding trajectory for an RNA sequence, i.e., a sequence of secondary structures connecting the unfolded state with the thermodynamic ground state.

We consider here only proper RNA secondary structures, i.e., structures without pseudoknots. Secondary structures are thus lists of base pairs (i, j), with $i < j$, such that (i) each nucleotide (nt) i takes part in at most one base pair, (ii) $j - i > 3$, and (iii) two base pairs (i, j) and (k, l) do not cross, i.e., $i < k < j$ implies $k < l < j$. A collection of adjacent base pairs (i, j), $(i + 1, j - 1)$, $\ldots$, $(i + \ell, j - \ell)$ is called a stack. Stacks encapsulate the dominant stabilizing contributions, while the loops that connect the stacks with each other are associated with destabilizing entropic contributions, see (38) for details on the standard energy model. The energies attributed to RNA secondary structures are free energies because they comprise both enthalpic and entropic contributions (arising from summing over different spatial conformations of the unpaired loop regions).

For each subsequence $(x_i, \ldots, x_j)$, dynamic programming algorithms are available that compute the corresponding most stable (minimum free energy, mfE) structure, subject to the condition that the delimiting bases i and j form a base pair. These energy values of c_{ij} are obtained recursively by explicitly considering the energetically different loop types (hairpin loops, bulges, interior loops, multi-branch loops) as well as stacked base pairs and are stored in a $C_{i,j}$-matrix. Standard backtracking can be used to retrieve the actual structures from the dynamic programming tables. For details we refer to (39; 40).

In general, optimal substructures on overlapping intervals will not be consistent with each other. We say that base pairs or substructures are in conflict when the "non-crossing" condition (ii) above is violated. As we shall see below, an important issue in the `Kinwalker` algorithm is the resolution of base pair conflicts when attempting to combine overlapping substructures.

2.2 Overview

The typical scenario is that the RNA sequence is gradually transcribed as the folding process progresses, although the same approach can also be employed starting from an arbitrary structure that already has full length. In the latter scenario, one will typically start from the denatured state (represented by the open chain). In the case of co-transcriptional folding, newly transcribed bases are initially unpaired.

For all subsequences of the RNA sequence the mfE structures and their energy values are precomputed by `Kinwalker` using the $C_{i,j}$ matrix for forward recursion of the standard dynamic programming algorithm for secondary structure prediction (39). In practice, we use there the implementation contained in the `ViennaRNA package`[1] (41). All subsequences are stored in a list L (see 2.3).

`Kinwalker` splits the folding process into a series of events where each event can either be a folding event or a transcription event. In each folding event a subsequence (i, j), $1 \leq i < j \leq n$, of the already transcribed RNA sequence is selected (details about the selection process can be found in Subsection 2.3) and a new structure is formed by combining base pairs from the current structure with base pairs from the mfE structure of the subsequence (i, j). This is done in such a way that the new structure includes base pairs from both structures in an energetically favorable manner (details are described in Subsection 2.5).

In each transcription event one base from the RNA sequence is appended to the already transcribed and (partially) folded subsequence. `Kinwalker` executes transcription events at regular time intervals. The number of bases transcribed per second is set by the user via the parameter *transcription_rate*. Typical values for the speed of the transcription process in nature are 10-20 nucleotides/sec for eukaryotes, 20-80 nucleotides/sec for bacteria and about 200 nucleotides/sec for bacteriophages, see (1).

Folding events occur both between transcription events and after the last transcription event when the full length RNA sequence is transcribed. `Kinwalker` estimates the waiting times for individual folding events depending on the

[1] `http://www.tbi.univie.ac.at/RNA`

height of the energy barrier between the current structure and the new structure into which the molecule is folded (details are explained in Subsection 2.4).

The current state of the folding process of a molecule can be visualized in terms of the upper triangular $C_{i,j}$-matrix, see figure 1. During the folding process, optimal structures on certain sequence intervals (i, j) are added to the growing structure. In this case we say that all subintervals (k, l) of (i, j) are *covered* (in the current structure). In the beginning all matrix entries are uncovered. The intervals covered by (i, j) correspond to the (red) triangle extending from (i, j) to the diagonal of the matrix. A covered element (i, j) is called non-dominated if there exists no covered element $(k, l) \neq (i, j)$ with $i \leq k$ and $j \leq l$. Otherwise it is called dominated. The set of all covered non-dominated matrix entries (i, j) describes the set of all maximal subsequences for which optimal substructures have been incorporated into the current structure so far. We use the term *front* for the set of all non-dominated matrix elements in figure 1.

When `Kinwalker` executes a folding step, a new matrix element is included into the front and elements of the previous front that become dominated are removed from the front. The extension of the front proceeds until the front consists of element $(1, n)$, i.e., until all matrix elements are covered. As `Kinwalker` is continuously trying to extend the front, i.e., to increase the number of covered matrix elements, it does not consider subsequences again where the corresponding matrix element is already covered. Thus, every time when a subsequence (i, j) is incorporated into the front, all subsequences that are proper subsequences of (i, j) are removed from L, as they cannot further contribute to the extension of the front.

In order to save CPU time, `Kinwalker` temporarily marks two types of subsequences in L as ineligible for front extension until the next folding event is performed: (i) subsequences that yield a structure which does not improve the free energy when integrated into the current structure, (ii) subsequences that are reachable only via energy barriers that are higher than the energy difference that (according to Arrhenius' law) corresponds to a time interval that exceed the time step between two consecutive transcription events.

2.3 Substructure Selection

All possible subsequences which are derived from the $C_{i,j}$-matrix used for thermodynamic prediction of secondary structures are stored in a list L and ordered according to the following criteria (in the given priority order):

1. Length $j-i$ of the subsequence. Short sequences are folded first since the ini-

tial nucleation step in hairpin formation — and presumably in the initiation of a new helical region — in general is entropy dominated. Consequently, local structure formation is favored (42).
2. Distance $\min\{i, n-(j+1)\}$ of the interval from the 5' and 3' ends of the sequence. We argue that "free" ends of the molecule can form local structures more readily than interior intervals which are already "anchored" in bulky substructures (or long tails) at both ends.
3. Sequences closer to the 5' end are selected preferentially. This rule is mostly included to break ties and is consistent with assumptions of co-transcriptional folding.

2.4 Transcription Rates and Energy Barrier Height

`Kinwalker` executes transcription events at regular time intervals until the entire RNA sequence has been transcribed. Since adding a base to the current sequence changes the energy landscape, changes of the secondary structure can occur between two transcription steps. In order to determine the secondary structure changes that are possible during a certain time interval, `Kinwalker` uses the following empirical relationship between barrier heights and first passage times:

$$t(\Delta G) = 10^{(\frac{8}{11}\Delta G - 7)}, \text{ for } \Delta G > 0 \quad (1)$$

This expression, which is derived from experiments for small hairpins (43) has also been used in `Kinfold` (18).

Equation (1) is used to compute a maximal barrier height E_{max} that can be traversed within a given time interval (until the next transcription event). It is assumed in `Kinwalker` that a barrier of height E_{max} can not be surpassed when taking a path from the current structure to a new structure. After each folding event Equation 1 is used to determine the corresponding first passage time for the energy that was traversed to move from the previous to the new structure. Then the first passage time is added to the current time. This reduces the time that is left until the next transcription event happens. Therefore, the maximum energy barrier E_{max} that a folding event can surpass before the next transcription event happens is reduced accordingly. If a transcription event occurs, the time counter is advanced to the next integer multiple of the transcription rate and the energy barrier E_{max} is reset to its maximal value as given by the inverse of function f (which exists at values greater zero) at 1 divided by the transcription rate. In the case that the entire RNA sequence has been transcribed, the transcription step is replaced by an energy barrier increment step where E_{max} is set to the smallest integer value at which a folding event can occur. This reflects the fact that after transcription structure transitions occur over progressively higher barriers.

To combine a structure S_2 with the current structure S_1, the algorithm considers all stacks in the set $S_2 \setminus S_1$. A single base pair that has no adjacent base pair is counted here as a stack. Starting with structure S_1 the stacks are considered from outside to inside and for each stack as many base pairs as possible are integrated into the current structure as long as this improves the free energy. This is done progressively as described in the following, see also figure. 2.

For a given stack s in $S_2 \setminus S_1$, all of its base pairs that do not conflict with the current structure are added. The resulting structure is denoted S'. Then, proceeding from the inside of the stack to the outside, the remaining base pairs of the stack s are added iteratively one at a time and those base pairs in the current structure that are in conflict with it are removed. In each step the resulting structure and its free energy are recorded. This process is repeated again starting at S' but with the difference that the remaining base pairs of the stack are now considered from outside to inside. Then the structure with the lowest free energy among all the structures that have been recorded for this stack, including the original structure S', is selected. If the free energy of the selected structure is lower than the free energy of S_1, the selected structure replaces S_1. Otherwise, S_1 remains unchanged and thus s makes no contribution, i.e. no base pair of the stack has been added to S_1. If refolding into the thus obtained structure does not succeed, either because of the height of the saddle point or because the structure's free energy is too high, refolding into another combined structure is attempted. That structure is comprised of all basepairs in S_2 as well as those basepairs in S_1 not in conflict with S_2.

2.6 Saddle Point Heuristics

A crucial step in the `Kinwalker` algorithm is the determination of the energy barrier between two locally optimal conformations. In the case of short sequences this problem can be solved by completely generating the lowest regions of the energy landscape using e.g. `RNAsubopt` (44) and subsequent explicit computation of the barrier tree (18; 24). This procedure, however, is too time-consuming for larger RNAs ($n > 100$, say). Hence heuristic approaches have to be employed which explicitly construct a (re)folding path between the two structures. The saddle height is then estimated as the highest point along the path.

The best known algorithm for approximating saddle heights between RNA conformations is the Morgan-Higgs heuristic (45), which tries to find a *direct*

folding path from an origin secondary structure to a target secondary structure where the maximum height along the path is minimal. In order to find such a path the heuristic iteratively adds base pairs from the set of base pairs in the target sequence that are not included in the current structure. For each structure that is obtained after an addition of a base pair the free energy is recorded. To avoid conflicts, immediately before a base pair is added, all base pairs in the current structure that conflict with the pair to be added are removed. The free energy of the structure that results from such a deletion is recorded as well. The height of the saddle point between the origin and target structure is estimated by the height of the highest point of the path whose maximum energy along all trajectories is the lowest of all paths tried.

The heuristic works in several rounds. In each round those base pairs that have the smallest number of conflicts with the current structure are added to it. The set of all these base pairs of the target structure is called the conflict group. For each permutation of the conflict group a folding path is calculated. For each such permutation the base pairs are added in the order as given by the permutation. Before a base pair (i, j) is added, all base pairs in the current structure that are in conflict with it are deleted. Then (i, j) is added. Every addition of a base pair and every deletion of a set of base pairs constitutes a step in the folding path. The best subpath with the lowest saddle point is accepted as a partial path and the next round starts with the new structure as origin. Once all base pairs occurring in the target structure have been added, any remaining base pairs in the current structure that are not present in the target structure are removed, thus yielding the target structure. The heuristic returns the concatenation of all partial paths as its estimate for the lowest folding path.

We have modified the original Morgan-Higgs heuristic by adding two parameters that affect the frequency of building and the treatment of conflict groups. Parameter *lookahead* denotes the maximum length of partial paths that is considered and thus the number of base pairs within a conflict group that is considered when creating a path. Thus, for each *lookahead*-tuple of members of the conflict groups a subpath is computed. Two possibilities how to handle the update of the conflict group after a partial path of length *lookahead* has been accepted are considered. Method *Standard* does not recalculate the conflict group after *lookahead* base pairs have been added to the current structure. Base pairs that are not in the conflict might have their number of conflicts reduced when a base pair is added to the structure and therefore the base pairs in conflict with the new base pair are removed from the structure. Therefore we also considered a method that always updates the conflict group after a step where *lookahead* base pairs were removed from the conflict group and have been added the current structure. This method is called *Regroup*. Note, that using a small value for parameter *lookahead* can save much computation time for a large conflict group because with method *Standard* only

$\lceil n/lookahead \rceil \times (n!/((n-lookahead)!lookahead!)$ partial paths have to be considered as opposed to the case in the standard Morgan-Higgs heuristics where $n!$ many subpaths — one for each permutation of the conflict group — are considered. If not stated otherwise, `Kinwalker` was used with the *lookahead* method and *lookahead*=1 and *Standard* in the experiments.

There are two further modifications to the heuristic that the user can choose. The first allows the folding of partial trajectories in the case that the entire trajectory between structures crosses an energy barrier that is too high. In this case, the last structure on the trajectory that lies below the allowed energy barrier is substituted for the target structure. This behavior is enabled with the *interrupt* switch. Furthermore, it is possible to make base pair transitions more realistic by only allowing one stack of less than 3 base pairs (a GC stack of less than 2 base pairs, or a single GC pair, resp.) at a time. In other words, if another stack of this type would be created by the Morgan-Higgs heuristic, the previous one is first removed entirely from the structure. This can be achieved either by adding base pairs to it or by removing all the stack's basepairs. Which action is chosen depends on whether the stack is part of the target structure or not. Stacks which occur in the target structure and already have the correct size are not counted towards the one stack maximum, as modifying them would make the target structure unattainable via a direct path.

An alternative to the Morgan-Higgs heuristic is an approximate algorithm introduced in (46) which in addition allows some "detours" in the paths. While it yields in general better approximations to the energy barriers, it is computationally more demanding and hence applicable only to shorter sequences of length $n < 200$, say.

3 Results

3.1 Runtime

Figure 3 summarizes the computational performance of the current implementation of `Kinwalker`. Examining the call graph of the algorithm in a profiler, we found that more than 90 % of the time is spent estimating saddle points. This requires frequent energy evaluations, which account for about 50% of the runtime. We have hence spent considerable efforts to optimize the evaluation of energy differences between adjacent secondary structures.

The figure shows the runtime for two parameter settings: the default parameter setting as well as the default setting with the windowsize parameter set to 100. The latter means that during the transcription process only local optimal

substructures (i, j) of length up to 100 are considered. Once transcription is completed, all structures are considered. The graph shows that the runtime T grows with about $n^{4.643}$ for the default parameter setting and with $n^{4.3}$ if the windowsize is limited to $w = 100$. For sequences of 800-1000 nts that means a reduction of one half to two thirds in calculation time.

The performance data show that `kinwalker` is able to study kinetic effects in RNAs of currently up to about 1500 nts, i.e. the size of mitochondrial SSU mRNAs. The sequences are listed in Table 4 This range covers most of the important regulatory RNA elements that are dependent on refolding effects such as naturally occurring riboswitches (35), self-induced RNA switches as the *hok* family mRNAs (36) or the fine-tuned system of retarded cloverleaf formation in the case of bacteriophage MS2 (37), and makes them accessible to computational prediction.

3.2 Bacteriophages MS2 and KU1

The genome of the enterobacteria phage MS2, a member of the family *Leviviridae*, genus *Levivirus*, is organized as single stranded positive-strand RNA of 3569nt length coding for only 4 genes (47; 48). While the expression of coat protein, lysis protein and replicase are coupled to each other (49; 50; 51), translation of the maturation (A) protein is independent. Every virion has only one copy of A-protein, which is required for the attachment of the phage to the pilus of the bacterium (52).

The coding region of the A-protein on the viral genome is preceded by a 130 nt long untranslated region (UTR), which was shown by Groeneveld (53) to fold into a cloverleaf structure. This structure hides the Shine-Dalgarno(SD) sequence in a long-distance interaction (LDI) with an upstream complementary sequence (UCS), and thus is essential for translational control. Folding of the cloverleaf structure takes up to several minutes (54), while tRNA cloverleaf structures — although similar in size and secondary structure — fold within milliseconds (55). Experimental work (37) shows that the folding of MS2 cloverleaf structure is delayed by the formation of a small stemloop containing the UCS.

`Kinwalker` folds into the intermediate structure described by (37) directly during translation (see fig. 4, red line marked as "known intermediate") and keeps this structure until transcription of the UTR is complete. After another eight minutes, it refolds into the clover-leaf structure. `Kinwalker` thus accurately reproduces the experimental data including the time frame in which the refolding process takes place.

The phage KU1 (53; 56) is a close relative of MS2, belonging to the same genus

but to a different species (MS2 is species I, KU1 is species II). While KU1 shares the genome organization of MS2, their sequences are quite different (clustalw score 51). Very similar to the MS2 5'UTR of the A-protein, KU1 folds into a cloverleaf structure. Similar to MS2, the trajectory predicted by `Kinwalker` includes the proposed kinetic trap (37) before a refolding into the clover-leaf structure takes place. `Kinwalker` estimates the folding time at a few seconds, which is still reasonably accurate. (For the folding trajectory see the supplement.)

3.3 SV11

SV11 is an RNA species of 115 nt that is replicated by Qβ replicase. As its sequence is nearly palindromic, it is believed to result as a recombinant of the plus and minus strand of MNV-11 by duplication of its high-melting domain (57). This palindromic sequence has a strong tendency to fold into a hairpin structure. In pulse-chase experiments Biebericher could show, that the active conformation is a metastable structure formed during translation, whereas the hairpin structure is unable to replicate. After prolonged standing or short boiling the activity of SV11 was irreversibly lost (57).

The mfE energy hairpin structure was proposed in (57) using thermodynamic folding algorithms. Melting experiments lead to the assumption, that two stems are present in the metastable conformation of SV11, recognized by Qβ replicase.

The `Kinwalker` trajectory for SV11 directly leads into the metastable structure within very short time. From there, refolding into the ground-state is about 6 orders of magnitude slower. Thus both the predicted structures and the estimated time frame describe the known behavior very well. (See supplement for details.)

3.4 Adenine Sensing Riboswitch

The adenine sensing riboswitch ASR of the *Bacillus subtilis pbuE* mRNA controls a gene product that is involved in adenine transport at the transcriptional level. It is the first example of an ON switch, i.e. a switch that, when bound to the target metabolite, up-regulates gene expression (58).

`Kinwalker` predicts that at short time scales a conformation is formed to which the metabolite, if present, could bind. Such binding stabilizes the kinetically controlled initial structure (58) which allows the formation of an anti-terminator and hence enables transcription. In the absence of the metabo-

lite, the molecule refolds into a more stable structure that exhibits a terminator hairpin and shuts down gene expression. As in the previous cases, the `Kinwalker` trajectory is consistent with the experimental evidence. The Morgan-Higgs heuristic estimates the folding time at about 10 hours, which is a clear overestimate. However, if we combine the three heuristics by taking the minimum of their estimates as energy barrier, the folding time is lowered to a few seconds, which is realistic. This approach, however, is computationally very expensive and may only be used for comparatively short sequences.

3.5 HOK

A particularly impressive example for kinetic control by means of RNA restructuring is the control of plasmid R1 maintenance in *Escherichia coli*. The R1 plasmid codes for two RNAs, the host-killing (hok) toxin and the suppression-of-killing (sok) RNA, acting as an "antidote" against hok. Both RNAs are constitutively expressed and hence regulated only at the post-transcriptional level (59; 5).

The hok mRNA initially forms a highly structured conformation that is translationally inactive. Upon (slow) processing of its 3' end, it structurally rearranges to a conformation with translational activity. The sok RNA, which has a considerably shorter life-time in the cell than the hok mRNA, is an antisense RNA targeting the translationally active conformation of the hok mRNA leading to quick degradation of the resulting duplex. If the plasmid R1 is lost during cell division of *E. coli*, the pool of labile sok RNA quickly depletes and can no longer suppress translation of the activated hok mRNA to the hok toxin and cell death is induced. Expression of the hok toxin must therefore be controlled in all stages of the life cycle of the hok mRNA to avoid premature killing of plasmid-containing cells. In particular, premature activation of hok mRNA during transcription is prevented by self-induced structural switching of the growing RNA chain (60) between metastable structures that conceal the ribosome binding site. A detailed description of the mechanism is also given in (36).

Different versions of the `Kinwalker` heuristic yield somewhat different structures along their trajectories. The variant of Morgan-Higgs heuristic, which allows only one stack of size less than 3 when constructing folding paths, (see Subsection 2.6), finds a *metastable nascent transcript* after transcription of 172 nts, that stays stable until the whole mRNA is transcribed. This metastable conformation (fig. 5(a)) differs from former predictions (61) in that here tac does not fold completely back into a stem loop with itself (stem I in (61)). This is due to a stem that was not proposed before. As a consequence of this, stem II (61), containing the upper complementary binding site (ucb)

is shifted somewhat downstream, thereby rendering stem III a little shorter than originally proposed. Thus, stem I and III are shorter than in (61), stem II is shifted downstream and stems IV and V agree with the structure proposed there. `Kinwalker` undergoes several structural rearrangements before folding into the intermediate structure shown in figure 5(c), in which it stays for a significant period of time. Here the sok RNA target site (sokT') is part of a multiloop, and thus single stranded. Refolding into the mfE structure (fig. 5(c)) happens quite fast.

The unmodified Morgan-Higgs heuristic finds the substructures labeled in (61) as II, IV, V, and most of III, which form immediately after transcription and stay stable until transcription is complete. Along the folding path into the mfE structure, `Kinwalker` folds the mRNA into an intermediate structure containing a stem that presents sokT'. This has been described as a crucial motive for antisense sok RNA binding to hok mRNA, which is supposed to happen when inactive mRNA is processed to its active form. To find this motive transiently already during the unprocessed mRNA folding path could explain observations that sok antisense RNA is able to bind to full length hok mRNA up to a certain extent (62).

4 Discussion

We have introduced here a novel approach to determine folding trajectories of large RNA molecules. The `Kinwalker` approach is based on the observation that important folding intermediates consist of locally optimal substructures or at least simple combinations of such local modules. As a consequence it is possible to restrict the conformation space dramatically, while at the same time the component structures can be efficiently obtained from the usual dynamic programming recursions of thermodynamic structure prediction. The folding process is conveniently visualized by a "folding front" that progresses from the open structure to the complete mfE structure. Since most natural examples link (re)folding to concurrent transcription, we have designed `Kinwalker` to interlace transcriptional steps and folding steps. In its current implementation, `Kinwalker` can be applied to RNAs of up to some 1500nt, i.e. the size of 16S rRNAs. It is hence suitable to investigate almost all RNAs for which kinetic effects are know to play a crucial role.

A comparison of several experimentally well-characterized folding pathways, including MS2, KU1, SV11, and ASR, with the `Kinwalker` predictions shows excellent qualitative agreement. Furthermore, the obtained folding times approximatly match the values predicted in the literature.

Hok mRNA was a most challenging test sequence for our algorithm as several

different features have to be considered at the same time. Although different heuristics yield folding pathways that slightly differ from each other, the experimental observations are reflected very well by our results. We find all of the previously described intermediate structures, yet not necessarily coexisting in a certain time interval. A main requirement for hok mRNAs must be that the host killing protein may not be translated at any state of mRNA transcription. All our heuristics consistently keep SD(hok) and SD(mok) hidden (i.e., in a paired and stacked conformation) and thus inaccessible for ribosomes during the whole folding path, although the pathways may differ considerably. Moreover, the observation that sokT' is accessible for sok antisense RNA in the unprocessed hok mRNA (62) can be explained by the transient building of the sokT' presenting stem or the multiloop containing the complete sokT' sequence as a single stranded region as suggested by the respective heuristics.

While `Kinwalker`'s heuristics typically tend to agree on a dominant folding pathway for short molecules, they highlight different plausible variants of folding pathways for long molecules. Hence we advise the user to consider the results of different combinations of `Kinwalker`'s parameters to assess the stability of the predicted folding pathway.

The `Kinwalker` approach shows that a combination of thermodynamic (dynamic programming) computations with coarse grained "local" kinetics is capable of describing kinetic effects in systems that go beyond the computational reach for both landscape-based approaches and direct stochastic simulations. Several aspects of the current implementation of `Kinwalker` seem to be amenable to further improvements: Refined techniques for conflict resolution should allow us to obtain better resolution (i.e., additional intermediates), where at present large rearrangements are predicted. As the performance of `Kinwalker` crucially depends on approximating saddle heights, further improvements to the Morgan-Higgs heuristic as well as alternative approaches will be investigated. A third possibility is to explicitly pre-compute additional types of structural building blocks from the dynamic programming tables.

Acknowledgments

This work was supported by the European Union as part the FP-6 *EMBIO* project, `http://www-embio.ch.cam.ac.uk/`, the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj. No. P-19411-B11, and the DFG Bioinformatics Initiative (BIZ-6/1-2).

References

[1] Pan, T. & Sosnick, T. (2006). RNA folding during transcription. *Annu. Rev. Biophys. Biomol. Struc.*, **35**, 161–175.

[2] Wong, T., Sosnick, T. R. & Pan, T. (2007). Folding of non-coding rnas during transcription facilitated by pausing-induced non-native structures. *Proc. Natl. Acad. Sci.* In press.

[3] Henkin, T. M. & Yanofsky, C. (2002). Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *Bioessays*, **24**, 700–707.

[4] Gollnick, P., Babitzke, P., Antson, A. & Yanofsky, C. (2005). Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. *Annu. Rev. Genet.*, **39**, 47–68.

[5] Gerdes, K. & Wagner, G. H. (2007). RNA antitoxins. *Curr. Op. Microbiol.*, **10**, 117–124.

[6] Baumstark, T., Schroder, A. R. & Riesner, D. (1997). Viroid processing: Switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation. *EMBO J.*, **16**, 599–610.

[7] Perrotta, A. T. & Been, M. D. (1998). A toggle duplex in hepatitis delta virus self-cleaving RNA that stabilizes an inactive and a salt-dependent pro-active ribozyme conformation. *J. Mol. Biol.*, **279**, 361–373.

[8] Schultes, E. A. & Bartel, D. P. (2000). One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.

[9] D. Thirumalai, N. Lee, S. A. W. & Klimov, D. K. (2001). Early events in RNA folding. *Annu. Rev. Phys. Chem.*, **52**, 751–762.

[10] Martinez, H. M. (1984). An RNA folding rule. *Nucl. Acid. Res.*, **12**, 323–335.

[11] Mironov, A. A., Dyakonova, L. P. & Kister, A. E. (1985). A kinetic approach to the prediction of RNA secondary structures. *Journal of Biomolecular Structure and Dynamics*, **2**, 953.

[12] Abrahams, J. P., van den Berg, M., van Batenburg, E. & Pleij, C. (1990). Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.*, **18**, 3035–3044.

[13] Gultyaev, A. P. (1991). The computer simulation of RNA folding involving pseudoknot formation. *Nucl. Acids Res.*, **19**, 2489–2493.

[14] Tacker, M., Fontana, W., Stadler, P. F. & Schuster, P. (1994). Statistics of RNA melting kinetics. *Eur. Biophys. J.*, **23**, 29–38.

[15] Higgs, P. G. (1995). Thermodynamic properties of transfer RNA: A computational study. *J. Chem. Soc. Faraday Trans.*, **91**, 2531–2540.

[16] Gultyaev, A. P., van Batenburg & Pleij, C. W. A. (1995). The computer simulation of RNA folding pathways using an genetic algorithm. *J. Mol. Biol.*, **250**, 37–51.

[17] Suvernev, A. A. & Frantsuzov, P. A. (1995). Statistical description of nucleic acid secondary structure folding. *J. Biomolec. Struct. Dyn.*, **13**,

135–144.

[18] Flamm, C., Fontana, W., Hofacker, I. & Schuster, P. (2000). RNA folding kinetics at elementary step resolution. *RNA*, **6**, 325–338.

[19] Isambert, H. & Siggia, E. D. (2000). Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. USA*, **97**, 6515–6520.

[20] Cao, S. & Chen, S.-J. (2007). Biphasic folding kinetics of RNA pseudoknots and telomerase RNA activity. *J. Mol. Biol.*, **367**, 909–924.

[21] Cao, S. & Chen, S.-J. (2005). Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*, **11**, 1884–1897.

[22] Olson, W. K. (1975). Configurational statistics of polynucleotide chains. A single virtual bond treatment. *Macromolecules*, **8**, 272–275.

[23] Chen, S.-J. & Dill, K. A. (1995). Statistical thermodynamics of double-stranded polymer molecules. *J. Chem. Phys.*, **103**, 5802–5813.

[24] Flamm, C., Hofacker, I. L., Stadler, P. F. & Wolfinger, M. T. (2002). Barrier trees of degenerate landscapes. *Z. Phys. Chem.*, **216**, 1–19.

[25] Klotz, T. & Kobe, S. (1994). "Valley Structures" in the phase space of a finite 3D Ising spin glass with $\pm i$ interactions. *J. Phys. A: Math. Gen*, **27**, L95–L100.

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

[27] Garstecki, P., Hoang, T. X. & Cieplak, M. (1999). Energy landscapes, supergraphs, and "folding funnels" in spin systems. *Phys. Rev. E*, **60**, 3219–3226.

[28] Wales, D. J., Miller, M. A. & Walsh, T. R. (1998). Archetypal energy landscapes. *Nature*, **394**, 758–760.

[29] Doye, J. P., Miller, M. A. & Welsh, D. J. (1999). Evolution of the potential energy surface with size for Lennard-Jones clusters. *J. Chem. Phys.*, **111**, 8417–8429.

[30] Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L. & Stadler, P. F. (2004). Exact folding dynamics of RNA secondary structures. *J. Phys. A: Math. Gen.*, **37**, 4731–4741.

[31] Schmitz, M. & Steger, G. (1996). Discription of RNA folding by simulated annealing. *J. Mol. Biol.*, **225**, 254–266.

[32] Danilova, L. V., Pervouchine, Dmitri, D., Favorov, A. V. & Mironov, A. A. (2006). Rnakinetics: a web server that models secondary structure kinetics of an elongating RNA. *J. Bioinf. Comp. Biol.*, **4**, 589–596.

[33] Morgan, S. R. & Higgs, P. G. (1996). Evidence for kinetic effects in the folding of large RNA molecules. *J. Chem. Phys.*, **105**, 7152–7157.

[34] Higgs, P. G. (2000). RNA secondary structure: Physical and computational aspects. *Quart. Rev. Biophys.*, **33**, 199–253.

[35] Vitreschak, A. G., Rodionov, D. A., Mironov, A. A. & Gelfand, M. S. (2004). Riboswitches: the oldest mechanism for the regulation of gene expression. *Trends Gen.*, **20**, 44–50.

[36] Micura, R. & Höbartner, C. (2003). On secondary structure rearrangements and equilibria of small RNAs. *Chem. Biochem.*, **4**, 984–990.

[37] van Meerten, D., Girard, G. & van Duin, J. (2001). Translational control by delayed RNA folding: identification of the kinetic trap. *RNA*, **7**, 483–494.

[38] Mathews, D. H., Sabina, J., Zuker, M. & Turner, H. (1999). Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

[39] Zuker, M. & Sankoff, D. (1984). RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, **46**, 591–621.

[40] Bompfünewerer, A. F., Backofen, R., Berhart, S. H., Hertel, J., , Hofacker, I. L., Stadler, P. F. & Will, S. (2007). Variations on RNA folding and alignment: Lessons from benasque. *J. Math. Biol.* In press.

[41] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

[42] Brion, P. & Westhof, E. (1997). Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, 113–137.

[43] Poerschke, D. & Eigen, M. (1971). Co-operative non-enzymic base recognition. 3. Kinetics of the helix-coil transition of the oligoribouridylic-oligoriboadenylic acid system and of oligoriboadenylic acid alone at acidic pH. *J. Mol. Biol.*, **62**, 361–381.

[44] Wuchty, S., Fontana, W., Hofacker, I. L. & Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.

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

[46] Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F. & Zehl, M. (2000). Design of multi-stable RNA molecules. *RNA*, **7**, 254–265. Santa Fe Institute Preprint 00-05-027.

[47] Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, **237**, 82–88.

[48] Fiers, W., Contreras, R., Duerinck, F., Haegmean, G., Merregaert, J., Jou, W. M., Raeymakers, A., Volckaert, G., Ysebaert, M., Van de Kerckhove, J., Nolf, F. & Van Montagu, M. (1975). A protein gene of bacteriophage MS2. *Nature*, **256**, 273–278.

[49] Berkhout, B., Schmidt, B. F., van Strien, A., van Boom, J., van Westrenen, J. & van Duin, J. (1987). Lysis gene of bacteriophage MS2 is activated by translation termination at the overlapping coat gene. *J Mol Biol*, **195**, 517–524.

[50] Skripkin, E. A., Adhin, M. R., de Smit, M. H. & van Duin, J. (1990). Secondary structure of the central region of bacteriophage MS2 RNA. conservation and biological significance. *J Mol Biol*, **211**, 447–463.

[51] van Himbergen, J., van Geffen, B. & van Duin, J. (1993). Translational control by a long range RNA-RNA interaction; a basepair substitution analysis. *Nucleic Acids Res*, **21**, 1713–1717.

[52] Paranchych, W. & Frost, L. S. (1988). The physiology and biochemistry of pili. *Adv Microb Physiol*, **29**, 53–114.

[53] Groeneveld, H., Thimon, K. & van Duin, J. (1995). Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? *RNA*, **1**, 79–88.

[54] Poot, R. A., Tsareva, N. V., Boni, I. V. & van Duin, J. (1997). RNA folding kinetics regulates translation of phage MS2 maturation gene. *Proc Natl Acad Sci USA*, **94**, 10110–10115.

[55] Crothers, D. M., Cole, P. E., Hilbers, C. W. & Shulman, R. G. (1974). The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *J Mol Biol*, **87**, 63–88.

[56] Groeneveld, H., Oudot, F. & van Duin, J. V. (1996). RNA phage KU1 has an insertion of 18 nucleotides in the start codon of its lysis gene. *Virology*, **218**, 141–147.

[57] Biebricher, C. K. & Luce, R. (1992). In vitro recombination and terminal elongation of RNA by Qβ replicase. *EMBO J.*, **11**, 5129–5135.

[58] Wickiser, J. K., Cheah, M. T., Breaker, R. R. & Crothers, D. M. (2005). The kinetics of ligand binding by an adenine-sensing riboswitch. *Biochemistry*, **44**, 13404–13414.

[59] Gerdes, K., Rasmussen, P. B. & Molin, S. (1986). Unique type of plasmid maintenance function: postsegregational killing of plasmid-free cells. *Proc. Natl. Acad. Sci. USA*, **83**, 3116–3120.

[60] Nagel, J. H. A. & Pleij, C. W. A. (2002). Self-induced structural switches in RNA. *Biochimie*, **84**, 913–923.

[61] Nagel, J. H., Gultyaev, A. P., Gerdes, K. & Pleij, C. W. (1999). Metastable structures and refolding kinetics in hok mRNA of plasmid R1. *RNA*, **5**, 1408–1418.

[62] Franch, T., Gultyaev, A. P. & Gerdes, K. (1997). Programmed cell death by hok/sok of plasmid R1: processing at the hok mRNA 3'-end triggers structural rearrangements that allow translation and antisense RNA binding. *J. Mol. Biol.*, **273**, 38–51.

[63] Winkler, W. C., Nahvi, A., Sudarsan, N., Barrick, J. E. & Breaker, R. R. (2003). An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nat. Struct. Biol.*, **10**, 701–707.

[64] Cannone, J. J. *et al.* (2002). The comparative rna web (crw) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. *BMC Bioinformatics*, **3**.

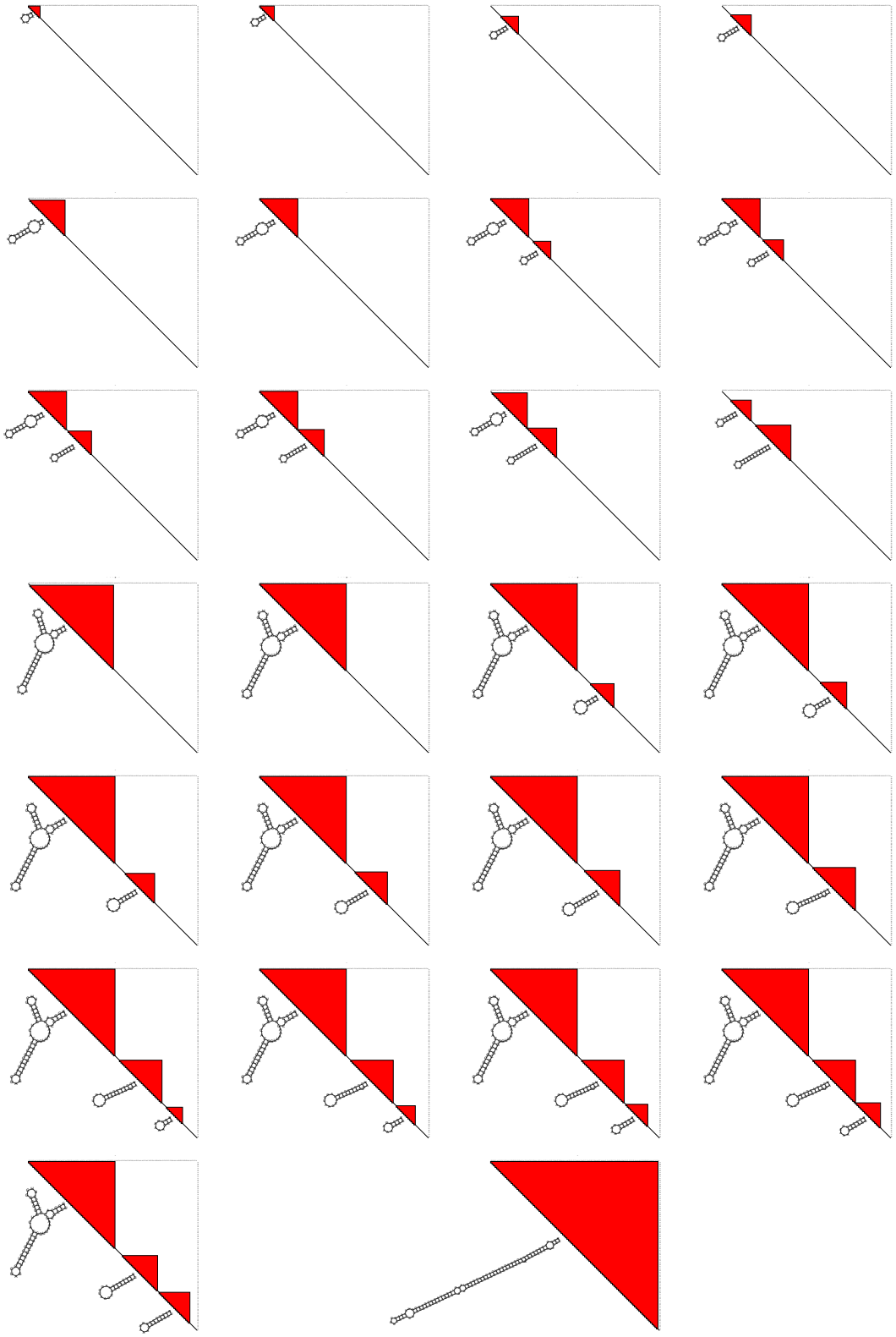

Fig. 1. Development of the front for the folding trajectory of SV11. Locally optimal substructures on the sequence interval (i, j) are represented by a triangle extending from the matrix entry (i, j) towards the diagonal. The colored areas correspond to regions over which the structure is already locally optimized. Initially the structure consists of separated local structure motives. In later stages partial refolding introduces long-range basepairs, hence the front gradually extends towards the upper-right corner $(1, n)$.

Algorithm 1 `Kinwalker`

Input: RNA sequence of length n
Output: Folding trajectory

Compute C_{ij} for (i, j) with $1 \leq i < j \leq n$
Create list L of ordered subintervals (i, j) /* Section 2.3 */
$S \leftarrow \emptyset$
$n_t \leftarrow 1$ /* sequence length at time t */
$E_{max} \leftarrow$ TimeToEnergy(Δt) /* Δt: time of one transcription step, Equation 1 */
$t_T \leftarrow 0$ /* time since last transcription event */
$E_{saddle} \leftarrow 0$
while $L \neq \emptyset$ **do**
    for all $(i, j) \in L \ \wedge \ j \leq n_t$ **do**
        $\sigma \leftarrow$BacktrackStructure(i, j) /* Section 2.5 */
        $S' \leftarrow S \cup \sigma$
        if $E_{S'} \leq E_S$ **then**
            $E_{saddle} \leftarrow$BarrierHeuristics(S', S) /* Section 2.6 */
            if $E_{saddle} \leq E_S + E_{max}$ **then**
                $t_{inc} \leftarrow$EnergyToTime$(E_{saddle} - E_S)$ /* Equation 1 */
                $t_T \leftarrow t_T + t_{inc}$
                $t \leftarrow t + t_{inc}$
                PrintOut(S', t)
                if $n_t < n$ **then**
                    $E_{max} \leftarrow$ TimeToEnergy$(\Delta t - t_T)$
                $L \leftarrow$RemoveCoveredIntervals(L, S') /* Section 2.2 */
                $S \leftarrow S'$
    if $n_t < n$ **then**
        $t \leftarrow n_t * \Delta t$
        $n_t \leftarrow n_t + 1$
        $t_T \leftarrow 0$
        $E_{max} \leftarrow$ TimeToEnergy(Δt)
    else
        $E_{max} \leftarrow$ IncreaseEnergyBarrier(E_{max}) /* Section 2.4 */

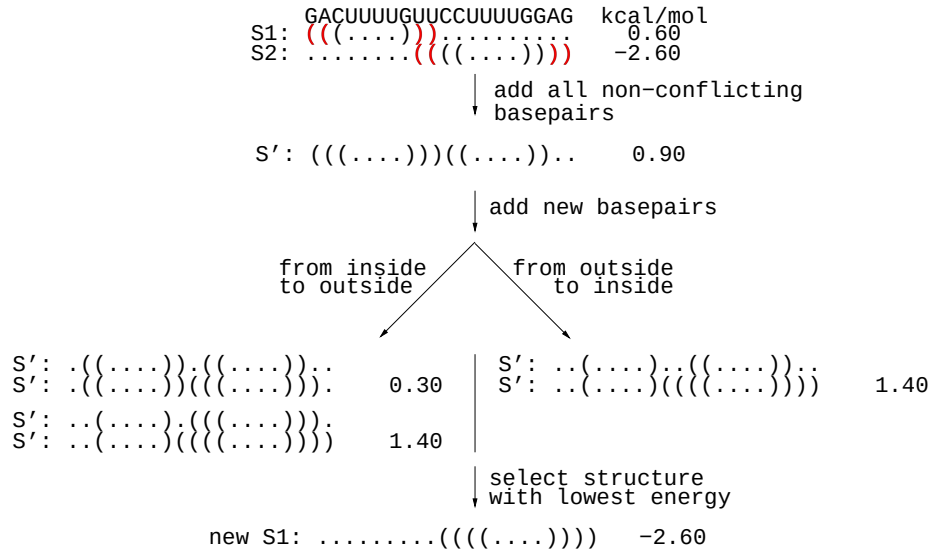

Fig. 2. Example of conflict resolution: S_1 is the current structure, S_2 is the target structure. Basepairs that conflict between S_1 and S_2 are depicted in red. Energy values for the respective structures are annotated as well. In a first step (a) all basepairs from S_2 that do not conflict with S_1 are added to an intermediate structure S'. Formation of stems may be considered in two ways: or from inside to outside, or vice versa. Although in the majority of the cases energy contributions of the fist alternative will be favorable, there exist examples that make the introduction of the alternative case reasonable. (b) Closing of the introduced stem is considered from inside to outside. `Kinwalker` now removes the first basepair of the old stem, introduces the next basepair of the new stem and evaluates the energy of the resulting structure. This procedure is repeated until no energetically favorable basepair can be added. (c) Now the new stem is closed from outside to inside. Therefore two basepairs of the old stem have to be opened. All possible basepairs of the new stem are introduced and the energy of the actual intermediate structure is evaluated. Finally `Kinwalker` sets the energetically most favorable structure from all intermediate structures S' or Structure S_2 to the new current structure S_1.

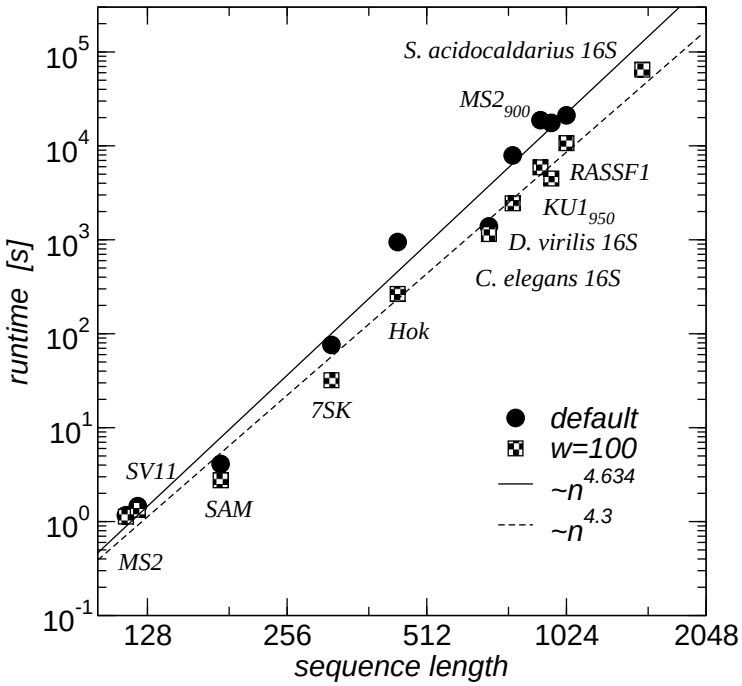

Fig. 3. Runtime of `Kinwalker` on different sequences. Computations were performed on a 64 bit machine with Intel Xeon 2.33 GHz processors and 32 GB RAM. The continuous line was produced by the standard parameter set, the dashed line was produced by the same parameter set with the windowsize $w = 100$. *S acidocaldarius* 16S was calculated only with windowsize $w = 100$, as calculation in the default mode would exceed rational calculation times. For details on used sequences see table 1 of the supplement.

sequence	length	ref
SV11	115	(57)
MS2	122	3'UTR from the whole mRNA (3569 nt) NCBI acc. no. NC_001417
$MS2_{900}$	900	from the whole mRNA (3569 nt) NCBI acc. no. NC_001417
KU1	136	3'UTR from the whole mRNA (3486 nt) NCBI acc. no. NC_002250
$KU1_{950}$	950	from the whole mRNA (3486 nt) NCBI acc. no. NC_002250
SAM	184	(63)
7SK	319	Fugu rubripes 7SK small nuclear RNA, Genbank acc. no. AJ890104.1
Hok	443	NCBI acc. no. AP000342
C. elegans 16S	697	C. elegans 16S rRNA, CompRNA acc. no. X54252
D. virilis 16S	784	D. virilis 16S rRNA, CompRNA acc. no. X05914
RASSF1	1024	H. sapiens antisense intronic RASSF1 transcript 2 mRNA, Genbank acc. no. AY545528.1
S. acidocaldarius 16S	1492	S. acidocaldarius 16S rRNA, CompRNA acc. no. D14876

Table 1
List of sequences used for testing *Kinwalker*. Accession numbers refer to NCBI, Genbank or the comparative RNA web (64) site as stated.

Fig. 4. Folding pathway of MS2 A-protein 5'UTR as it results from the Morgan-Higgs variant, which does not allow more than one stack of length less than 3 at the time. `Kinwalker` identifies the "trap structure" described by Meerten *et al.* (37), depicted here in red color and assigned as "known intermediate". The target structure corresponds to the mfE structure.

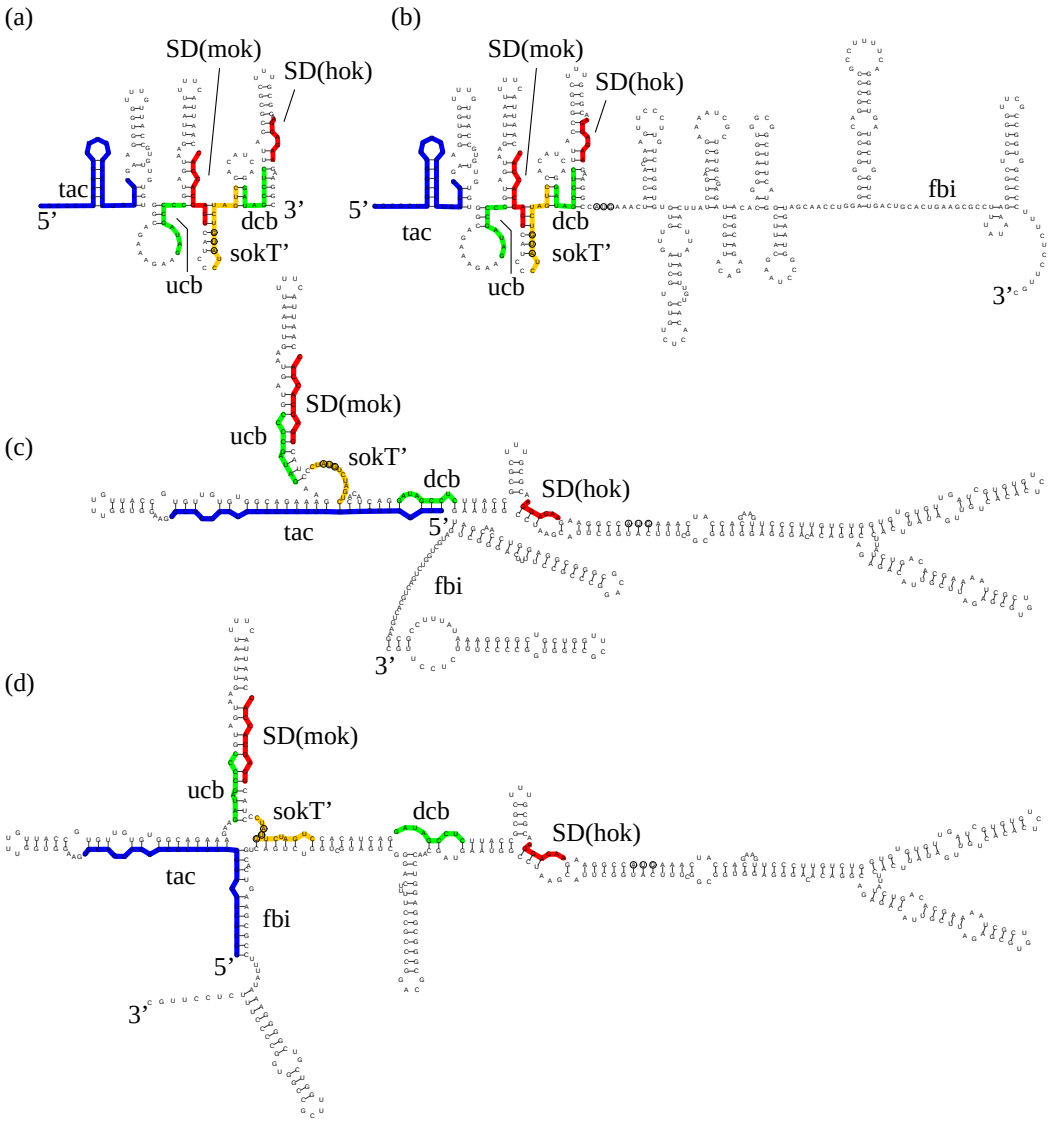

Fig. 5. Folding pathway of HOK.(a) Shows the metastable conformation when 170 nts are transcribed. In (b) transcription has terminated. The overall structure is composed by mainly short-range interactions. Several smaller rearrangements take place, leading to an intermediate structure (c), which consequently folds into (d), the mfE structure.