Cotranscriptional kinetic folding of RNA secondary structures including pseudoknots

Vo Hong Thanh[1,2*] Dani Korpela[1†]

Pekka Orponen[1‡]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

March 9, 2021

Abstract

Computational prediction of RNA structures is an important problem in computational structural biology. Studies of RNA structure formation often assume that the process starts from a fully synthesized sequence. Experimental evidence, however, has shown that RNA folds concurrently with its elongation. We investigate RNA secondary structure formation, including pseudoknots, that takes into account the cotranscriptional effects. We propose a single-nucleotide resolution kinetic model of the folding process of RNA molecules, where the polymerase-driven elongation of an RNA strand by a new nucleotide is included as a primitive operation, together with a stochastic simulation method that implements this folding concurrently with the transcriptional synthesis. Numerical case studies show that our cotranscriptional RNA folding model can predict the formation of conformations that are favored in actual biological systems. Our new computational tool can thus provide quantitative predictions and offer useful insights into the kinetics of RNA folding.

Keywords: RNA secondary structure; Cotranscriptional folding; Kinetic simulation.

[*] thanh.vo@certara.com

[†] dani.korpela@aalto.fi

[‡] pekka.orponen@aalto.fi

1 Introduction

Ribonucleic acid (RNA) is a biopolymer constituted of nucleotides with bases adenine (A), cytosine (C), guanine (G) and uracil (U). The synthesis of an RNA molecule from its DNA template is initiated when the corresponding RNA polymerase binds to the DNA promoter region. RNA has been shown to serve diverse functions in a wide range of cellular processes such as regulating gene expression and acting as an enzymatic catalyst Collins and Penny (2009); Storz (2002), and has also recently been used as an emerging material for nanotechnology Jasinski et al. (2017).

Computational prediction of RNA secondary structures given their sequences is often based on the estimation of changes in free energy, which postulates that thermodynamically an RNA strand will fold into a conformation that yields the minimum free energy (MFE) (see e.g., Fallmann et al. (2017) for a review on the topic). The energy of an RNA secondary structure can be modeled as the sum of energies of strand loops flanked by base pairs. The loop energy parameters have been measured experimentally and are detailed in a nearest neighbor parameter database (NNDB) Turner and Mathews (2009). Methods grounded in the thermodynamic framework, e.g., the Zuker algorithm by Zuker and Stiegler (1981) and its extensions Zuker (1989); Mathews et al. (1999), can be used to compute pseudoknot-free MFE secondary structures effectively in a bottom-up manner. Recent attempts to extend the Zuker algorithm to find MFE secondary structures with certain classes of pseudoknots are also proposed Rivas and Eddy (1999); Reeder and Giegerich (2004); Dirks and Pierce (2003); Akutsu (2000); Chen et al. (2009); however, finding MFE structures with pseudoknots given a general energy model is a NP-complete problem Lyngsø and Pedersen (2000).

The kinetic approach Flamm et al. (2000) is an alternative way to study the RNA folding process. It models the folding as a random process where the additions/deletions of base pairs in the current structure are assigned probabilities proportional to the respective changes in free energy values. A folding pathway of a sequence is then generated by executing stochastic simulation Flamm et al. (2000); Mironov and Lebedev (1993); Dykeman (2015). We refer to Marchetti et al. (2017) for a comprehensive review on stochastic simulation and recent work Thanh et al. (2014, 2016); Marchetti et al. (2016); Thanh et al. (2017) for state-of-the-art stochastic simulation techniques. Each simulation run on a given RNA sequence can produce a list of possible structures that it can fold into. Such dynamic view of RNA folding allows one to capture cases where local conformations are progressively folded to create metastable structures that kinetically trap the folding, thus complementing the prediction of equilibrium MFE structures produced by the thermodynamic approach.

The study of RNA structure formation often assumes that the folding process starts from a fully synthesized open strand, the *denatured state*. However, experimental evidence Watters *et al.* (2016); Pan and Sosnick (2006) has shown that RNA starts folding already concurrently with the transcription. The nucleotide transcription speed varies from 200 nt/sec (nucleotides per second) in phages, to 20-80 nt/sec in bacteria, and 5-20 nt/sec in humans Pan and Sosnick (2006). The RNA dynamics also occur over a wide range of time-scales where base pairing takes about 10^{-3} msec; structure formation is about 10-100 msec; and kinetically trapped conformations can persist for minutes or hours Al-Hashimi and Walter (2008). One consequence of considering cotranscriptional folding is that the base pairs at the 5' end of the RNA strand will form first, while the ones at the 3' end can only be formed once the transcription is complete, which leads to structural asymmetries. Cotranscriptional folding can thus form *transient* structures that are only present for a specific time period and involved in distinct roles. For instance, gene expression when considering such transient conformations of RNA during cotranscriptional folding can exhibit oscillation behavior Bratsun *et al.* (2005). We refer to the review by Lai *et al.* (2013) for further discussion on the importance of cotranscriptional effects.

In this work, we extend the kinetic approach to take into account cotranscriptional effects and pseudoknots on the folding of RNA secondary structures at single-nucleotide resolution. Our contribution is twofold. First, we explicitly consider the elongation of RNA during transcription as a primitive action in the model. The time when a new nucleotide is added to the current RNA chain is specified by the transcription speed of the RNA polymerase enzyme. The RNA strand in our modeling approach can elongate with newly synthesized nucleotides added to the sequence and fold simultaneously. To handle the transcription events, we propose an exact stochastic simulation method, the CoStochFold algorithm, to correct the folding pathway. Our method is thus able capture the effects of cotranscriptional folding at single-nucleotide resolution instead of approximating it as in previous approaches Flamm *et al.* (2000); Proctor and Meyer (2013); Mironov and Kister (1986); Zhao *et al.* (2011); Hua *et al.* (2018). Second, our algorithm allows the formation of pseudoknots, which are important for understanding RNA functions. To cope with the challenge in evaluating the energy of pseudoknotted RNA structures, we adapt the NNDB model Dirks and Pierce (2003); Andronescu *et al.* (2010) to calculate their energy values. It is worth noting that determining a reasonable energy model for RNA structures with pseudoknots is still an open question Lyngsø and Pedersen (2000); Chen *et al.* (2009). However, the advantage of our strategy in comparison with other approaches, e.g., adapting polymer theory in protein folding Dill (1999) to evaluate energy of pseudoknots Isambert and Siggia (2000), is that in the future when experimental data for pseudoknot parameters are established we can readily apply the simulation without revalidating parameters of the energy model. In addition, we facilitate the computation of energy of RNA structures with pseudoknots by employing a motif tree representation. This concept extends the coarse-grained tree representation of pseudoknot-free RNA structures, e.g., Hofacker and Stadler (2005), to allow also pseudoknotted motifs.

The rest of the paper is organized as follows. Sec. 2 reviews some background on kinetic folding of RNA. In Sec. 3, we present our work to extend the model of RNA folding to incorporate the transcription process and handle the formation of pseudoknots. Sec. 4 reports our numerical experiments on case studies. Concluding remarks are in Sec. 5.

2 Background on kinetic folding

Let S_n be a linear sequence of length n of four bases A, C, G, and U in which the 5' end is at position 1 and the 3' end is at position n. A base at position i may form a pair with a base at j, denoted by (i, j), if they form a Watson-Crick pair A-U, G-C or a wobble pair G-U. A secondary structure formed by intra-molecular interactions between bases in S_n is a list of base pairs (i, j) with $i < j$ satisfying constraints: a) the ith base and jth base must be separated by at least 3 (un-paired) bases, i.e., $j - i > 3$; b) for any base pair (k, l) with $k < l$, if $i = k$ then $j = l$; and c) for any base pair (k, l) with $k < l$, if $i < k$ then $i < k < l < j$. The first condition prevents the RNA backbone from bending too sharply. The second one prevents the forming of tertiary structure motifs such as base triplets and G-quartets. The last constraint ensures that no two base pairs intersect, i.e., there are no pseudoknots. We will relax this constraint in Sec. 3 to allow for the formation of pseudoknots during the folding.

Figure 1

Let Ω_{S_n} be the set of all possible secondary structures formed by S_n. Consider a secondary structure $x \in \Omega_{S_n}$. It can be represented compactly as a string of dots and brackets (see Fig. 1). Specifically, for a base pair (i, j), an opening parenthesis '(' is put at ith position and a closing parenthesis ')' at jth position. Finally, unpaired positions are represented by dots '.'. The dot-bracket representation is unambiguous because the base pairs in a secondary structure do not cross each other. An alternative method of representing RNA secondary structures is arc diagram. The arc diagram depicts an RNA structure as a horizontal line from 5' end (left) to 3' end (right) with arcs connecting nucleotides at positions in the sequence to show respective base pairs in the structure. The advantage of the arc diagram is that it can represent RNA structures, e.g., pseudo-knotted structures, that are difficult or impossible to visualize as planar diagrams. Fig. 1 a), b) and c), respectively, show the dot bracket, the arc diagram and the graphical visualization of tRNA molecule.

The free energy of x can be estimated by the nearest neighbor model Mathews et al. (1999), in which the free energy of an RNA secondary structure is taken to be the sum of energies of components flanked by base pairs. Formally, for a base pair (i, j) in x, we say that base k, $i < k < j$, is accessible from (i, j) if there is no other base pair (i', j') such that $i < i' < k < j' < j$. The set of accessible bases flanked by base pair (i, j) is called the loop $\mathbf{L}(i, j)$. The number of unpaired bases in a loop $\mathbf{L}(i, j)$ is its size, while the number of enclosed base pairs determines its degree. Based on these properties, loops $\mathbf{L}(i, j)$ can be classified as stacks (or stems), hairpins, bulges, internal loops and multi-loops (or multi-branch loops). The unpaired bases that are not contained in loops constitute the exterior (or external) loop $\mathbf{L}_e$.

A secondary structure x is thus uniquely decomposed into a collection of loops $x = \cup_{(i,j)} \mathbf{L}(i,j) \cup \mathbf{L}_e$. Based on this decomposition, the free energy G_x (in kcal) of secondary structure x is computed as:

$$G_x = \sum_{(i,j)} G_{\mathbf{L}(i,j)} + G_{\mathbf{L}_e} \tag{1}$$

where $G_{\mathbf{L}(i,j)}$ is the free energy of loop $\mathbf{L}(i,j)$. Experimental energy values for $G_{\mathbf{L}(i,j)}$ are available in the nearest neighbor database Turner and Mathews (2009).

Let $y \in \Omega_{S_n}$ be a secondary structure derived directly from x by an intramolecular reaction between bases i and j in x. Commonly, three operations on a pair of bases, referred to as the move set (see Fig. 2), are defined Flamm et al. (2000):

- Addition: y is derived by adding a base pair that joins bases i and j in x that are currently unpaired and eligible to pair.

- Deletion: y is derived by breaking a current base pair (i, j) in x.

- Shifting: y is derived by shifting a base pair (i, j) in x to form a new base pair (i, k) or (k, j).

Let $k_{x \to y}$ be the rate (probability per time unit) of the transition from x to y. In a conformation x, the RNA molecule may wander vibrationally around its energy basin for a long time, before it surmounts an energy barrier to escape to a conformation y in another basin. The dynamics of the transition from x to y characterizes a rare event in Molecular Dynamics (MD). Here, we adopt the coarse-grained kinetic Monte Carlo approximation Metropolis et al. (1953); Kawasaki (1966), and model the transition rate $k_{x \to y}$ as:

$$k_{x \to y} = k_0 e^{-\Delta G_{xy}/2RT} \tag{2}$$

where T is absolute temperature in Kelvin (K), $R = 1.98717 \times 10^{-3} (kcal \cdot K^{-1} \cdot mol^{-1})$ is the gas constant and $\Delta G_{xy} = G_y - G_x$ denotes the difference between free energies of x and y. The constant k_0, normally taking values in the range 10^{-4} to 10^{-3}, provides a calibration of time.

Let $P(x, t)$ be the probability that the system is at conformation x at time t. The dynamics of $P(x, t)$ is formulated by the (chemical) master equation Marchetti et al. (2017) as:

$$\frac{dP(x,t)}{dt} = \sum_{y \in \Omega_{S_n}} \Big[k_{y \to x} P(x,t) - k_{x \to y} P(x,t) \Big] \tag{3}$$

Analytically solving Eq. 3 requires to enumerate all possible states x and their neighbors y. The size of the state space $\|\Omega_{S_n}\| \sim n^{-3/2}\alpha^n$ with $\alpha = 1.8488$ increases exponentially with the sequence length n, and the number of neighbors of x is in order of $O(n^2)$ Hofacker et al. (1998). Thus, due to the high dimension of the state space, solving Eq. 3 often involves numerical simulation.

Let $P(y, \tau|x, t)$ be the probability that, given current structure x at time t, x will fold into y in the next infinitesimal time interval $[t + \tau, t + \tau + d\tau)$. We have

$$P(y, \tau|x, t) = k_{x\to y}e^{-k_x\tau}d\tau \tag{4}$$

where $k_x = \sum_{y\in\Omega_{S_n}} k_{x\to y}$ is the sum of transition rates to single-move neighbors of x. Eq. 4 lays down the mathematical framework for stochastic RNA folding. Integrating Eq. 4 with respect to τ from 0 to ∞, the probability that x moves to y is $k_{x\to y}/k_x$. Summing Eq. 4 over all possible states $y \in \Omega_{S_n}$, it shows the waiting time τ until the transition occurs follows an exponential distribution $Exp(k_x)$. These facts are the basis for our kinetic folding algorithm called StochFold presented as Algorithm 1. We note that StochFold shares the structure of the earlier algorithm Kinfold Flamm et al. (2000) and its improvements Dykeman (2015); Thanh and Zunino (2014).

Algorithm 1 StochFold

```
Require: initial RNA conformation $s_0$ and ending time $T_{max}$
initialize $x = s_0$ and time $t = 0$
repeat
    enumerate next possible conformations of the current conformation $x$ and put into set $Q$
    compute the transition rate $k_{x\to y}$ for each $y \in Q$ and total rate $k_x = \sum_{y\in Q} k_{x\to y}$
    select next conformation $y \in Q$ with probability $k_{x\to y}/k_x$
    sample waiting time to the next folding event $\tau \sim Exp(k_x)$
    set $x = y$ and $t = t + \tau$
until $t \geq T_{max}$
```

3 Cotranscriptional kinetic folding of RNA

The folding of an RNA strand adapts immediately to new nucleotides synthesized during the transcription. The kinetic approach described in Sec. 2 cannot capture the effects of such cotranscriptional folding, because it considers only interactions between bases already present in the sequence. We outline in this section an approach to incorporating these effects in the simulation. The transcription process is explicitly taken into account by extending the move set with the new operation of *elongation*. Our extended move set thus comprises four operations: addition, deletion, shifting and elongation. The first three operations are defined as in the previous section. In elongation, the current RNA chain increases in length and a newly synthesized nucleotide is added to its 3' end. Figure 2 illustrates the extended move set.

Figure 2

Under the extended move set, we define two event types: folding event and transcription event. A folding event is an internal event that occurs when one of the three operations addition, deletion or shifting, is applied to a base pair of the current sequence. A transcription event happens when the elongation operation is applied. It is an external event whose rate is specified by the transcription speed of the RNA polymerase enzyme. The occurrences of transcription events break the Markovian property of transitions between conformations. This is because when a new nucleotide is added to the current RNA conformation, the number of next possible conformations increases. The waiting time of the next folding event also changes and thus a new folding event has to be recomputed.

Algorithm 2 outlines how the CoStochFold algorithm handles this situation. The key element of CoStochFold (lines 8 - 15) is a race where the event having the smallest waiting time will be selected to update the current RNA conformation. More specifically, suppose the current structure is x at time t. Let τ_e be the waiting time to the next folding event and τ_{trans} the waiting time to the next transcription event. Assuming that no events occur earlier, τ_e has an exponential distribution with rate k_x which is the sum of all transition rates of applying addition, deletion and shifting operations to base pairs in x. For simplifying the computation of τ_{trans}, we assume that it is the expected time to transcribe one nucleotide. Let N_{trans} be the (average) transcription speed of the polymerase. We compute τ_{trans} as:

$$\tau_{trans} = 1/N_{trans} \tag{5}$$

Thus, given current time t, the next folding event will occur at time $t_e = t + \tau_e$ and, respectively, the transcription event where a new nucleotide will be added to the current sequence is scheduled at time $t_{trans} = t + \tau_{trans}$. We decide which event will occur by comparing t_e and t_{trans}. If $t_e > t_{trans}$, then a new nucleotide is first transcribed and added to the current RNA conformation. Otherwise, a folding event is performed where a structure in the set Q of neighboring structures is selected to update the current conformation.

Algorithm 2 CoStochFold

```
Require: initial RNA conformation $s_0$, transcription speed $N_{trans}$, and ending time $T_{max}$
 1: initialize $x = s_0$ and time $t = 0$
 2: set $\tau_{trans} = 1/N_{trans}$
 3: compute the next transcription event $t_{trans} = t + \tau_{trans}$
 4: repeat
 5:   enumerate next possible conformations by applying addition, deletion and shifting operations
      on the current conformation $x$ and put into set $Q$
 6:   compute the transition rate $k_{x \to y}$, for $y \in Q$, and total rate $k_x = \sum_{y \in Q} k_{x \to y}$
 7:   sample waiting time to the next folding event $\tau_e \sim Exp(k_x)$ and set $t_e = t + \tau_e$
 8:   if $(t_e > t_{trans})$ then
 9:     elongate $x$
10:     set $t = t_{trans}$
11:     compute the next transcription event $t_{trans} = t + \tau_{trans}$
12:   else
13:     select next conformation $y \in Q$ with probability $k_{x \to y}/k_x$
14:     set $x = y$ and $t = t_e$
15:   end if
16: until $t \geq T_{max}$
```

We remark that one can easily extend Algorithm 2 to allow modeling τ_{trans} as a random variable without changing the steps of event selection. Specifically, one only needs to change step 3 in Algorithm 2 to generate the waiting time of the next transcription event, while keeping the simulation otherwise unchanged.

3.1 Handling pseudoknots

This section extends the CoStochFold algorithm to include structures with pseudoknots during the enumeration of neighbor structures (see step 5, Algorithm 2). A pseudoknot occurs if there exists a crossing between two base pairs. Here we restrict to the two most common pseudoknots: the H-type and K-type (kissing hairpin) Reidys et al. (2011). We use the extended dot-bracket notation, i.e., augment the original dot-bracket with additional types of bracket pairs, e.g., [], {} and ⟨⟩, to denote the crossing base pairs. Fig. 3 depicts examples of RNA structures with H-type and K-type pseudoknots and their corresponding extended dot-bracket notations and arc diagrams. Figure 3

Let $\mathbf{L}(i,j)$ be a pseudoknot flanked by the bases i and j. We compute its energy $G_{\mathbf{L}(i,j)}$ by adapting the NNDB model Dirks and Pierce (2003); Andronescu et al. (2010); Reidys et al. (2011). The energy of a pseudoknot consists of an initiation penalty and structural penalties. The initiation penalty depends on whether the pseudoknot is unnested or nested within another multiloop or pseudoknot. The structural penalty takes into account the number of unpaired bases, nested substructures and the energy of the pseudoknotted stems. Specifically, the energy of $\mathbf{L}(i,j)$ is calculated by the formula:

$$G_{\mathbf{L}(i,j)} = \beta_{\mathbf{L}(i,j)} + P * \beta_2 + U * \beta_3 \qquad (6)$$

where $\beta_{\mathbf{L}(i,j)}$ is an initiation energy term that penalizes the formation of the pseudoknot, and P and U, respectively, denote the numbers of paired bases that flank the interior of the pseudoknot and unpaired bases inside the pseudoknot. The corresponding parameters β_2 and β_3 are used to penalize the formation of base pairs P and unpaired bases U correspondingly.

To facilitate the evaluation of the energy of an RNA structure x with pseudoknots, we first parse x to closed regions Rastegari and Condon (2007). A set of bases $\{i, i+1, ..., j\}$ is called a closed region if a) no base in the region pairs to a base outside of the interval $\{i, i+1, ..., j\}$, and b) such region cannot be partitioned into smaller closed regions. We then decompose each closed region into loops and pseudoknots. Such structural motifs will form a tree that we called a motif tree. An example of a motif tree is depicted in Fig. 4. Having the motif tree for structure x, we can traverse it from the leaves to the root to obtain the energy value G_x. Specifically, we evaluate energy values of motifs at the leaves and send them to their parents. At each inner node, we compute the sum of its energy and those of the child nodes, then propagate to the upper level. The process is done recursively until reaching the root where the total energy sum value G_x is returned. Figure 4

4 Numerical experiments

We illustrate the application of our cotranscriptional kinetic folding method on four case studies: a) the E. coli signal recognition particle (SRP) RNA Watters et al. (2016), b) the switching molecule Flamm et al. (2000), c) the Beet soil-borne virus Taufer et al. (2008) and d) the SV-11 variant in Qβ replicase Biebricher and Luce (1992). We use these examples to manifest the characteristics of our method that thermodynamic/kinetic methods Zuker and Stiegler (1981); Gultyaev et al. (1995); Flamm et al. (2000) would fail to capture if initiated from fully denatured sequences. Our cotranscriptional folding method is not only able to produce these structures, but also provides insight into mechanisms that biological systems may use to guide the structure formation process. Finally, we assess the computational performance of the proposed simulation algorithm on sequences of varying lengths. The code for the implementation of our CoStochFold algorithm is available at: `https://github.com/vo-hong-thanh/stochfold`.

4.1 Signal recognition particle (SRP) RNA

Figure 5

This section studies the process of structural formation of the E. coli SRP RNA during transcription. SRP is a 117nt long molecule, which recognizes the signal peptide and binds to the ribosome locking the protein synthesis. Its active structure is a long helical structure containing interspersed inner loops (see S3 in Fig. 5). Experimental work Watters et al. (2016) using SHAPE-seq techniques has suggested a series of structural rearrangements during transcription that ultimately result in the SRP helical structure. In particular, the 5' end of SRP forms a hairpin structure during early transcription. The structure persists until the transcript reaches a length of 117nt. The unstable hairpin then rearranges to its active structure. Fig. 5 depicts three structural motifs at 25nt (S1), 86nt (S2), and 117nt (S3), respectively, in the formation of SRP. Specifically, the hairpin motif S1 emerges at transcript length 25nt, and the transcript then continues elongating to form structure S2 at length 86nt. When reaching transcript length 117nt, SRP rearranges into its persistent helical conformation S3.

Figure 6

We validated the prediction of the CoStochFold algorithm against the experimental work in Watters et al. (2016). To do that, we performed 10000 simulation runs of the algorithm to fold SRP cotranscriptionally. The average transcription speed was set to 5 nt/sec. Fig. 6 shows the frequency of occurrences of the considered structures during the simulated time of 30 seconds. Kinetic folding starting from the denatured state was carried out by the StochFold algorithm, while cotranscriptional folding was conducted by the CoStochFold algorithm. The plot on the left shows the cotranscriptional folding of SRP and the plot on the right presents the folding of SRP starting from the denatured state. The figures clearly show that the CoStochFold algorithm can capture the folding pathway of SRP. Specifically, the hairpin motif S1 starts to form at about $t = 4$s when the transcript length is 20nt and peaks at about $t = 8$s when 40nt have been transcribed. At about $t = 18$s, Structure S2 appears and then rearranges to S3 at about $t = 24$s. We note that in the simulated folding without considering transcription only the conformation S3 is encountered.

4.2 Switching molecule

We consider the dynamic folding of an artificial RNA sequence $S =$ ”GGCCCCUUGGGGGCCAGACCCCUAAAGGGGUC” Flamm et al. (2000). Two stable conformations of the sequence are: the MFE structure $x =$ “((((((((((((((.....))))))))))))))” (−26.20 kcal), and a suboptimal structure $y =$ “((((((....)))))).((((((....))))))” (−25.30 kcal). We use this example to demonstrate how by tuning the transcription speed we can change the ratio of occurrences of structures x and y. Here we focus on the number of first-hitting time occurrences of a target structure. The number of first-hitting time occurrences of a structure in a time interval divided by the total number of simulation runs approximates the first-passage time probability of the structure, i.e., its folding time Flamm et al. (2000). Figure 7

Fig. 7 plots the number of first-hitting time occurrences of the MFE structure x and the suboptimal y with varying transcription speeds. We performed 10000 simulation runs of the CoStochFold algorithm on the sequence S in which each simulation ran until a target structure was observed or the ending time $T_{max} = 1000$ seconds was reached. The constant $k_0 = 1$ in Eq. 2 is used in this case study to scale the time. Fig. 7 shows that changing the transcription speed of the polymerase significantly affects the folding characteristics of the sequence. Specifically, cotranscriptional folding with slow transcription speed favors the suboptimal structure y. It increases the number of occurrences of y, while reducing the number of occurrences of the MFE structure x. Figure 8

Fig. 8 compares the total number of first-hitting time occurrences of the MFE structure x with respect to the suboptimal conformation y up to time $T_{max} = 1000$. We note that if the simulation starts from the fully denatured state, the occurrence ratio of the suboptimal conformation y to the MFE structure x is about 2:1, as also observed by Flamm et al. (2000). However, the ratio increases noticeably when the transcription speed decreases. For example, the occurrence ratio of the suboptimal conformation y to the MFE structure x is about 6.5:1 in the case of transcription speed 5 nt/sec.

4.3 Beet soil-borne virus

We use the beet soil-borne virus S = ”CGGUAGCGCGAACCGUUAUCGCGCA” from the PseudoBase++ database Taufer et al. (2008) to demonstrate the application of our simulation in predicting RNA structures with pseudoknots. The folding of the sequence S was simulated with 10000 runs. We evaluated the energy of pseudoknots using the energy parameters from Andronescu et al. (2010), which were estimated by fitting the standard NNDB parameters by Mathews et al. (1999) and pseudoknotted parameters by Dirks and Pierce (2003) over a large data set of both pseudoknotted and pseudoknot-free secondary structures. We compare two simulation settings: a) cotranscriptional folding of S with transcription speed 200 nt/sec, and b) the folding starting from the denatured initial state (i.e., a fully synthesized open strand). Figs. 9 depicts the occurrence frequency of the H-type pseudoknotted structure C_1 =”.(((.[[[[[[)))...]]]]]].” with an energy of -12.39 (kcal). We also consider two intermediate structures C_2 =”.(((...[[[[)))...]]]]...” and C_3 =”.(((((........))))).....” having energies of -7.25 (kcal) and -4.52 (kcal), respectively. Figure 9

Figs. 9a - 9b clearly show that the dominant structure of the beet soil-borne virus sequence S is the H-type pseudoknotted structure C_1. We also see from these figures that the folding starting from the denatured state misses the formation of intermediate structures C_2 and C_3, which appear in the cotranscriptional folding. After the transcription phase, intermediate structures will rearrange to C_1 and remain in this stable form. Figs. 9a shows that the frequency of C_1 is more than 82% in the simulation.

We conclude this section with a note about the energy parameters for RNA structures with pseudoknots. In particular, we also simulated the beet soil-borne virus S with the energy model by Reidys et al. (2011), which is another extension of the NNDB model for pseudoknots. The occurrence frequency of pseudoknotted structure C_1 estimated by the Reidys et al. (2011) model was significantly lower than by the Andronescu et al. (2010) model. This prediction discrepancy is because the energy model by Reidys et al. (2011) penalizes the formation of pseudoknots significantly more than the model by Andronescu et al. (2010). In fact, all pseudoknotted structures will be unfavourable with such high penalties for the pseudoknots. An interesting prediction from our cotranscriptional folding simulation using both energy models is the occurrence of the intermediate hairpin structure C_3. The persistence of C_3 before rearranging to the pseudoknot C_1 depends on how much penalty is applied to the formation of pseudoknots.

4.4 SV-11

SV-11 is a 115 nt long RNA sequence. It is a recombinant between the plus and minus strands of the natural Qβ template MNV-11 RNA Biebricher and Luce (1992). The result of the recombination is a highly palindromic sequence whose most stable secondary structure is a long hairpin-like structure, the MFE structure in Fig. 10a). The MFE structure, however, disables Qβ replicase because its primer regions are blocked. Experimental work Biebricher and Luce (1992) has shown that an active structure of SV-11 for replication is when it folds into a metastable conformation depicted in Fig. 10b). This is a hairpin-hairpin-multi-loop motif with open primer regions that serve as templates for replication. Transition from the metastable structure to the MFE structure has been observed experimentally but is rather slow Biebricher and Luce (1992), indicating long relaxation time to equilibrium.

Figure 10

We plot in Fig. 11 the energy vs. occurrence frequency of structures by the cotranscriptional folding of SV-11. The result is obtained by 10000 simulation runs of our CoStochFold algorithm for $t = 50$ simulated seconds and average transcription speed 5nt/sec. To determine the frequency of occurrence of a structure, we discretize the simulation time into intervals and record how much time was spent in each structure within each interval. The frequency of occurrence of a structure in each time interval is then averaged over 10000 runs. The figure shows that the folding favors metastable structures, and disfavors the MFE structure. In particular, cotranscriptional folding quickly folds SV-11 to its metastable conformations with the mode of the energy distribution at about -63kcal.

Figure 11

Fig. 12 shows the long-term occurrence frequencies of structures at different energy levels in the SV-11 folding and Fig. 13 compares the occurrence frequencies of the specific metastable structure depicted in Fig. 10b) with the MFE structure and two randomly selected suboptimal structures in the energy level of MFE structure. Fig. 13 shows that the SV-11 molecule interestingly prefers the metastable structure over the MFE structure. Specifically, the metastable structure in the cotranscriptional folding regime is in the time interval $[0, 10000]$ about tenfold more frequent than the MFE structure.

Figure 12

Figure 13

4.5 Simulation performance

This section reports the performance of our stochastic folding algorithm with RNA sequences of varying lengths from 25 to 5000nt. To estimate the computational cost of a single simulation move, we executed 10 independent simulation runs of 1000 simulation steps, each with a random sequence of the given length. The average runtime for each sequence length was computed, and then divided by the number of simulation steps to assess the single-step computation cost.

Fig. 14 plots the resulting estimated single-step computational cost of our folding algorithm in two settings: 1) simulation without pseudoknots, executed on Intel an i5-7300U dual-core CPU with a clock speed of 2.6 GHz, on the left and 2) simulation with pseudoknots, executed on an Intel i5-8365U quad-core CPU with a clock speed of 1.6 GHz, on the right. As witnessed by the figure, the simulation is quite computation intensive, especially for long sequences. For example, the simulation without pseudoknots for a sequence of length 1000nt took on average 0.1 seconds of processor time per simulation step. Thus, a single simulation run of 1000 simulation steps would take on average 100 seconds, and 10000 repeats of this would take $1M$ seconds, that is, 11.6 days of processor time.

The single-step computational cost increases with increasing sequence length: the runtime in the case of a pseudoknot-free simulation for sequences of length 5000nt is about 11 times higher than for sequences of length 1000nt. This increase is mostly due to the quadratically increasing number of possible moves in the locality of a conformation. Our detailed breakdown analysis of the computational cost of simulations shows that the cost of enumerating the possible moves contributes more than 95% of the total cost in each simulation step. (Note that the cost of enumerating the moves depends on both the number of possible moves and the algorithmics of the enumeration process.) The regression lines depicted in Fig. 14 indicate that the computational cost per single move of our folding algorithm without pseudoknots grows as $O(N^{2.25})$ and with pseudoknots as $O(N^{2.89})$, as a function of the sequence length N.

Figure 14

5 Conclusions

We propose a kinetic model of RNA folding that takes into account the elongation of an RNA chain during transcription as a primitive structure-forming operation alongside the common base-pairing operations. We developed a new stochastic simulation algorithm CoStochFold to explore RNA structure formation, including pseudoknots, in the cotranscriptional folding regime. We showed through numerical case studies that our method can quantitatively predict the formation of (metastable) conformations in an RNA folding pathway. The simulation method thus promises to offer useful insights into RNA folding kinetics in real biological systems. However, it also poses a great computational challenge for long sequences due to the huge number of possible moves in the locality of a conformation. Furthermore, many simulation runs must be performed in order to obtain a reasonable statistical estimation of the system dynamics. Several improvements are possible in future work. For instance, we can reduce the enumeration of possible moves by localizing the computation. The motif tree, a coarse-grained representation for pseudoknotted structures developed in the paper, could be useful also in this context. We decompose an RNA structure into motifs and then enumerate new conformations related to each motif. To reduce the cost for executing many simulation runs, we can employ high performance computing to run simulations in parallel.

Acknowledgements

This work has been supported by Academy of Finland grant no. 311639, "Algorithmic Designs for Biomolecular Nanostructures (ALBION)". The work of VHT has been partially done when he was at Aalto University.

References

Akutsu, T., 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. Discrete Appl. Math. 104, 45–62.

Al-Hashimi, H. M. and Walter, N. G., 2008. RNA dynamics: It is about time. Curr. Opin. Struct. Biol. 18, 321–329.

Andronescu, M. S., Pop, C., and Condon, A. E., 2010. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. RNA 16, 26–42.

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

Bratsun, D., Volfson, D., Tsimring, L. S., and Hasty, J., 2005. Delay-induced stochastic oscillations in gene regulation. PNAS 102, 14593–14598.

Chen, H.-L., Condon, A., and Jabbari, H., 2009. An $O(n^5)$ algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. J. Comp. Biol. 16, 803–815.

Collins, L. J. and Penny, D., 2009. The RNA infrastructure: Dark matter of the eukaryotic cell? Trends Genet. 25, 120–128.

Dill, K. A., 1999. Polymer principles and protein folding. Protein Science 8, 1166–1180.

Dirks, R. M. and Pierce, N. A., 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. J. Comp. Chem. 24, 1664–1677.

Dykeman, E. C., 2015. An implementation of the Gillespie algorithm for RNA kinetics with logarithmic time update. Nucleic Acids Res. 43, 5708–5715.

Fallmann, J., Will, S., Engelhardt, J., Grüning, B., Backofenc, R., and Stadler, P. F., 2017. Recent advances in RNA folding. J. Biotechnol. 261, 97–104.

Flamm, C., Fontana, W., Hofacker, I. L., and Schuster, P., 2000. RNA folding at elementary step resolution. RNA 6, 325–338.

Gultyaev, A. P., van Batenburg F. H. D., and Pleij, C. W. A., 1995. The computer simulation of RNA folding pathways using a genetic algorithm. J. Mol. Biol. 250, 37–51.

Hofacker, I. L., Schuster, P., and Stadler, P. F., 1998. Combinatorics of RNA secondary structures. Discrete Appl. Math. 88, 207–237.

Hofacker, I. L. and Stadler, P. F., 2005. RNA secondary structures. In Meyers, R. A., ed., Encyclopedia of Molecular Cell Biology and Molecular Medicine, Volume 12, 581–603. Wiley-VCH Verlag GmbH.

Hua, B., Panja, S., Wang, Y., Woodson, S. A., and Ha, T., 2018. Mimicking co-transcriptional RNA folding using a superhelicase. J. Am. Chem. Soc. 140, 10067–10070.

Isambert, H. and Siggia, E. D., 2000. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. PNAS 97, 6515.

Jasinski, D., Haque, F., Binzel, D. W., and Guo, P., 2017. Advancement of the emerging field of RNA nanotechnology. ACS Nano 11, 1142–1164.

Kawasaki, K., 1966. Diffusion constants near the critical point for time-dependent Ising models. Phys. Rev. 145, 224–230.

Kerpedjiev, P., Hammer, S., and Hofacker, I. L., 2015. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 31, 3377–3379.

Lai, D., Proctor, J. R., and Meyer, I. M., 2013. On the importance of cotranscriptional RNA structure formation. RNA 19, 1461–1473.

Lyngsø R. B. and Pedersen, C. N. S., 2000. RNA pseudoknot prediction in energy-based models. J. Comp. Biol. 7, 409–427.

Marchetti, L., Priami, C., and Thanh, V. H., 2016. HRSSA–efficient hybrid stochastic simulation for spatially homogeneous biochemical reaction networks. J. Comp. Phys. 317, 301–317.

Marchetti, L., Priami, C., and Thanh, V. H., 2017. Simulation Algorithms for Computational Systems Biology. Springer.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H., 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288, 911–940.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092.

Mironov, A. and Kister, A., 1986. RNA secondary structure formation during transcription. Journal of Biomolecular Structure and Dynamics 4, 1–9.

Mironov, A. A. and Lebedev, V. F., 1993. A kinetic model of RNA folding. Biosystems 30, 49–56.

Pan, T. and Sosnick, T. R., 2006. RNA folding during transcription. Annu. Rev. Biophys. Biomol. Struct. 35, 161–175.

Proctor, J. R. and Meyer, I. M., 2013. COFOLD: an RNA secondary structure prediction method that takes co-transcriptional folding into account. Nucleic Acids Res. 41, e102.

Rastegari, B. and Condon, A., 2007. Parsing nucleic acid pseudoknotted secondary structure: Algorithm and applications. Journal of Computational Biology 14, 16–32.

Reeder, J. and Giegerich, R., 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics 5.

Reidys, C. M., Huang, F. W. D., Andersen, J. E., Penner, R. C., Stadler, P. F., and Nebel, M. E., 2011. Topology and prediction of RNA pseudoknots. Bioinformatics 27, 1076–1085.

Rivas, E. and Eddy, S. R., 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. 285, 2053–2068.

Storz, G., 2002. An expanding universe of noncoding RNAs. Science 296, 1260–1263.

Taufer, M., Licon, A., Araiza, R., Mireles, D., van Batenburg, F. H. D., Gultyaev, A. P., and Leung, M.-Y., 2008. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. Nucleic Acids Research 37, D127–D135.

Thanh, V. H., Priami, C., and Zunino, R., 2014. Efficient rejection-based simulation of biochemical reactions with stochastic noise and delays. J. Chem. Phys. 141, 10B602.

Thanh, V. H. and Zunino, R., 2014. Adaptive tree-based search for stochastic simulation algorithm. Int. J. Comput. Biol. Drug. Des. 74, 341–357.

Thanh, V. H., Zunino, R., and Priami, C., 2016. Efficient constant-time complexity algorithm for stochastic simulation of large reaction networks. IEEE/ACM Trans. Comput. Biol. Bioinform. 14, 657–667.

Thanh, V. H., Zunino, R., and Priami, C., 2017. Efficient stochastic simulation of biochemical reactions with noise and delays. J. Chem. Phys. 146, 084107.

Turner, D. H. and Mathews, D. H., 2009. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res. 38, D280–D282.

Watters, K. E., Strobel, E. J., Yu, A. M., Lis, J. T., and Lucks, J. B., 2016. Cotranscriptional folding of a riboswitch at nucleotide resolution. Nat. Struct. Mol. Biol. 23, 1124–1131.

Zhao, P., Zhang, W., and Chen, S.-J., 2011. Cotranscriptional folding kinetics of ribonucleic acid secondary structure. J. Chem. Phys 135, 245101.

Zuker, M., 1989. On finding all suboptimal foldings of an RNA molecule. Science 244, 48–52.

Zuker, M. and Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 9, 133–148.

List of Figures

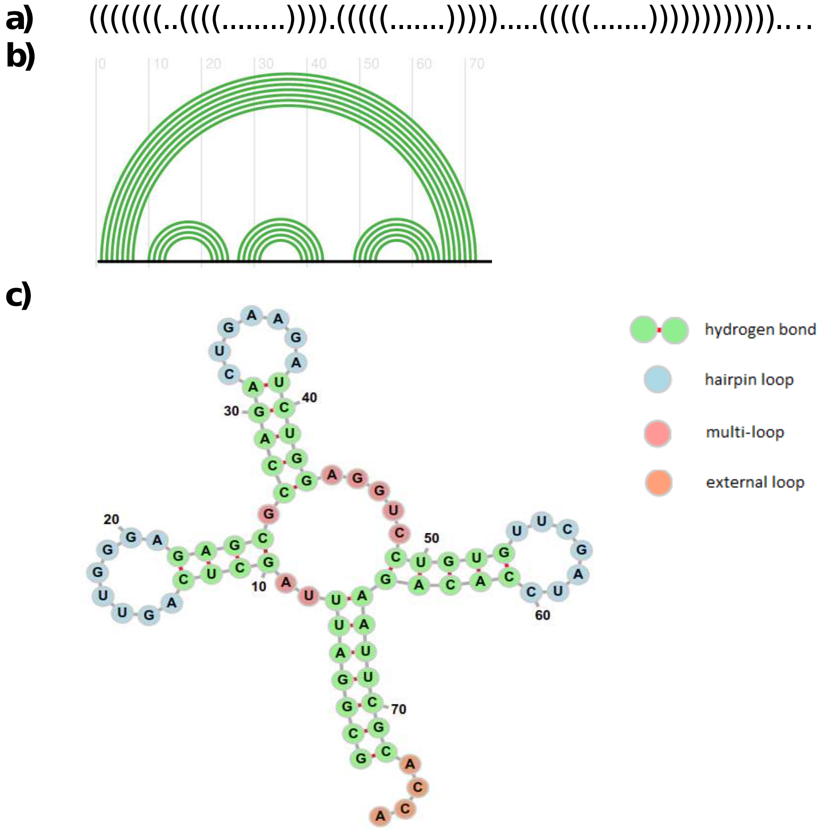


Figure 1: Representation of the tRNA molecule in a) dot-bracket notation, b) arc diagram and c) graphical visualization. The graphical visualization is made by the Forna tool Kerpedjiev et al. (2015)

.

⇑

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 1 (of 14)

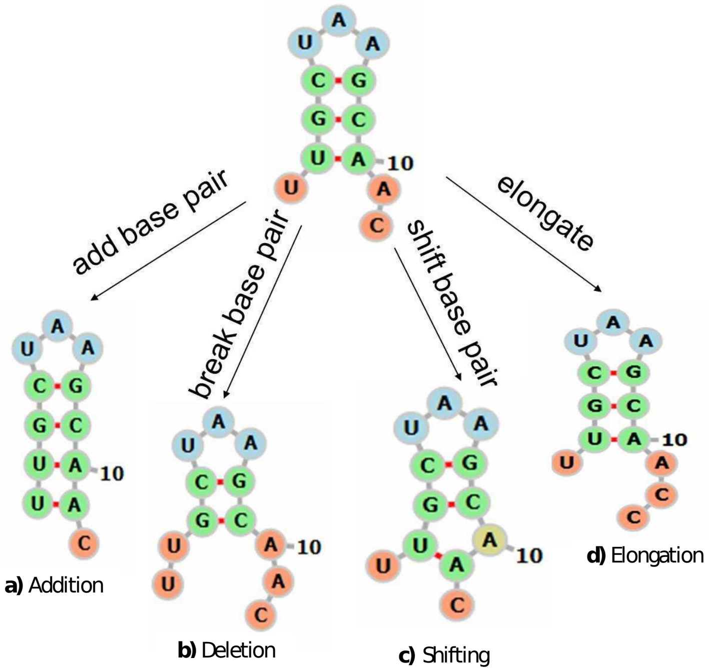


Figure 2: Extended move set consisting of a) addition, b) deletion, c) shifting and d) elongation. The elongation move models the transcription process extending the current RNA chain with a new nucleotide at the 3' end.

$\Uparrow$

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 2 (of 14)

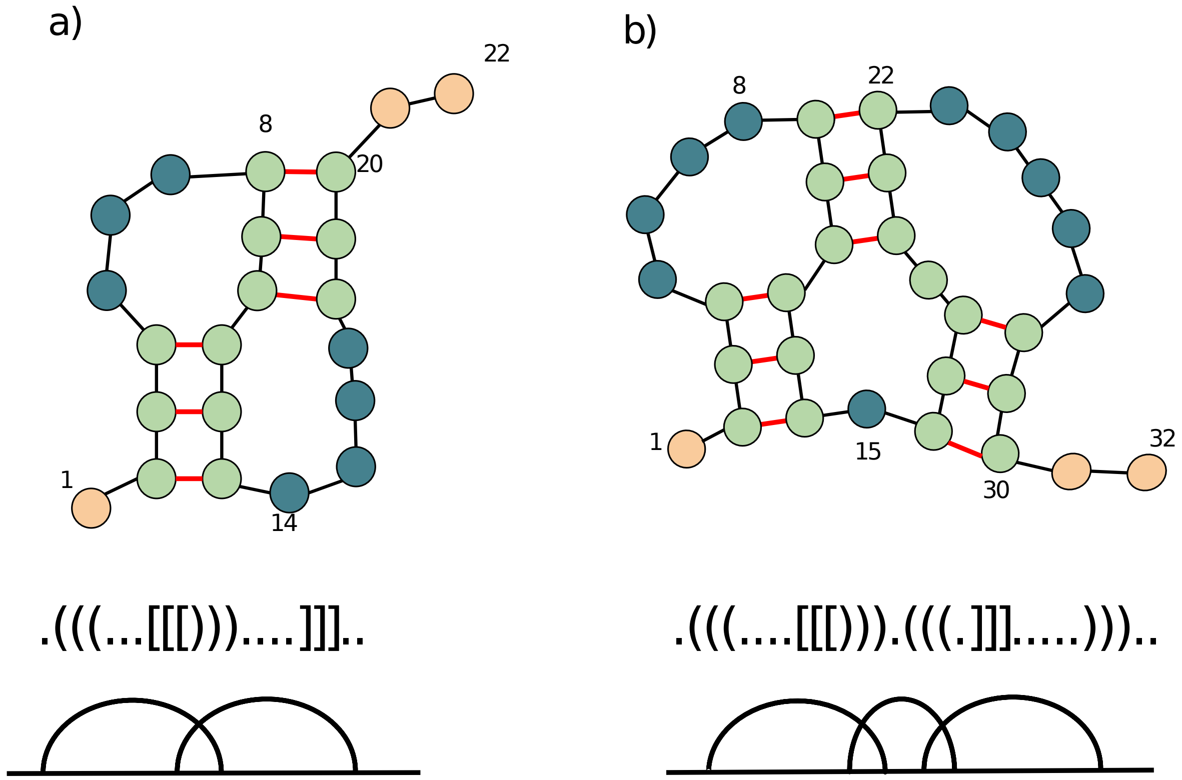


Figure 3: a) H-type pseudoknot and b) K-type pseudoknot depicted with extended dot-bracket notation and arc diagrams.

$\Uparrow$

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 3 (of 14)

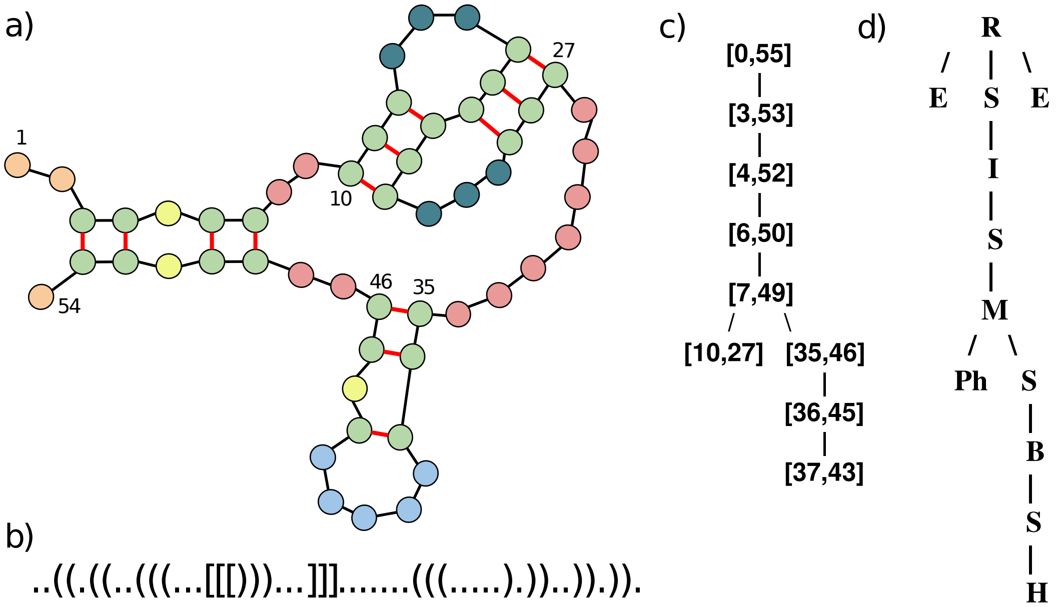


Figure 4: An example of a motif tree. a) Secondary structure with pseudoknot, b) its extended dot-bracket form, c) closed region tree, and d) motif tree. Starting from the root R (a dummy node) the motif tree represents the relationship of loops: exterior (E), stem (S), hairpin (H), multi-branch (M), pseudoknot (Ph), bulge (B) in the structure.

⇑

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 4 (of 14)

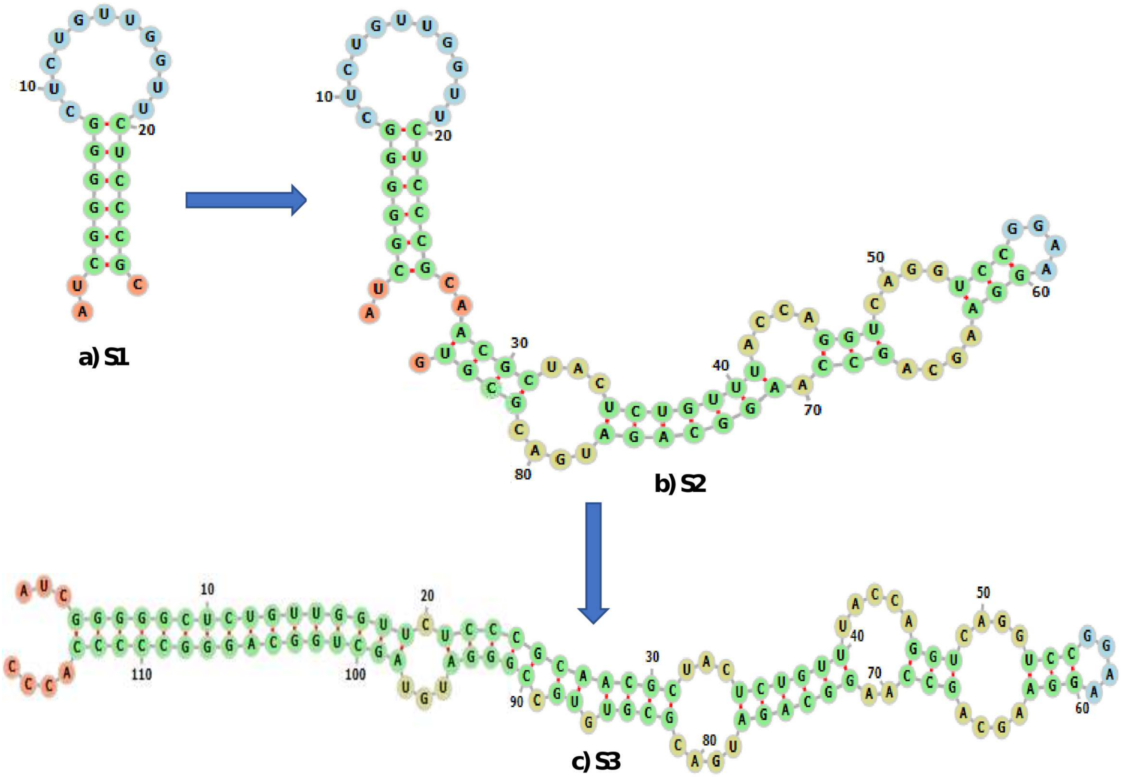


Figure 5: The folding pathway of secondary structures of the E. coli signaling recognition particle (SRP) RNA. The hairpin motif S1 is formed at transcript length 25nt and form S2 completed at length 86nt. When reaching transcript length 117nt, SRP rearranges into its stable helical shape S3. The visualization of structures is made by the Forna tool Kerpedjiev et al. (2015).

⇑

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 5 (of 14)

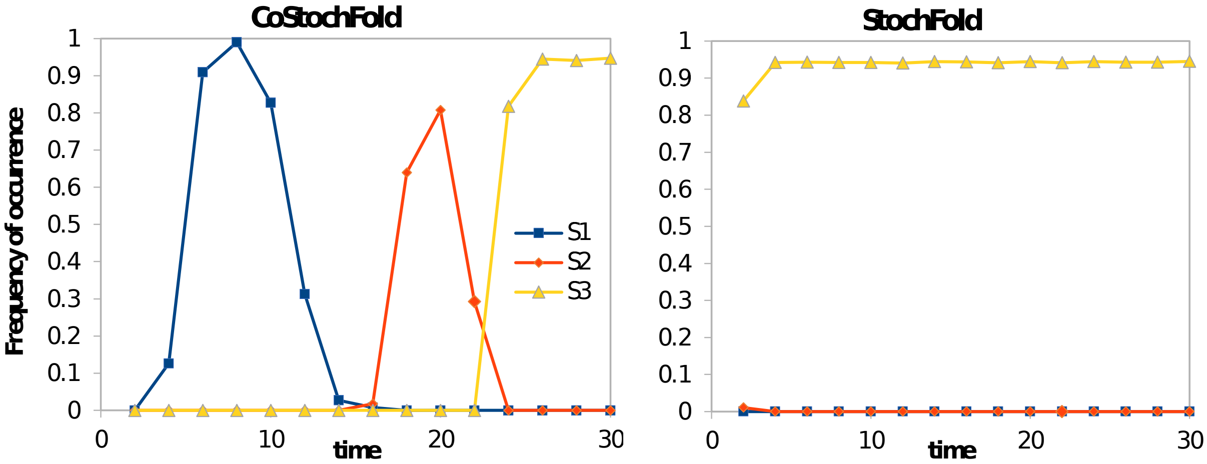


Figure 6: Prediction of the structural formation of SRP. Left: cotranscriptional folding. Right: folding from denatured state without transcription. The frequency of occurrence of a motif on the y-axis is computed as the numbers of occurrences over total 10000 simulation runs. Time on the x-axis is in seconds of simulated time.

$\Uparrow$

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 6 (of 14)

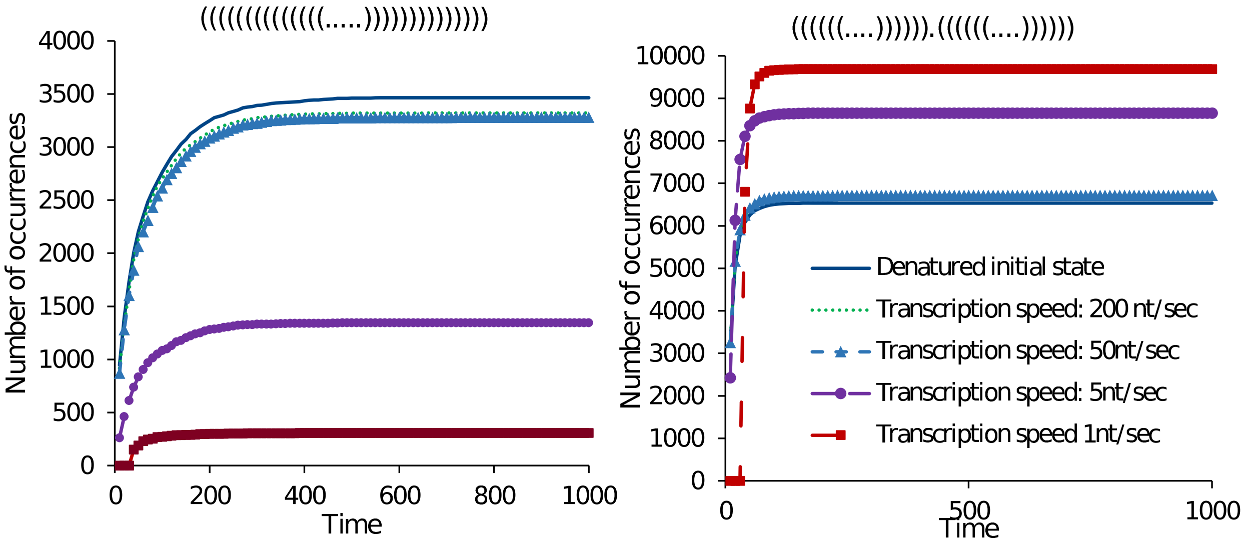


Figure 7: Cumulative first-hitting time occurrences of MFE structure x = “((((((((((((((.....))))))))))))))” (−26.20 kcal, left) and suboptimal y = “((((((....)))))).((((((....))))))” (−25.30 kcal, right). Time on the x-axis is in seconds of simulated time.

⇑

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 7 (of 14)

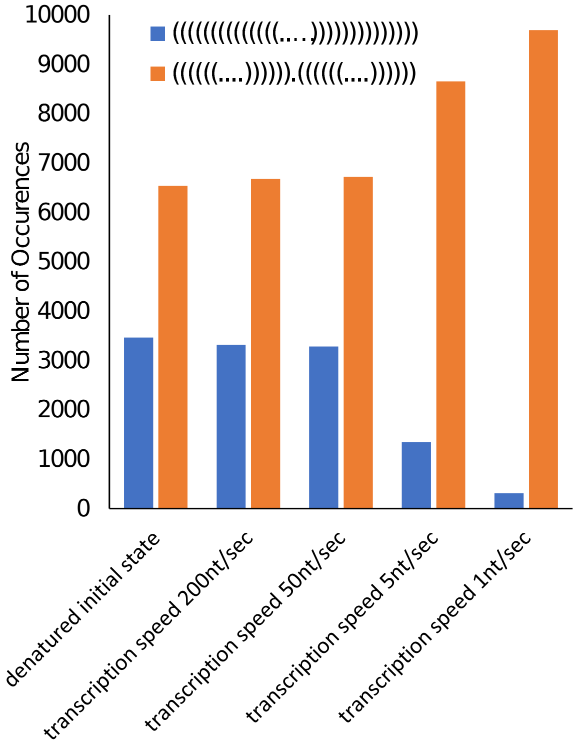


Figure 8: Total number of occurrences of MFE structure $x =$ ”((((((((((((((.....))))))))))))))” (−26.20 kcal) and suboptimal $y =$ ”((((((....)))))).((((((....))))))” (−25.30 kcal) with simulated time $T_{max} =$ 1000 seconds by varying transcription speeds.

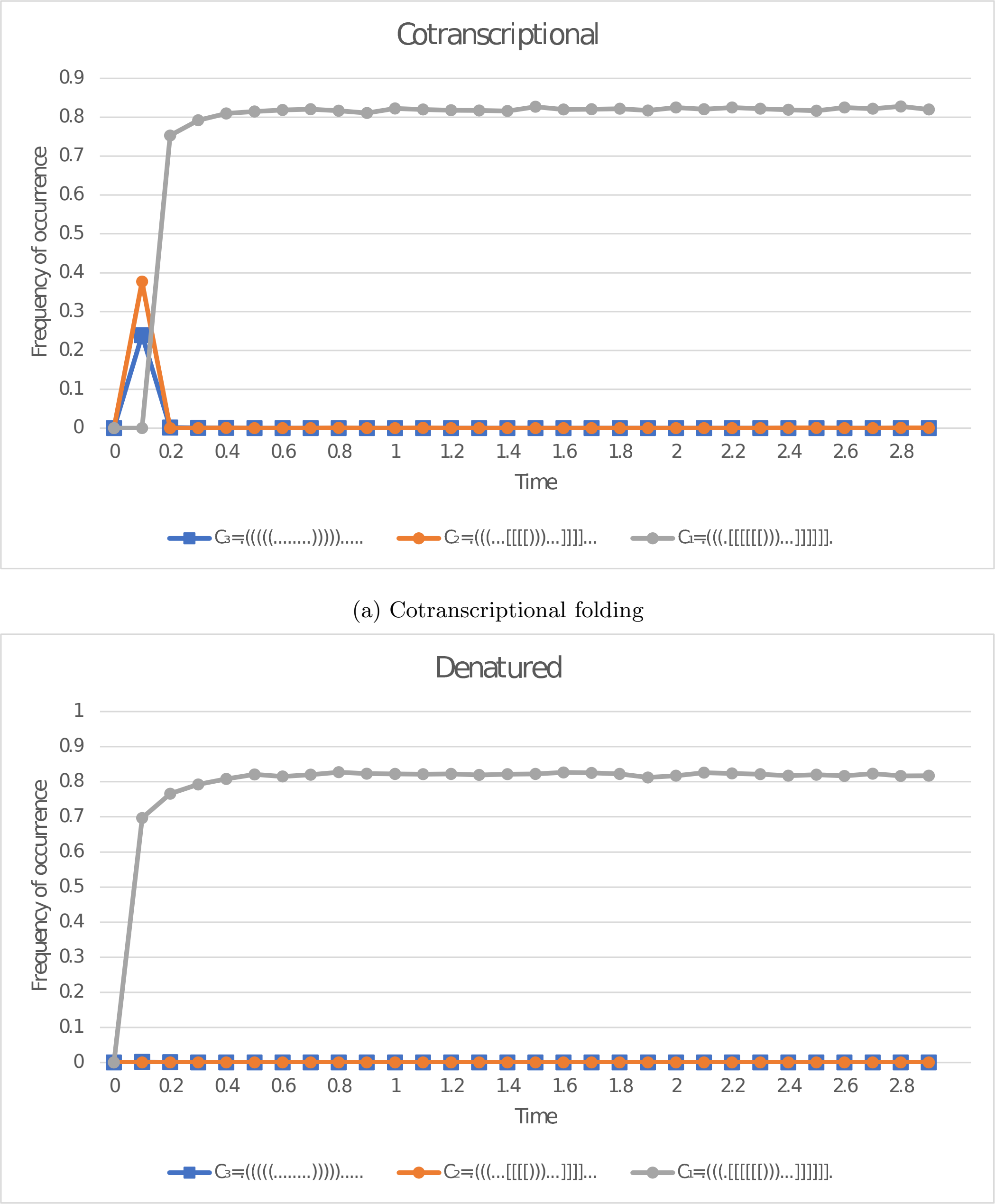
$\Uparrow$

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 8 (of 14)



(a) Cotranscriptional folding

(b) Folding from denatured state

Figure 9: Structural formation of the Beet soil-borne virus in a) Cotranscription folding and b) Folding from denatured initial state. Time on the x-axis is in seconds of simulated time.

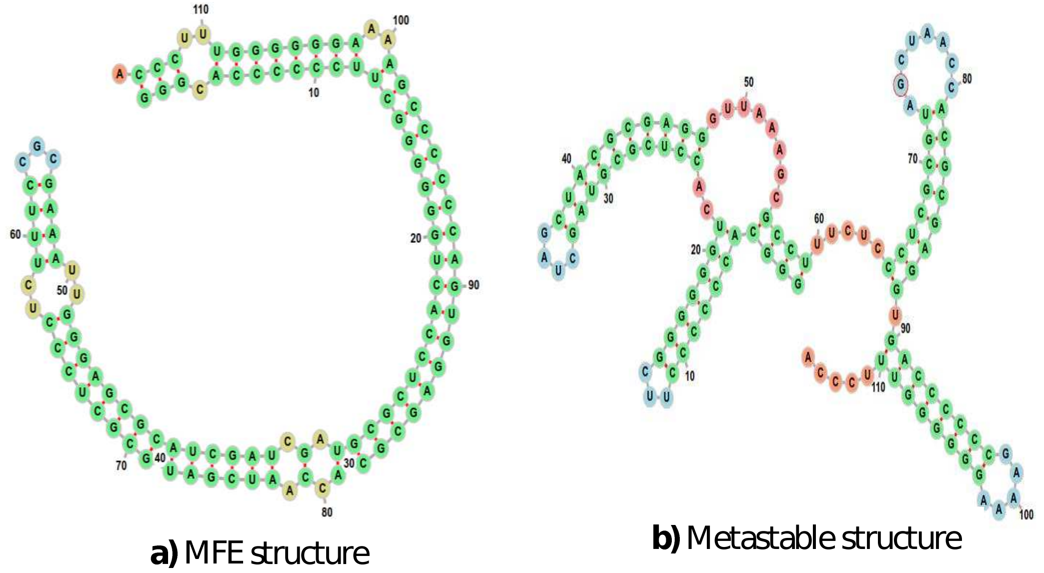
$\Uparrow$

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 9 (of 14)



a) MFE structure

b) Metastable structure

Figure 10: SV-11 with two conformations a) MFE structure (−95.90 kcal) and b) metastable structure (−63.60 kcal). The visualization of structures is made by the Forna tool Kerpedjiev et al. (2015).

⇑

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 10 (of 14)

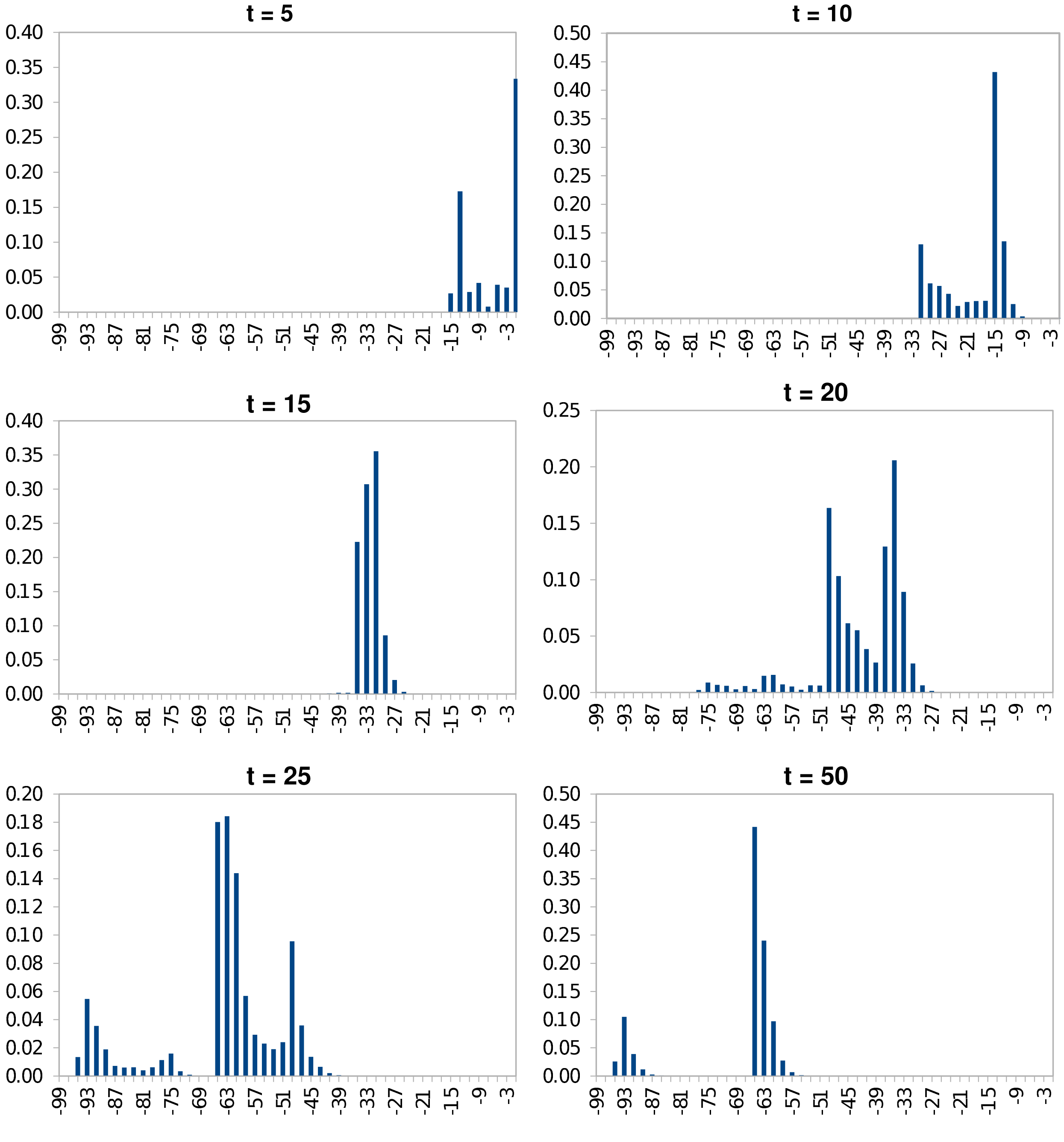


Figure 11: Cotranscriptional folding of SV-11. The x-axis denotes the energy level in kcal, and y-axis shows the frequency of structures at a given energy level.

$\Uparrow$

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 11 (of 14)

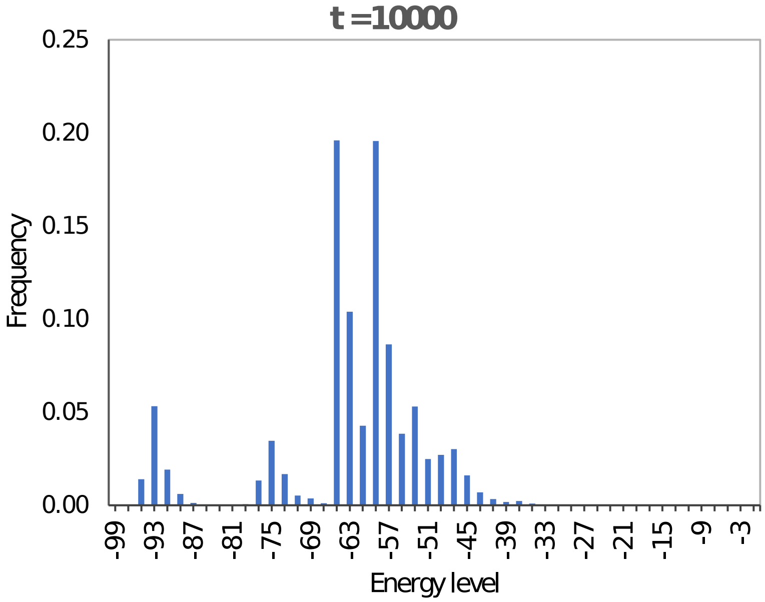


Figure 12: Frequency of structures in folding SV-11.

⇑

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 12 (of 14)

Metastable: (((((((((((...)))))))..((((((((((....))))))))))........)))).....((((((((........)))))))).((((((((.....)))))))).....

MFE: (((.(((((((((((((((((((((((((((.(.((((((((((((((..((((...))))..)))))))))))))).).))))))))))))))))))).))))))))..))).

Suboptimal 1: (((.(((((((((((((((((((((((((((...((((((((((((((..((((...))))..))))))))))))))...)))))))))))))))))))..))))))))..))).

Suboptimal 2: (((.(((((((((((((((((((((((((((...((((((((((((((..((((...))))..))))))))))))))...))))))))))))))))))..)))))))))..))).

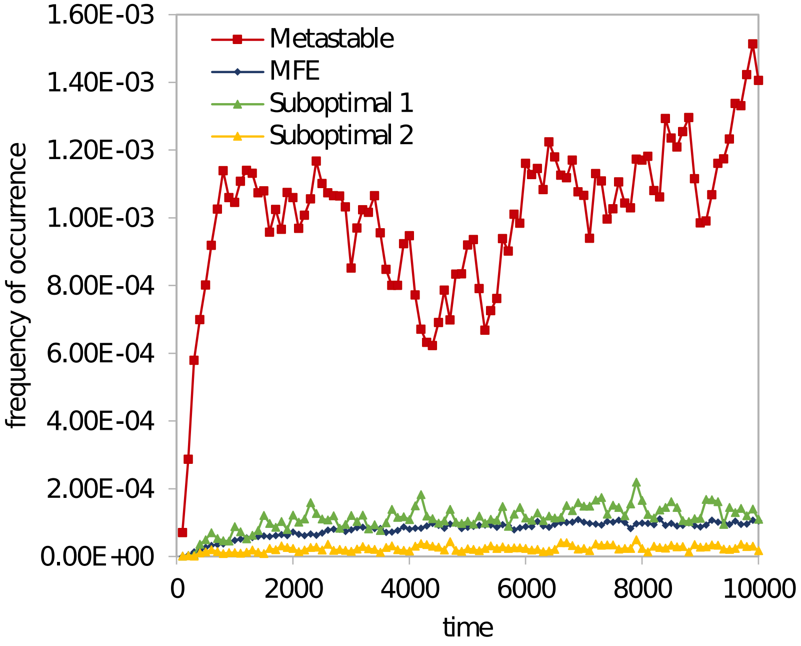


Figure 13: Frequency of the metastable structure in comparison with the MFE structure and two randomly selected suboptimal structures in the locality of the energy level of MFE. Time on the x-axis is in seconds of simulated time.

⇑

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 13 (of 14)

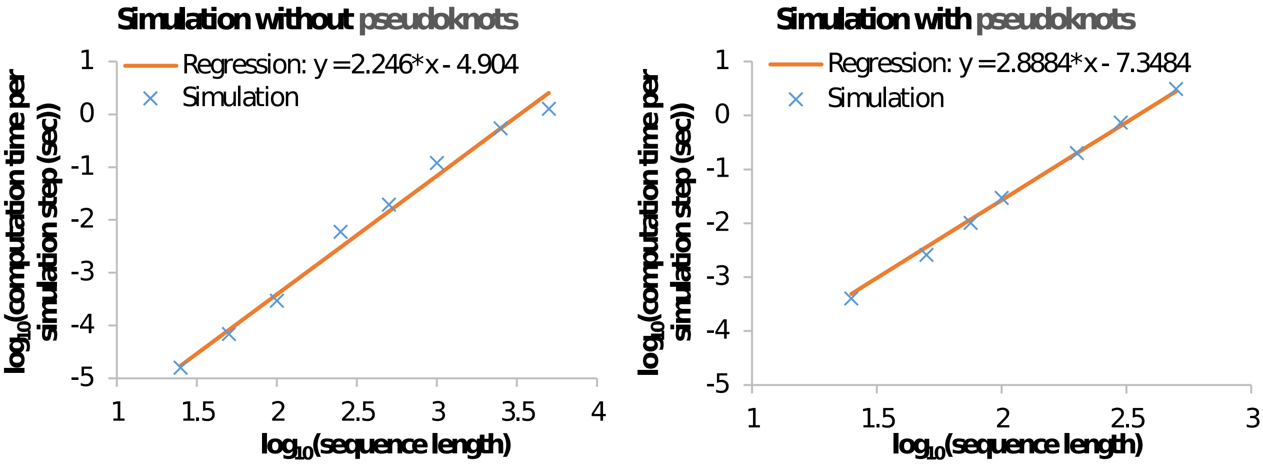


Figure 14: Computational runtimes of stochastic folding with sequences of varying lengths. Left: simulation without pseudoknots. Right: simulation with pseudoknots. Values on the x-axis and y-axis are in logarithmic scale.

$\Uparrow$

Vo Hong Thanh[1,2], Dani Korpela[1], Pekka Orponen[1]

[1]Department of Computer Science, Aalto University

[2] Certara, Simcyp Division

Figure 14 (of 14)