Cotranscriptional folding kinetics of ribonucleic acid secondary structures

Peinan Zhao, Wenbing Zhang, and Shi-Jie Chen

Citation: J. Chem. Phys. **135**, 245101 (2011); doi: 10.1063/1.3671644
View online: http://dx.doi.org/10.1063/1.3671644
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v135/i24
Published by the American Institute of Physics.

Cotranscriptional folding kinetics of ribonucleic acid secondary structures

Peinan Zhao,[1] Wenbing Zhang,[1,a)] and Shi-Jie Chen[2]
[1]*Department of Physics, Wuhan University, Wuhan 430072, People's Republic of China*
[2]*Department of Physics and Astronomy and Department of Biochemistry University of Missouri, Columbia, Missouri 65211, USA*

(Received 12 August 2011; accepted 2 December 2011; published online 22 December 2011)

We develop a systematic helix-based computational method to predict RNA folding kinetics during transcription. In our method, the transcription is modeled as stepwise process, where each step is the transcription of a nucleotide. For each step, the kinetics algorithm predicts the population kinetics, transition pathways, folding intermediates, and the transcriptional folding products. The folding pathways, rate constants, and the conformational populations for cotranscription folding show contrastingly different features than the refolding kinetics for a fully transcribed chain. The competition between the transcription speed and rate constants for the transitions between the different nascent structures determines the RNA folding pathway and the end product of folding. For example, fast transcription favors the formation of branch-like structures than rod-like structures and chain elongation in the folding process may reduce the probability of the formation of misfolded structures. Furthermore, good theory-experiment agreements suggest that our method may provide a reliable tool for quantitative prediction for cotranscriptional RNA folding, including the kinetics for the population distribution for the whole conformational ensemble. © *2011 American Institute of Physics*. [doi:10.1063/1.3671644]

I. INTRODUCTION

One of the important problems in gene regulation is how mRNA structure controls transcription and translation.[1–3] The folding of functional RNA structures are often coupled with the transcription process.[4–6] For instance, in the auto-catalyzed splicing reaction of *tetrahymena* group I intron, the functional native structure may form within the timescale of transcription, which is much faster than the refolding of the complete chain *in vitro*.[7–12] And in the cases of riboswitches, by adding or removing ligand, the RNA can form distinct structures which determine whether the polymerase promotes or terminates the transcription.[13–18] It has been proposed that natural RNAs can effectively avoid the formation of misfolded structures during the cotranscriptional folding process,[19,20] but the mechanism is still not completely clear. Experimental findings indicate that the transcription speed, pausing, and RNA-protein interactions are three main factors that can greatly influence how the RNA structure is formed during transcription.[21–24] A recent experiment with designed RNAs (Ref. 25) showed that transient intra- and inter-molecular base pair interactions can effectively regulate the folding of nascent RNA molecules through the formation of the different structures. A clear and detailed understanding of the kinetic process of RNA folding during transcription is crucial for uncovering the mechanism of RNA functions in gene regulation and the design and artificial control of the folding pathways, folding products and their population distributions at the end of the transcription.

Monte Carlo simulations on cotranscriptional RNA folding has led to many insights into folding mechanisms,[26,27] although direct interpretation of the kinetic timescales from the traditional Monte Carlo move sets can be convoluted. The (real) time scale of the synthesis of a nucleotide could greatly influence the folding production and population distribution of different conformations. Recently, several RNA folding kinetics algorithms were developed in connection with the thermodynamic energetics of the folding system. For instance, coarse-grained landscapes in conjunction with a stochastic sampling algorithm was used to study the RNA folding kinetics.[28] By using barrier trees and assuming that the basins of individual local minima are in quasi-equilibrium, the folding kinetics under transcription was studied.[29] Combining the thermodynamic properties with coarse-grained local folding kinetics, a heuristic approach was also developed to successfully predict cotranscriptional folding for large RNAs.[30] It is important to note that the local equilibration is determined by the competition between the transition rate between the states and the transcription speed, so the local minima may not always reach quasi-equilibrium. Motivated by the importance to have a direct account of the (real) time scale, we develop an analytical (non-simulational) method to predict and analyze the cotranscriptional folding kinetics. We focus on the sequential folding scenarios. In the theory, the elongation process is divided into many steps, each corresponding to the elongation of the chain by one nucleotide. Based on our recently developed RNA folding kinetic theory,[31–33] we establish a systematic algorithm to provide the conformational ensemble, the transition pathways, and the population distribution in every single transcriptional step. By integrating these steps, we predict the complete

a)Author to whom correspondence should be addressed. Electronic mail: wbzhang@whu.edu.cn.

scenario of the RNA folding during the entire transcription process.

II. THEORY AND METHOD

The Theory and Method section is arranged as follows: we will first present the folding kinetics theory including the master equation method and the rate model for the elementary kinetic steps. We will then focus on the method for predicting cotranscription folding.

A. Master equation

Consider an ensemble of conformational states. The population $p_i(t)$ for each state i at time t can be described by the following equation (master equation): $dp_i/dt = \sum[k_{j\to i}p_j(t) - k_{i\to j}p_i(t)]$, where $k_{j\to i}$ and $k_{i\to j}$ are the rate of the respective transitions, Ω is the total number of conformations. The above rate equation can be written as the matrix form: $d\mathbf{p}/dt = \mathbf{M}\cdot\mathbf{p}$, where $\mathbf{p}$ is the vector for the population distribution, $\mathbf{M}$ is the $\Omega\times\Omega$ rate matrix with the matrix elements defined by $M_{ij} = k_{j\to i}$ $(i\neq j)$ and $M_{ii} = -\sum_{j\neq i} k_{i\to j}$. The population kinetics is given by the eigenvalue spectrum:

$$p(t) = \sum_{m=1}^{\Omega} C_m n_m e^{-\lambda_m t}, \tag{1}$$

where $-\lambda_m$ and n_m are the mth eigenvalue and eigenvector of the rate matrix M, and C_m is the coefficient as determined by the initial condition.

B. Helices as building blocks

A fundamental process in RNA folding is to form a helix, which consists of consecutive base stacks. In our model,[31] we assume the following rate model for the formation and disruption of a base stack: $k_+ = k_0 e^{-\Delta S_{stack}/k_B T}$, $k_- = k_0 e^{-\Delta H/k_B T}$, where $-\Delta S_{stack}$ and $-\Delta H$ are the entropy and enthalpy change upon the formation of the stack. Under the folding condition, the rate for the formation of a base stack is usually larger than that of disrupting the stack. Hence, once the first few stacks in a helix are closed and stabilized, zipping of the subsequent stacks in the helix can be fast (10–100 μs). Except the initial (nucleation) states, any other partially formed helical intermediates would quickly slip into the fully folded helix through fast pathways such as branch migration or helix-helix exchange.[33] This suggests the use of helices as building blocks for the study of the overall (slower) folding kinetics. Using helices as the building blocks can lead to a drastic decrease in the conformational space for the kinetics calculation. In the following, we will first illustrate our rate theory for the creation/annihilation of a helix and the conversion between two helices.

C. Formation and disruption of a helix

If two conformations differ only in one helix, the transition between them would be the formation and disruption of the helix. We assume that after the first stack is closed

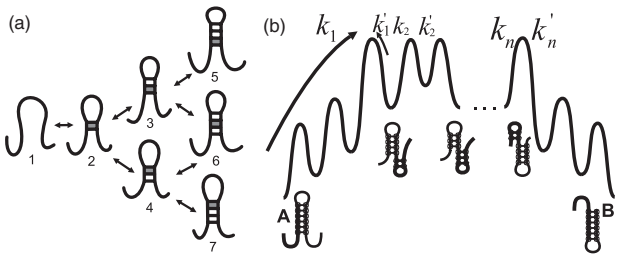

FIG. 1. (a) The zipping pathways for the formation a helix after the first (loop-closing) stack is formed. After structure 2 (state 2) is formed, there exist two folding pathways: $1\to 2\to 3$ and $1\to 2\to 4$. (b) The free energy landscape of the tunneling pathway that connects two overlapping helices A and B. The unfolding of A is accompanied by the folding of B. k_1 denotes the transition rate for the unfolding of helix A to form the first stack of helix B. k_1', k_2, k_2', …, k_n denote the transition rates between the neighboring intermediates along the tunneling pathways.

(with the concurrent formation of a loop), the helix will form along the zipping pathway.[23] We can estimate the rate of helix formation along this zipping pathway. After the first base stack formed, as shown in the Fig. 1, the zipping pathway is branched into two directions (corresponding to the two neighbors of the first base stack). From the empirical thermodynamic parameters,[34,35] we found that for most RNA helices, the free energy landscape for a zipping pathway shows a downhill profile after the formation of the third base stack. Therefore, the rate k_f of the helix formation (along a specific pathway) is equal to the rate for the formation of the three-stack state. Considering the (slow) breaking of the stacks, for zipping along the $1\to 2\to 3$ pathway in Fig. 1(a), we have[33]

$$\begin{aligned} k_f &= k_{12}K_1\left(1 - K_2'K_1'\sum_0^{\infty} K_2'K_1\right) \\ &= k_{12}K_1\left(1 - K_2'K_1'\frac{1}{1-K_2'K_1}\right), \end{aligned} \tag{2}$$

where k_{ij} denotes the rate for the transition from state i to state j, K_1 and K'_1 are the forward (state $2\to 3$) and reverse (state $2\to 1$) probability for state 2, K_2 and K'_2 are the forward (state $3\to 5$ and $3\to 6$) and reverse (state $3\to 2$) probability for state 3,

$$\begin{aligned} K_1 &= \frac{k_{23}}{k_{23}+k_{21}+k_{24}},\ K_1' = \frac{k_{21}}{k_{23}+k_{21}+k_{24}}, \\ K_2 &= \frac{k_{35}+k_{36}}{k_{32}+k_{35}+k_{36}},\ K_2' = \frac{k_{32}}{k_{35}+k_{36}+k_{32}}. \end{aligned} \tag{3}$$

For a given RNA molecule, the first base stack can be formed anywhere inside the helix. Therefore, the net rate k_F for the formation of a helix is the sum of the rates (Eq. (2)) along the two pathways (Fig. 1(a)) with the different first (nucleation) base stacks. The rate for the disruption of the helix can be estimated from the equilibrium constant of the helix: $k_U = k_F e^{\Delta G/k_B T}$, where ΔG is the folding free energy of the helix.

D. Exchange between two helices

If two helices A and B overlap with each other, they cannot coexist in the same structure. The conversion of helix A

to helix B through complete unfolding of helix A followed by refolding to B is extremely slow due to the high energy barrier to disrupt all the base stacks in helix A. A much faster pathway is through stepwise pathway where at first helix A is partially disrupted, and in each subsequent step, disruption of a base stack in A is accompanied by a concurrent formation of a base stack in B (see Fig. 1(b)), which lowers the free energy. The pathway is fast because the formation of the base stacks in B tends to cause an overall downhill shape of the free energy landscape. This (tunneling) pathway involves a much lower energy barrier to unwinding the helix than the complete unfolding pathway. Based on the tunneling pathway, we can estimate the rate for helix exchange:[33]

$$k_{A\to B} = \frac{\prod_i^n k_i}{\sum_{j=0}^{n-1}\left(\prod_{i=1}^{j} k_i' \prod_{m=j+2}^{n} k_m\right)},$$

$$k_{B\to A} = k_{A\to B} e^{-\frac{\Delta G_{AB}}{k_B T}}. \quad (4)$$

In the above formula, k_n and k_n' are the rate constants for the process to formation (disruption) and disruption (formation) of a base stack in A (B), respectively.

E. Predicting cotranscriptional folding

We treat the transcription of a single nucleotide as an elementary time step. The real time for each step is a constant or variable if the nucleotides are synthesized at a constant or variable speed, respectively. For example, to simulate transcriptional pausing at a specific site, we can assign a large number of effective time steps for the corresponding (paused) step. In the present study, we assume a constant transcription speed. If the transcription speed of an RNA sequence is ν nucleotides per second, the (real) time window for each step would be $1/\nu$ s, i.e., the polymerase spends $1/\nu$ s to synthesis a nucleotide.

From time t when an l-nt chain is (newly) transcribed to time $t + 1/\nu$ when the $(l + 1)$th nucleotide is (newly) transcribed, the l-nt chain samples the conformational space and its population distribution is relaxed from $[P_1(l)_{begin}, P_2(l)_{begin}, \ldots, P_\Omega(l)_{begin}]$ to $[P_1(l)_{end}, P_2(l)_{end}, \ldots, P_\Omega(l)_{end}]$. Here Ω is the number of conformations for a l-nt chain. We call the above process as the lth step. For each such step, we generate the complete conformational ensemble of the chain, compute the transition rates between the different structures (see Eqs. (2) and (4)) and calculate the population kinetics from time t to $t + 1/\nu$(see Eq. (1)). The population distribution at the end of the lth step is dependent on the initial population distribution $[P_1(l)_{begin}, P_2(l)_{begin}, \ldots, P_\Omega(l)_{begin}]$ at the beginning of the lth step and the time duration ($\sim 1/\nu$) for the synthesis of the $(l + 1)$th nucleotide.

The beginning population of the $(l + 1)$th step is inherited from the ending population of the lth step. However, the RNA chain in the $(l + 1)$th step is one nucleotide longer than in the lth step. How to determine the beginning population for the $(l + 1)$th step? According to the possible changes of the structures upon the elongation of the chain by one nucleotide, we classify four types of structures.

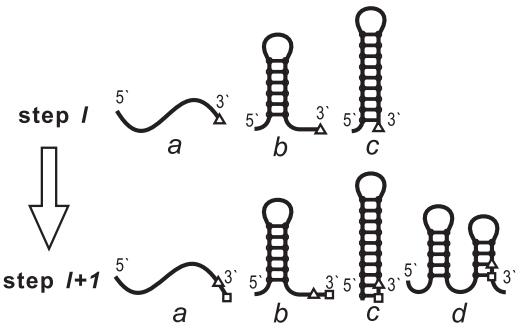

FIG. 2. Four types of relationships between l-nt and $(l + 1)$-nt structures: elongation of an open chain (a), a dangling tail (b), a helix (c), and the formation of a new structure (d). The triangle denotes the last transcribed nucleotide in step l and the square denotes the last transcribed nucleotide in step $l + 1$.

As shown in Fig. 2, for type *a* and type *b,* the newly transcribed nucleotide is added to the 3′ end of an open chain and the dangling tail of a helix, respectively. In these two categories, the addition of the new nucleotide will not cause change of the existing structure so the new $(l + 1)$-nt chain can retain the same secondary structure as the l-nt chain. For category *c,* however, the newly transcribed nucleotide can pair with an upstream nucleotide to form an elongated helix by one base pair. Since the zipping of the new base pair (stack) (with time $\sim$ (rate)$^{-1}$ $\sim 10^{-6}$ s) is much faster than transcribing a nucleotide ($1/\nu$ from 2.5×10^{-3} s to 5×10^{-2} s),[23] these two structures can be treated as "directly inherited" and thus have the same population. Category *d* denotes all the new $(l + 1)$-nt structures that cannot be formed in an l-nt chain and thus have zero population at the beginning of the $(l + 1)$th step. The population distribution at the beginning of step $l + 1$ can be summarized by the equations below:

$$P(l+1)_{begin} = P(l)_{end} \text{ for } a,\ b,\ \text{and } c,$$
$$P(l+1)_{begin} = 0 \text{ for } d.$$

For consecutive steps, the folding results of the previous step turn into the initial condition of the next step. Conformations for the chain of the $(l + 1)$th step can be calculated from conformations of the lth step. To construct the $(l + 1)$-nt conformations, we enumerate all the possible new helices in which the newly transcribed $(l + 1)$th nucleotide base pairs with other already transcribed nucleotide. In particular, if such a new helix is compatible with the existing helices of the preceding l-nt chain, we can generate an $(l + 1)$-nt conformation by simply adding the new helix to the existing (l-nt) conformations. Applying this method from the first step to the end of transcription, we compute the folding kinetics for the RNA chain during transcription. In the whole process, we need to solve about $(n - 7)$ master equations for an n-nt sequence (7-nt is the minimum foldable sequence length in the model). The total computer time is much less than $(n - 7)$ times the computer time required for an n-nt full chain, as the size of the rate matrix for an elongating chain is less than (or equal to at the last step) that of the full chain and the most time consuming part is for the conformational sampling and rate matrix calculation.

III. RESULTS

A. Competition of the different rate processes determines the cotranscriptional folding pathway

Kramer and Mills studied cotranscriptional formation of MDV-1 RNA secondary structures using Q-beta replicase.[36] Their results suggested that the structural reorganization of MDV-1 RNA can be divided into three stages: by the time the first 45 nucleotides are synthesized, the chain folds into a branched structure that contains two hairpins A and B (denoted as A + B); after over 60 nucleotides are synthesized, hairpin B dissociates and a new, longer (more stable) hairpin T is formed (denoted as A + T); when the full 72-nt sequence is synthesized, a more branched structure (denoted as A + B + C) is formed. The experimental data also suggested that the chain elongation speed can influence the formation of the temporary structures of the nascent RNA chain. In the experiment, electrophoresis was used to deduce the stable structures. Since the product strands were melted first to free them of their templates prior to electrophoresis, the experiment could not show directly that these structures actually form during transcription. The theoretical analysis here, however, can directly provide the population distributions for RNA structures formed at every chain elongation step, and the theory may provide the microscopic mechanism not directly accessible in the Kramer and Mills experiment. We note more recent experiments such as single molecular methods,[37,38] FRET,[39] hydroxyl radical footprinting[40] can provide direct measurements for the intermediate structures and can hence provide data for direct theory-experiment comparisons.

We assume the transcription speed of 20 nucleotides per minute. From Fig. 3, which shows the fractional populations of all the populated structures (cut-off population: 1%) as the chain grows, we find transitions between a series of discrete intermediates of the nascent chain. As the first 21 nucleotides are synthesized, hairpin structure A is quickly formed, as the 34th nucleotide is synthesized, the population of the structure A + D begins to increase. Almost all the population is occupied by the branched structure A + D as the 37th nucleotide is synthesized. The structure A + D persists for a short period of time (from step 37 to 39). As the 40th nucleotide is synthesized, the population of structure A + D begins to decrease and structure A + B emerges. The whole population is the sum of the population of the two structures, suggesting the conversion from hairpin D to hairpin B. The conversion stops when the 45th nucleotide is synthesized. The structure A + B remains for a long period of time from step 45 to step 58. From step 58, structure A + T is formed with the disruption of structure A + B. The conversion from B to T stops as the nucleotide 64 is synthesized and the chain folds into A + B + C. The experiment did not show the structure A + D. This may be caused by the fact that the excised fragments in the experimental analysis were 32, 45, 46, 52, 53, 61, 62, while A + D is formed from the 34th nucleotide to the 44th nucleotide.

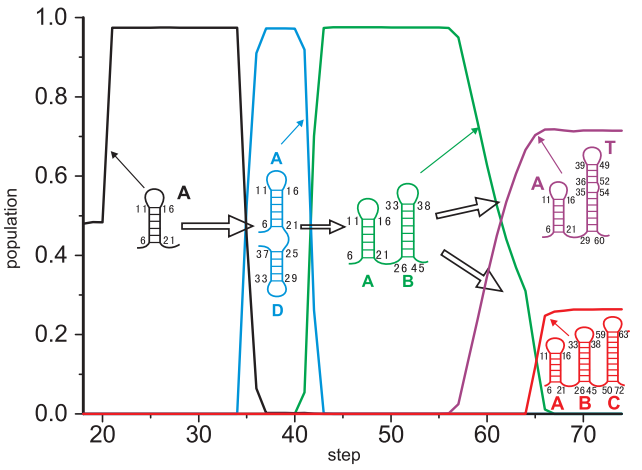

FIG. 3. The population kinetics with the formation and disruption of the different intermediate structures in the cotransciptional folding with a chain elongation speed of 20 nt/min. Strucuture A: black line, structure A + D: blue line, structure A + B: green line, structure A + T: purple line, structure A + B + C: red line.

Our kinetic theory gives the rate constants for the transitions between these different states (Table I) and may provide a microscopic mechanism for the structural rearrangement during the transcription. The formation of hairpin A is a zipping process with a high rate, so when the first 21 nucleotides were synthesized, the hairpin structure A would quickly form. No other structure could form till the 34th nucleotide is synthesized. As the 34th nucleotide is synthesized, a new base stack (with base pairs 28-34 and 29-33) is formed and zipping of the new hairpin D is initialized. The formation of helices A + D from A is a fast zipping process with a rate 2.43 $\times$ 10^4 s^{-1} (Table I), which is much faster than the transcription speed. So the formation of the A and D helices are rate-limited by the chain elongation. When the 37th nucleotide is synthesized, the fully hairpin D is formed. The nucleotide 39 stabilizes the loop-closing base stack (with base pairs 33-38 and 32-39) and initializes zipping of a new, more stable hairpin B through unwinding of helix D. The conversion from helix D to helix B is predicted to have a rate constant of 0.789 s^{-1} (Table I and Eq. (3)) for the tunneling pathway. So the A + D to A + B conversion occurs at step 39, and as the 45th nucleotide is synthesized, the hairpin D would be fully converted to B and the structure A + B occupies about 100% of the population.

Once nucleotide 50 is synthesized, zipping of hairpin T (Fig. 3) can occur. The notable increase of the A + T population emerges from step 58. This can be attributed to the higher free energy of T than B before step 58 and the slow rate for

TABLE I. The transition rates between the intermediate structures.

Transition	/ → A	/ → C	/ → D	D → B	B → T	A + T → A + B + C
Rate (s^{-1})	2.78×10^4	5.49×10^4	2.46×10^4	0.789	0.108	1.97×10^{-6}

the B to T transition, which involves closing the large hairpin loop (nts 39–49) and the disruption of the base stacks in helix B. We note that hairpins B and T cannot coexist, hence hairpin B must dissociate to release nucleotides for the formation of hairpin T. Our calculation shows a B to T transition rate of 0.108 s^{-1}. In zipping process of T, the free energy of the hairpin decreases. In step 58, the free energy of T (−12.40 kcal/mol) drops to that of B (−12.55 kcal/mol). After step 58, T outcompetes B and the population distribution shows a notable increase of the A + T population (Fig. 3). From nucleotide 58 to nucleotide 64 (about a time duration of 21 s), a significant fraction of population (71%) of the chain has been folded into the A + T structure when zipping of helix C is initialized. The conversion from B to T stops as the nucleotide 64 is synthesized and zipping of helix C is initialized. After step 65, the formation of C is a fast downhill process and the formation of T is overwhelmed by the zipping of C. As a result, the formation of A + T is stopped and it folds into A + B + C. The population of A + T and A + B + C is the result of the competition between three processes: transitions B to T, B to C, and the chain elongation.

The above results show that cotranscriptional folding of RNA may be kinetically controlled, namely, RNA conformations at the end of transcription has not yet reached thermal equilibrium. Instead, RNA conformational distribution is a result of the kinetics. From step 65 to the end of the transcription, the two structures A + T and A + B + C remain constant fractional populations of 71% and 26%, respectively. Although thermodynamically, A + B + C is much more stable than A + T($\Delta G_{A+B+C} = -40.97$ kcal/mol and $\Delta G_{A+T} = -29.66$ kcal/mol), the transition from A + T to A + B + C is extremely slow because the transition involves a high energy barrier corresponding to the disruption of a set of base pairs in T. Our calculation indicates a slow rate of 1.2 $\times 10^{-5}$ s^{-1} for the transition. For such a slow rate, direct interconversion can hardly occur in the timescale of transcription.

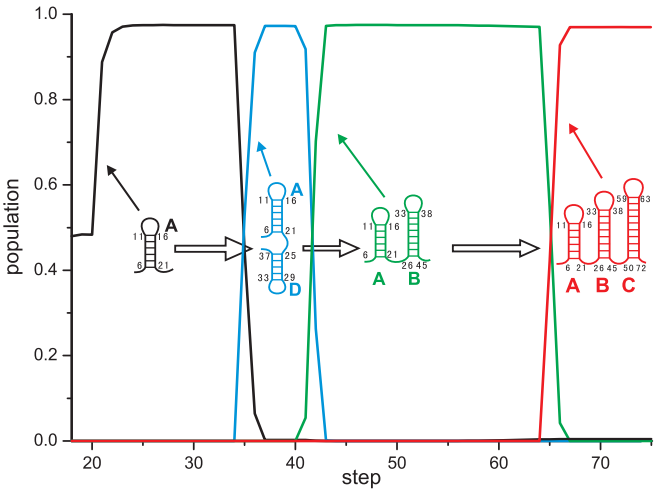

FIG. 4. The population kinetics with the formation and disruption of the different intermediate structures in the cotransciptional folding with a chain elongation speed of 100 nt/s. Strucuture A: black line, structure A + D: blue line, structure A + B: green line, structure A + B + C: red line.

B. Transcription speed affects the end product of transcription

To investigate the effect of the transcription speed, we calculated folding with different elongation speeds. For transcription with an increased speed of 100 nucleotides per second, as shown in Fig. 4, the folding process is from structure A to A + B and then to structure A + B + C. The structure A + T does not appear.

The dependence on the transcription speed of RNA folding is a result of the competition between the different timescales. From Table I for the predicted rate constants for the different transitions, we find that the conversions from A + D to A + B and from A + B to A + B + C are both very fast, while the conversion from A + B to A + T is much slower. The population conversion from A + B to A + T is given by $1 - \exp(-k_{A+B\to A+T}n/v)$, where $k_{A+B\to A+T}$ is the rate for the transition from A + B to A + T, v is the transcription speed, and n is the number of the nucleotides from nucleotide 58 to nucleotide 65 (the critical nucleotide for the initiation of the zipping of hairpin C). The transcription is so fast that nucleotide 64 is quickly transcribed and zipping of hairpin C starts before structure A + B is disrupted (to form A + T). Therefore, nearly no population of A + B would go to A + T, although A + T is thermodynamically much more stable than A + B ($\Delta G_{A+B} = -22.2$kcal/mol and $\Delta G_{A+T} = -29.66$kcal/mol). As a result, nearly 100% of the end product of the transcription is in the form of the branched structure A + B + C (Fig. 4).

In comparison with the refolding kinetics of the fully transcribed chain (Fig. 5), we find that a proper transcription speed can cause the mRNA chain to avoid the misfolded intermediates.

C. Transcription direction affects the folding pathway and folding products

To study the cotranscriptional folding paths, Xayaphoummine *et al.*[25] designed a pair of 73-nt RNA sequences that are exactly reversed: the "direct" sequence ("D"): 5′-GGAACCGUCUCCCUCUGCCAAAAGGUAGA GGGAGAUGGAGCAUCUCUCUACGAAGCAGAGAG AGACGAAGG-3′, and the "reverse" sequence ("R"). A reverse sequence with a mutation U38/C38 was also utilized to investigate the folding mechanism of nascent RNA chain. The transcription rate is about 200–400 nucleotides per second with T7 RNA polymerase. In the experiment, it was found that by the end of transcription, 100% of the "direct" sequence folds into the structure 1D (see Fig. 6(c)) and despite of the strong symmetry between the two sequences, only 10% of the "reverse" sequence folds into the structure 1R and 90% of the population forms structure 2R (see Fig. 6(c)) by the end of transcription. The experimental results can be related to stabilities of the different helices formed on the different transition pathways.[27,41] Our kinetic theory can provide a more detailed analysis for the microscopic mechanism including the population distribution and the rate constants of the RNA chain (see Fig. 6).

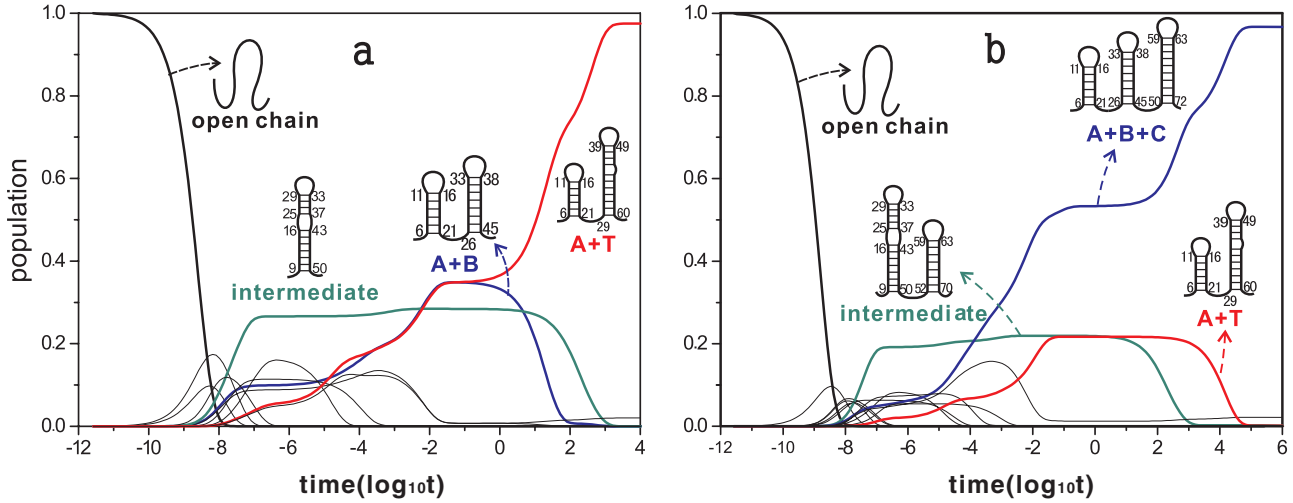

FIG. 5. The population folding kinetics for the refolding of the MDV-1 RNA when 62 nucleotides have been transcribed. Structure A + B: blue line, structure A + T: red line. (b) The population folding kinetics for the refolding of the fully transcribed MDV-1 RNA. Structure A + B + C: blue line, structure A + T: red line

The predicted population kinetics for the "direct" (see Fig. 6(a)) and the "reverse" sequence (Fig. 6(a)) are shown in Fig. 6 with RNA elongation speed of 250 nt/s. For the "direct" sequence, it shows that hairpin A is quickly folded when the first 28 nucleotides are transcribed. The hairpin is stable for a long lifetime till the 60th nucleotide is synthesized, which causes the formation of structures A + B1, A + B2, and A + B (structure 1D). Structures A + B1 and A + B2 persist for a short period of time before these structures are converted to structure A + B as the chain grows. By the end of the transcription, nearly 100% of the population goes to structure A + B (1D). For the reverse sequence (Fig. 6(b)), hairpins B, B1, and B2 are formed when the first 25 nucleotides are synthesized. Hairpins B1 and B2 persist for a short period of time before converting to hairpin B as the chain grows to 30 nt. Nearly all the population is converted to B when the chain grows to 35 nt. Hairpin B persists till the chain grows to 45 nt, by then hairpin B began to convert to hairpin D. As the 58th nucleotide was synthesized, about 93.6% of B converted to D, since then structures D + C (2R) and B + A (1R) appear, the sum of the population of D and D + C and the sum of B and B + A kept constant as 93.6% and 6.4%, respectively, denoting that structure D converted to structure D + C and B converted to B + A. As the chain elongated to 60 nt, the structure D + C (2R) and B + A (1R) occupied about 93.6% and 6.4% of the population, respectively, and they kept constant till the end of the transcription.

The theoretical results show good agreements with the experimental data, although the predicted A + B1 and A + B2 (for the direct sequence) and B1 and B2 (for the reverse sequence) were not reported in the experiments. This discrepancy may be due to the fact that these structures have a short lifetime and relatively low population during the transcription. The microscopic mechanism for the cotranscription folding for the direct and reverse sequence can be explained as follows.

For the "direct" sequence, by the time the first 38 nucleotides are transcribed, hairpin A is the most stable structure for the nascent chain and its formation is a zipping process with a fast rate. As nucleotide 42 is transcribed, the loop-closing base stack for hairpin D (base pairs 37-41 and 36-42) can be formed through the disruption of the terminal base pairs 5-38 and 6-37 in hairpin A (with the tunneling pathway for helix-helix exchange as discussed above). However, because helix A is more stable than helix D ($\Delta G_A = -28.55$ kcal/mol and $\Delta G_D = -18.09$ kcal/mol), the A to D uphill

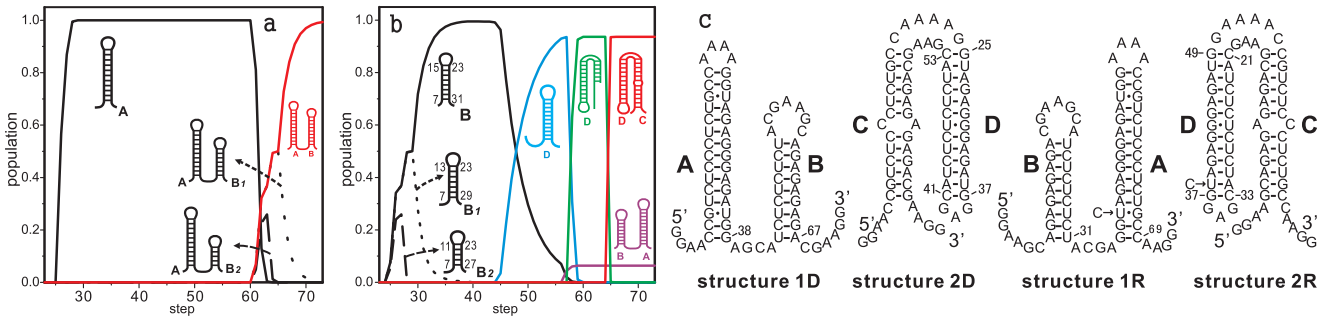

FIG. 6. (a) The population kinetics of the "direct" sequence at an elongation speed of 250 nt/s. The dashed and the dotted lines represent the populations for two intermediate structures A+B_1 and A+B_2, respectively. (b) The population kinetics of the "reverse" sequence at an elongation speed of 250 nt/s. (c) The conformational switch during the transcription of the "direct" sequence (1D and 2D) and the "reverse" sequence (1R and 2R).

transition can hardly occur. While helix B, B1, and B2 are compatible with A, and zipping of these helices is very fast once the first few loop-closing base pairs are stabilized. By the time the 60th nucleotide is transcribed, the first loop-closing base stacks (base pairs 51-59 and 50-60 for helix B, 49-59 and 48-60 for helix B1, 47-59 and 46-60 for helix B2) can be formed and the subsequent zipping of the helix would be fast. Zipping of the helices B1 and B2 would stop as the 65th and 63th nucleotides are transcribed, respectively, because further transcribed nucleotides cannot form base pairs for further helix elongation. We note that helix B can continue to grow. As a result, helices B1 and B2 would be converted to helix B as the chain grows and nearly 100% of the population is in the form of A + B (structure 1D), the native structure for the fully transcribed chain.

For the "reverse" sequence, compared to the "direct" sequence, helix B instead of A is first formed. As the 24th nucleotide is transcribed, the loop-closing base stacks for helix B1 (base pair 13-23 and 12-24), helix B2 (base pair 11-23 and 10-24) and helix B (base pair 15-23 and 14-24) can be formed. Zipping of these helices is very fast. B1 and B2 would convert to B as the chain grows due to the same reason as for the direct sequence. As shown in our predicted population kinetics (Fig. 6(b)), the nascent chain keeps folding into the hairpin B and finally occupied a 100% population until about nucleotide 44 is transcribed. From step 44 to step 57 (time duration $\sim$ 0.05 s), hairpin B is gradually converted to the hairpin D through the aforementioned tunneling pathway for helix-helix exchange (see Fig. 7). In the B to D tunneling pathway, the reaction is initiated by the formation of the loop-closing base stack (base pairs 33-37 and 32-36). Further zipping of the helix D is accompanied by the disruption of the corresponding base pairs in helix B. The free energy of D is higher than that of B at the beginning because of fewer base pairs than B. With the chain growth, the free energy of D is decreased as more base pairs in D are formed. Up till step 45, the free energy of D (-10.45 kcal/mol) will drop to the same level of the free energy of B (-11.87kcal/mol), and there are notable population transited from B to D. As shown in Fig. 7(b) for the energy profile along the pathway, in each step, the free energy barrier for the disruption of a base pair in B is offset by the free energy decrease for the formation of a base pair in D. Through such a helix-helix exchange "tunneling" pathway, the free energy barrier is drastically lower than the barrier for the complete unfolding of the whole helix B. The transition rates from B to D and from D to B are 62.4 s^{-1} and 0.0027 s^{-1}, respectively. Such helix-helix exchange pathway is crucial for the conformational switches in the cotranscriptional folding. Without such tunneling pathway, structural rearrangement could hardly occur in the transcriptional timescale.

By the time the "reverse" sequence is 49 nucleotides long, a fully zipped helix D can be formed. After step 58, hairpin D can continue the zipping process to form structure 2R by adding helix C. In the meantime, the chain is long enough to initialize the folding of helix A. The residual population of hairpin B can fold into structure 1R through the (fast) formation of helix A. These two parallel transitions are both very fast. Because the initiation of the B to D transition occurs prior

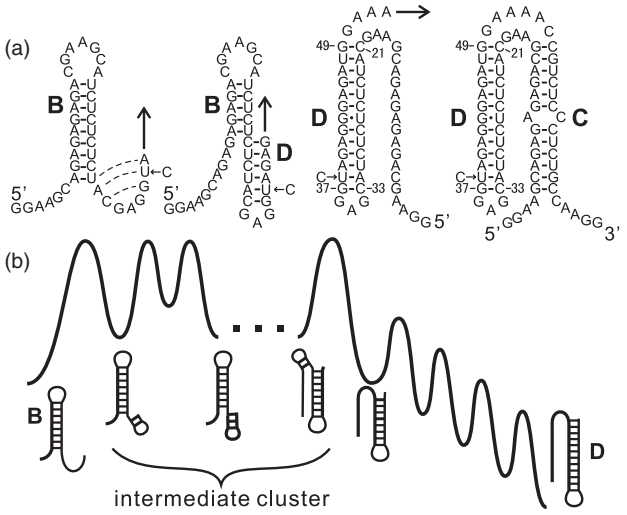

FIG. 7. (a) The main folding pathway of the "reverse" sequence during the transcription. The nascent hairpin B converts into hairpin D and finally folds into the intermediate rod-like structure C+D. (b) The free energy landscape of the tunneling pathway of conformational transition from hairpin B to hairpin D.

to the B to B + A transition, a larger portion of the population would flow to the D + C (i.e., 2R) structure than to the B + A (i.e., 1R) structure. We note that the transition between structure 1R and 2R is much slower and can hardly occur during the transcription. The cotranscriptional folding leads to a populous structure 2R and less populous structure 1R, though structure 1R is thermodynamically more stable than the structure 2R. The result highlights the significance of the kinetic (instead the thermodynamic) effects in cotranscriptional folding.

To investigate the effect of the transcription speed, we predict the folding kinetics with a different elongation speed (400 nucleotides per second). We found the same overall kinetic pathways, except that only 70.1% of the population (instead of 93.6%) folds into structure 2R. Why does the increased transcription speed cause a reduced population flow to structure 2R? Because with the increased elongation speed, the time duration from step 44 to step 57 is now decreased to 0.0325s, which is 62.5% of the time interval with a transcription speed of 250 nucleotides per second. The shorter time window causes less population converted from helix B to helix D (to D + C, i.e., 2R) and accordingly, more population available for the formation of B + A (i.e., 1R). With even higher transcription speed, all the population would fold into B + A. The cotranscriptional folding kinetics was also different from the refolding of a fully synthesized chain, in which the folding of the "direct" and the "reverse" sequence showed similar behavior (Fig. 8): the branched and rod-like structures can coexist at intermediate time, and after a long time the chain folds to the more stable branched structure. In addition, the folding pathway involves more intermediates than cotranscriptional folding.

Crucial to the folding pathway is the transition from helix B to helix D. In order to further examine the influence of the stability of helix D on the folding pathway, Xayaphoummine *et al.* used a single mutation U38/C38 in the experiment.[25] The kinetics of the mutant sequence is

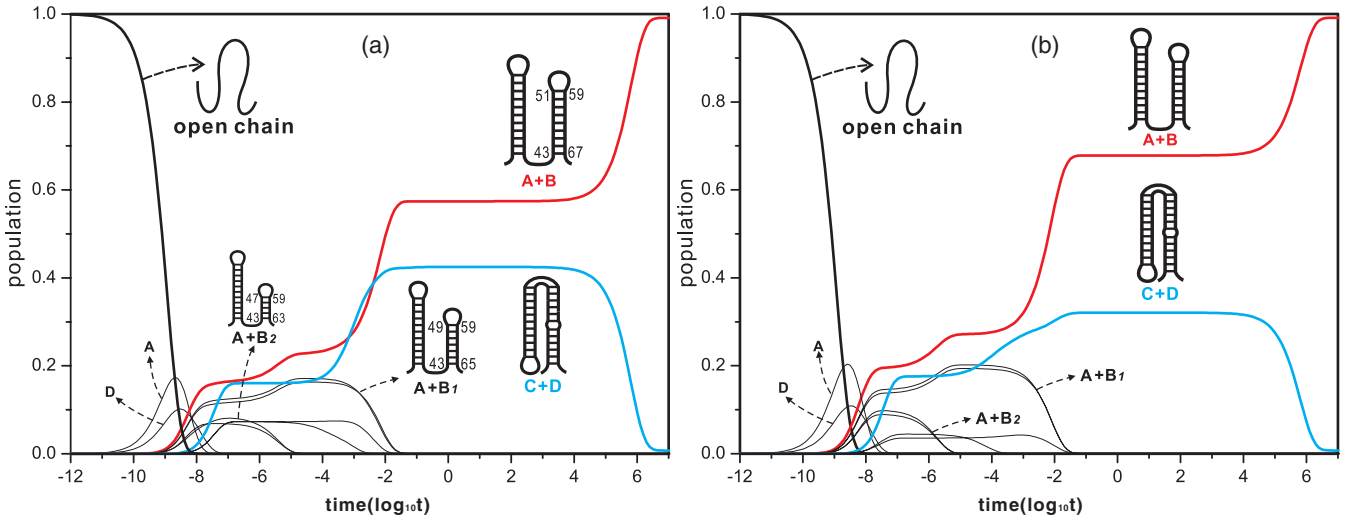

FIG. 8. (a) The populational folding kinetics of the fully transcribed "direct" sequence. (b) The populational folding kinetics of the fully transcribed "reverse" sequence.

the same as the wild sequence when the transcribed sequence is less than 38 nt. The predicted population kinetics for the mutant sequence is shown in Fig. 9 with RNA elongation speed at 250 nt/s. Our calculation shows that by the end of transcription, only 36% of the population folds to the structure C + D (2R) and the rest 63% folds into the structure A + B (1R). The results are consistent with the experimental findings.[25] The mutant sequence shows the same folding pathway as the wild-type sequence. The mutation causes destabilization of helix D as the A32-U38 base pair in hairpin D is replaced with A32-C38. Our calculation shows that because the mutant intermediate cluster is less stable than the wild-type, the transition from B to D has to surmount a higher energy barrier and the transition rate is smaller. As a result, the population of hairpin D is reduced and a higher population of B would remain to allow the formation (addition) of helix A. Therefore, at the end of transcription, in contrast to the wild-type sequence, the mutant sequence would have a higher/lower population for the A + B/C + D structures.

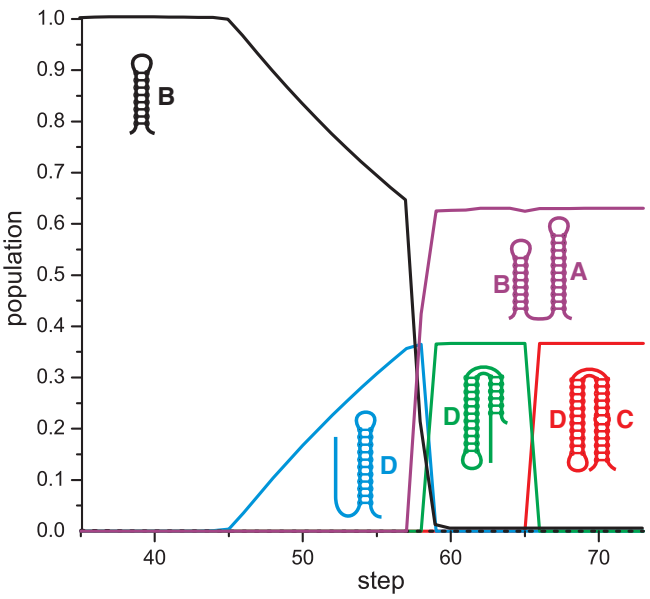

FIG. 9. The population kinetics of the "muatant" sequence at an elongation speed of 250 nt/s.

IV. DISCUSSION

Using an RNA folding kinetics theory based on the creation/disruption and exchange of helices, we investigate the cotranscriptional RNA folding pathways, rate constants, and the conformational populations for several model systems. To account for the change of the conformational ensemble in the chain elongation process, we divide the transcription process into elementary steps, each corresponding to the time duration of transcribing a nucleotide. For the nth step, the chain length is n-nt. We enumerate all the n-nt conformations and calculate the transition rates between the different conformations. From the rate constants, we solve for the time-evolution of the populational distribution. The populational distribution at the end of the $(n-1)$th step serves as the initial condition of the nth step. Such a systematic computational procedure leads to the full folding kinetics (the time-dependent conformational distribution, kinetic intermediates, pathways, folding rates) from the beginning of transcription to the end.

From the detailed studies for a set of paradigmatic simple systems, we find several important features for RNA cotranscriptional folding.

1. In sequential folding, if the transcription speed is not too slow (compared to the detrapping rate of the misfolded states), the RNA molecule tends to form a local structure rather than structures with long-range contacts. This is because (a) the folding of local structures (helices) is fast and (b) once a helix is formed, to disrupt the structure is usually slow. The (local) folding of helices results in "branch-like" structures. By contrast, "rod-like" structures involve long-range base pairs and can be stabilized only for long chains. This chain length requirement determines that the (fast) folding of local structures occurs

before the formation of a "rod-like" structure. Therefore, folding a "rod-like" structure often requires unfolding a local structure, which can be slow. Compared to the refolding of the full-length chain, RNA in sequential folding is more prone to the formation of "branch-like" structures.

2. Chain elongation may help the RNA avoid misfolded states. We assume that a native helix A_1 is formed in the nascent chain. As more nucleotides are transcribed, a misfolded helix NN could form. However, if the rate for the transition from A_1 to NN requires disruption of A_1 and the rate is slower than chain elongation, then NN will not form and the RNA would further fold into a native-like structure that does not require unfolding of A_1. Furthermore, even if the nascent structure A_1 is a non-native structure, as more nucleotides are transcribed, if there exists a (fast) tunneling pathway between A_1 and the native structure, then the RNA can fold to the native state and avoid the formation of other misfolded structures. The product of sequential folding is influenced by the competition between the transition rate between the different nascent structures and the rate of chain elongation. In general, if the transcription speed is significantly faster or slower than the transition rates, the folding product is not very sensitive to the small variations of the transcription speed. However, if the rate for the transition (structural rearrangement) between the intermediate structures is comparable with the rate of chain elongation, the folding result may become very sensitive to the transcription speed.

Because the conformational ensemble is dependent on the chain length, in each step of the chain elongation, the RNA chain samples the different conformational ensemble. A nascent chain may fold into transient structures that may later be displaced by more stable structures as a longer chain is transcribed. Therefore, the cotranscriptional folding kinetics is a result of the interplay between the transition rates for the conformational switches and the timing and accessibility of the conformational space. We find that cotranscriptional folding can have contrastingly different pathways than refolding of a fully transcribed chain. For example, the RNA may form non-native structures that cannot be explained by the thermodynamic and simple kinetics analysis for the refolding of the fully transcribed RNA. For RNA folding *in vivo*, in addition to the competition between the conformational switch and the chain elongation, proteins as RNA chaperones can accelerate the structural rearrangements and cause a rapid preequilibration between the different conformation species.[24] All these suggest that folding *in vivo* during transcription can be quite different from folding of a fully synthesized chain *in vitro*.[8,10,11] For instance, in the auto-catalyzed splicing reaction of tetrahymena group I intron, the functional native structure may form within the timescale of transcription, which is much faster than the refolding of the complete chain *in vitro*.[7,9] The transcription speed also influence the folding due to the kinetic competition of the conformation transition rate, the folding/unfolding time scales of the intermediate structures, and the transcription speed. For example, some riboswitches rely on the speed of RNA transcription and ligand binding kinetics to trigger riboswitch function.[42]

The current theory can be generalized to investigate the effect of transcriptional pausing at specific sites by making the specific step pause for proper time duration. In addition, the model allows us to simulate the effect of protein binding by stabilizing or destabilizing pertinent RNA structures. The current form of the theory involves several limitations. First, the theory does not treat folding/unfolding of tertiary folds such as pseudoknots. Second, the theory cannot treat, at the explicitly atomistic level, the effects of cofactors such as magnesium ions, ligands, and proteins. RNA interactions with these cofactors are important for understanding RNA folding *in vivo*. Nevertheless, the present model enables predictions of cotranscriptional folding for simple systems studied in this work. The theory-experiment agreements suggest that the theory may be a reliable first step for further systematic development of a more complete theory.

ACKNOWLEDGMENTS

This work was partly supported by the Program for New Century Excellent Talents at Wuhan University under Grant No. NCET-06-0623, the National Natural Science Foundation of China under Grant Nos. 10774115 and 30670487 (to W.Z.), the National Institutes of Health Grant No. R01-GM063732, and the National Science Foundation Grant Nos. MCB-0920411 and MCB-0920067 (to S.-J.C.).

[1] A. S. Mironov, I. Gusarov, R. Rafikov, L. E. Lopez, K. Shatalin, R. A. Kreneva, D. A. Perumov, and E. Nudler, Cell **111**, 747 (2002).
[2] D. P. Bartel, Cell **116**, 281 (2004).
[3] O. C. Uhlenbeck, RNA **1**, 4 (1995).
[4] T. M. Henkin and C. Yanofsky, Bioessays **24**, 700 (2002).
[5] E. Merino and C. Yanofsky, Trends Genet. **21**, 260 (2005).
[6] T. Franch, A. P. Gultyaev, and K. Gerder, J. Mol. Biol. **273**, 38 (1997).
[7] S. L. Heilman-Miller and S. A. Woodson, RNA **9**, 722 (2003).
[8] S. L. Brehm and T. R. Cech, Biochemistry **22**, 2390 (1983).
[9] F. Zhang, E. S. Ramsay, and S. A. Woodson, RNA **1**, 284 (1995).
[10] D. K. Treiber and J. R. Williamson, Curr. Opin. Struct. Biol. **11**, 309 (2001).
[11] S. A. Woodson, Biochem. Soc. Trans. **30**, 1166 (2002).
[12] L. B. Zhang, P. Bao, J. L. Michael, and Y. Zhang, RNA **15**, 1986 (2009).
[13] M. Mandal and R. R. Breaker, Nat. Rev. Mol. Cell Biol. **5**, 451 (2004).
[14] J. R. Wickiser, W. C. Winkler, R. R. Breaker, and D. M. Crothers, Mol. Cell **18**, 49 (2005).
[15] J.-F. Lemay, J. C. Penedo, R. Tremblay, D. M. J. Lilley, and D. A. Lafontaine, Chem. Biol. **13**, 857 (2006).
[16] R. Renate, L. Kathrin, G. Dagmar, and M. Ronald, ChemBioChem **8**, 896 (2007).
[17] H. Groeneveld, K. Thimon, and J. van Duin, RNA **1**, 79 (1995).
[18] A. P. Gultyaev, F. H. van Batenburg, and C. W. Pleij, J. Mol. Biol. **276**, 43 (1998).
[19] J. Boyle, G. Robillard, and S. Kim, J. Mol. Biol. **139**, 601 (1980).
[20] R. Nussinov and I. Tinoco, Jr., J. Mol. Biol. **151**, 519 (1981).
[21] T. Pan, I. Artsimovitch, X. Fang, R. Landick, and T. R. Sosnick, Proc. Natl. Acad. Sci. U.S.A. **96**, 9545 (1999).
[22] T. N. Wong, T. R. Sosnick, and T. Pan, Proc. Natl. Acad. Sci. U.S.A. **104**, 17995 (2007).
[23] T. R. Sosnick and T. Pan, Annu. Rev. Biophys. Biomol. Struct. **35**, 161 (2006).
[24] M. M. Elisabeth, Y. W. Peter, W. C. Joseph, and J. F. Martha, PloS Biol. **8**(2), e1000307 (2010).
[25] A. Xayaphoummine, V. Viasnoff, S. Harlepp, and H. Isambert, Nucleic Acids Res. **35**, 614 (2007).
[26] H. Isambert and E. D. Siggia, Proc. Natl. Acad. Sci. U.S.A. **97**, 6515 (2000).

[27]A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert, Proc. Natl. Acad. Sci. U.S.A. **100**, 15310 (2003).
[28]X. Tang, S. Thomas, L. Tapia, D. P. Giedroc, and N. M. Amato, J. Mol. Biol. **381**, 1055 (2008).
[29]I. L. Hofacker, C. Flamm, C. Heine, M. T. Wolfinger, G. Scheuermann, and P. F. Stadler, RNA **16**, 1308 (2010).
[30]M. Geis, C. Flamm, M. T. Wolfinger, A. Tanzer, I. L. Hofacker, M. Middendorf, C. Mandl, P. F. Stadler, and C. Thurner, J. Mol. Biol. **379**, 242 (2008).
[31]W. B. Zhang and S. J. Chen, Proc. Natl. Acad. Sci. U.S.A. **99**, 1931 (2002).
[32]W. B. Zhang and S. J. Chen, Biophys. J. **90**, 765 (2006).
[33]P. N. Zhao, W. B. Zhang, and S. J. Chen, Biophys. J. **98**, 1617 (2010).
[34]T. B. Xia, J. SantaLucia, and D. H. Turner, Biochemistry **37**, 14719 (1998).
[35]D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, J. Mol. Biol. **288**, 911 (1999).
[36]F. Kramer and D. Mills, Nucleic Acids Res. **9**, 5109 (1981).
[37]R. J. Davenport, G. J. Wuite, R. Landick, and C. Bustamante, Science **287**, 2497 (2000).
[38]S. F. Tolic-Norrelykke, A. M. Engh, R. Landick, and J. Gelles, J. Biol. Chem. **279**, 3292 (2004).
[39]Y. Sei-Iida, H. Koshimoto, S. Kondo, and A. Tsuji, Nucleic Acids Res. **28**, E59 (2000).
[40]M. Brenowitz, M. R. Chance, G. Dhavan, and K. Takamoto, Curr. Opin. Struct. Biol. **12**, 648 (2002)
[41]V. Viasnoff, A. Meller, and H. Isambert, Nano Lett. **6**, 101 (2006).
[42]J. K. Wickiser, M. T. Cheah, R. R. Breaker, and D. M. Crothers, Biochemistry **44**, 13404 (2005).