Biophysical Journal, Volume 96

Supporting Material

A new computational approach for mechanical folding kinetics of RNA hairpins

Song Cao and Shi-Jie Chen

Supplementary Material for "A new computational approach for mechanical folding kinetics of RNA hairpins"

Song CAO and Shi-Jie CHEN
Department of Physics and Astronomy and Department of Biochemistry
University of Missouri, Columbia, MO 65211

I. Thermodynamic analysis

Following Ref. (1), we consider a mechanical folding experiment system that consists of four parts (see Fig. S2) (1) :

1. The spring, which models the force measuring device, such as the potential well of the optical trap in optical tweezer experiments.

2. The double-stranded (ds) DNA linker which separates RNA from the substrate to avoid RNA-substrate surface interaction. In fact, previous studies (2) showed that the linker can distort the force-extension curve and perturb the folding thermodynamics (3). The dsDNA is described as a wormlike chain with the force (f)-extension (x_{DNA}) relation given by (4) :
$$f = \frac{k_B T}{l_p}\left(\frac{1}{4(1 - x_{\mathrm{linker}}/L)^2} + \frac{x_{\mathrm{linker}}}{L} - \frac{1}{4}\right) \tag{1}$$
where l_p = 3.57 nm is the persistence length and L is the length of the dsDNA.

3. The double-stranded RNA (dsRNA) between nucleotides a and b; see Fig. S2a.

4. The single-stranded RNA. For an N-nt ssRNA, the force (f)- extension (x_{RNA}) relation can be given by the following equation (1).
$$x_{\mathrm{RNA}} = N\ l_{\mathrm{ss}}\left(\coth\frac{fb}{k_B T} - \frac{k_B T}{fb}\right) \tag{2}$$
where b = 15 Å is the Kuhn length and l_{ss} =5.6 Å is the nucleotide length.

A. Constant distance ensemble

If we consider only hairpin structures for the RNA, the partition function $Z(x)$ for a given total end-to-end distance (see Fig. S1a) is given by:
$$Z(x) = \sum_{a,b}\int_0^x Z_{\mathrm{linker}}(x_{\mathrm{linker}})\ Z_{\mathrm{ssRNA}}(x_{\mathrm{RNA}})\ Z_{\mathrm{dsRNA}}(a,b)\ \frac{e^{-\frac{1}{2}\lambda x_{\mathrm{sp}}^2/k_{\mathrm{B}}T}}{\sqrt{2\pi k_{\mathrm{B}}T/\lambda}}\ dx_{\mathrm{RNA}} \tag{3}$$
where λ is the spring constant, $x_{\mathrm{sp}} = x - R$ is the displacement of the spring and $R = x_{\mathrm{RNA}} + x_{\mathrm{linker}}$ (see Fig. S2a). In Eq. 3, Z_{dsRNA} is the partition function for all the possible chain conformations closed by base pair (a, b). The calculation of Z_{dsRNA} for a given sequence requires a statistical mechanical model. We compute Z_{dsRNA} for a given RNA sequence from our virtual bond-based RNA folding model (Vfold) (5).

The partition function for the single-stranded (ss) RNA segments or the linker $Z_{\text{linker or ssRNA}}$ is related to the Gibbs free energy $\Delta G_{\text{linker or ssRNA}}$ through:

$$Z_{\text{linker or ssRNA}} = e^{-\Delta G_{\text{linker or ssRNA}}/k_B T}. \tag{4}$$

Here the Gibbs free energy $\Delta G_{\text{linker or ssRNA}}$ is equal to the quasi-static work performed on the chain during extension:

$$\Delta G_{\text{linker or ssRNA}} = \int f(x')\, dx' \tag{5}$$

where the force $f(x)$ are given by Eqs. 1 and 2 for the linker and the ssRNA, respectively.

The total partition function $Z(x)$ determines the total free energy $F(x) = -k_B T \ln Z(x)$ of the system, which gives the mean pulling force $\overline{f}$ and the mean extension $\overline{R}$:

$$\begin{aligned} \overline{f} &= -\partial F(x)/\partial x \\ \overline{R} &= x - \overline{x}_{\text{sp}} = x - \frac{\overline{f}}{\lambda} \end{aligned}$$

Here we have made use of the relation $\overline{f} = \lambda \overline{x}_{\text{sp}}$ for the spring.

B. Constant force ensemble

For an ensemble with constant force f, the partition function of the system can be written as:

$$Z(f) = \sum_{a,b} Z_{\text{linker}}(f)\, Z_{\text{ssRNA}}(f)\, Z_{\text{dsRNA}}(a,b) \tag{6}$$

where $Z_{\text{linker}}(f)$ and $Z_{\text{ssRNA}}(f)$ are the partition functions of the linker and the single-stranded RNA under constant force f, respectively:

$$Z_{\text{linker or RNA}}(f) = e^{-\Delta G_{\text{linker or RNA}}(f)/k_B T} \tag{7}$$

where the free energy $\Delta G_{\text{linker or RNA}}(f) = \int_0^f x(f')df'$ with $x(f)$ given by Eqs. 1 and 2 for the linker and the ssRNA, respectively. According to Eq. 2, the free energy of ssRNA is proportional to the length N of ssRNA. The free energy per nucleotide is (see also Fig. S2b)

$$g_s(f) = \Delta G_{\text{ssRNA}}(f)/N = l_{\text{ss}} \int_0^f \left(\coth \frac{f'b}{k_B T} - \frac{k_B T}{f'b} \right) df' \tag{8}$$

The partition function Z(f) gives the average extension for a given constant force f:

$$\overline{R}(f) = -k_{\text{B}} T \frac{\partial}{\partial f} \ln Z(f) \tag{9}$$

C. Theory-experiment comparisons

A key factor that determines the accuracy of the theoretical predictions is the partition function for the RNA. Our recently developed virtual bond-based (Vfold) model for RNA conformational statistics enables accurate calculation

for RNA conformational statistics and partition functions (5). Extensive experimental tests indicated that the Vfold model is reliable (6, 7). Therefore, we use the Vfold model to treat RNA hairpin folding here.

We compute the force-extension curve (FEC) (f vs. x) based on the theory described above and compare the theoretical predictions with the experimental results. Specifically, we choose TAR RNA, whose mechanical folding was recently measured by Li and Tinoco et al (8), to test our model. In the experiment, the ion concentration is 0.1 M KCl. However, our available entropy and enthalpy parameters for the base stacks are under standard 1M NaCl (9) condition. We use the previously developed ion electrostatic theory (10, 11) to account for the corrections to the base stacking parameters due to the different ionic conditions. Theory-experiment comparison shows satisfactory agreement; see Fig. S1. We find that the force-extension curve is sensitive to the spring constant. For $\lambda = 0.01$ pN/nm, the FEC is in agreement with the soft spring limit for the constant force ensemble. The critical force for unzipping the RNA is about 13 pN, which is in good agreement with the experimental value (13.4 pN) (8). We choose the experimental data with the slowest pulling rate 0.4 pN/s, which may be the closest to the quasistatic thermodynamic equilibrium process.

II. Master equation and kinetic cluster theory

In a master equation description, the kinetics for the fractional population (or the probability) $p_i(t)$ for the i-th state ($i = 0, .., \Omega - 1$, where Ω is the total number of chain conformations) is described as the difference between the rates for transitions entering and leaving the state:

$$\frac{d}{dt}p_i(t) = \sum_{j=0}^{\Omega-1} \left(k_{j\to i}\ p_j(t) - k_{i\to j}\ p_i(t) \right)$$

where $k_{j\to i}$ and $k_{i\to j}$ are the rate constants for the respective transitions. In terms of the rate matrix, defined as $\mathbf{M}$ is the rate matrix defined as $M_{ij} = k_{i\to j}$ for $i \neq j$ and $M_{ij} = -\sum_{l\neq i} k_{il}$ for $i = j$, we can write the above master equation in a matrix form: $d\mathbf{p}(t)/dt = \mathbf{M} \cdot \mathbf{p}(t)$, where $\mathbf{p}(t)$ is the fractional populational vector col $[p_0(t), p_1(t), ..., p_{\Omega-1}(t)]$,

For a given initial folding condition at $t = 0$, by diagonalizing the rate matrix M, we obtain the populational kinetics $\mathbf{p}(t)$ as a sum over all the possible eigenmodes of the rate matrix:

$$\mathbf{p}(t) = \sum_{m=0}^{\Omega-1} C_m \mathbf{n}_m e^{-\lambda_m t} \tag{10}$$

where $-\lambda_m$ and $\mathbf{n}_m$ are the m-th eigenvalue and eigenvector of the rate matrix $\mathbf{M}$, and C_m is the coefficient that is dependent on the initial condition. The eigenvalue spectrum contains a static equilibrium state corresponding to the eigenvalue $\lambda_0 = 0$. We denote the first and second nonzero eigenvalues as λ_1 and λ_2, respectively. A large gap between λ_1 and λ_2 indicates a single exponential kinetics with the rate constant equal to λ_1.

In our calculation, a kinetic move is defined as the formation/disruption of a stack or a stacked base pair. We use the transition state theory to calculate the rate constant for each kinetic move. We assume that the transition state for the formation of a base stack is the state that the bases have been juxtaposed to the restricted base pairing (stacking) positions, but have not yet "reacted" to form the stabilizing base stacking and base pairing interactions. Similarly, the transition state for the disruption of a base stack is the state that the stabilizing stacking and base pairing interactions have been disrupted but the nucleotides have not yet been liberated from the (restricted) base pairing/stacking positions.

According to the transition state model, we obtain the following (force-free) transition rate for the formation ($k_+(0)$) and disruption ($k_-(0)$) of a base stack:

$$k_+(0) = k_0\ e^{-\Delta S/k_B}; \quad k_-(0) = k_0\ e^{-\Delta H/k_BT}. \tag{11}$$

where ΔS and ΔH are the entropy and enthalpy changes for the formation/disruption of a base stack or a stacked base pair, respectively. k_0 is a prefactor. Fitting from a series of experimental data gives $k_0 = 6.6 \times 10^{12}$ for a AU base pair and 6.6×10^{13} for a GC base pair (15).

Due to large number of possible states for a long chain, the master equation is practically limited to short sequences. The kinetic cluster theory (12–16), however, based on the classification of conformations into reduced states, can effectively reduce the size of the conformational ensemble. From the rate constants (Eq. 11), we can identify groups of conformations such that their inter-conversion rates are fast so they will quickly reach local preequilibration. We define such a group of conformations as a kinetic cluster (a macrostate). Transitions between conformations belonging to different clusters have slow rates. In terms of the preequilibrated clusters, the overall kinetics corresponds to the equilibration between the clusters. For example, if there exists a single outstandingly slow rate constant that corresponds to the formation/disruption of a particular base stack denoted as s^*, the original conformational ensemble can be reduced to two clusters U and N:

$$\begin{aligned} \text{cluster } N &= \text{all the conformations with } s^* \text{ formed;} \\ \text{cluster } U &= \text{all the conformations without } s^*. \end{aligned}$$

The formation/disruption of $s*$ is the rate-limiting step for the formation/disruption of the N states. Transitions between the different clusters ($U \leftrightarrow N$) are through micro-transitions $U_i \leftrightarrow N_i$ between conformations U_i (in cluster U) and N_i (in cluster N), where $U_i \leftrightarrow N_i$ corresponds to a single kinetic move. There usually exist many such (U_i, N_i) pairs, i.e., many pathways for the inter-cluster transitions. The sum over all these (micro)inter-cluster pathways ($\sum_{U_i \leftrightarrow N_i}$) determines the inter-cluster transition rate:

$$k_{U\to N} = \sum_{U_i \to N_i} \mathrm{p}_{U_i}\ k_{U_i \to N_i}; \quad k_{N\to U} = \sum_{N_i \to U_i} \mathrm{p}_{N_i}\ k_{N_i \to U_i}. \tag{12}$$

where p_{U_i} and p_{N_i} are the equilibrium fractional populations of U_i and N_i in the respective clusters:

$$\mathrm{p}_{U_i} = e^{-(G_{U_i} - G_U)/k_BT}; \quad \mathrm{p}_{N_i} = e^{-(G_{N_i} - G_N)/k_BT} \tag{13}$$

G_{U_i} and G_{N_i} are the free energies of conformations U_i and N_i, respectively, and G_U and G_N are the free energies of clusters U and N, respectively:

$$G_U = -k_BT \ln(\sum_j e^{-G_{U_j}/k_BT}); \quad G_N = -k_BT \ln(\sum_j e^{-G_{N_j}/k_BT}) \tag{14}$$

Here $\sum_j$ is the sum over all the possible conformations in the respective clusters. An important conclusion from Eq. 12 is that the inter-cluster transition rate $k_{U\leftrightarrow N}$ and hence the overall folding rate is determined not only by the rate for each individual transition $k_{U_i \leftrightarrow N_i}$ between two conformations, but also by the populational distribution of the conformations (p_{U_i} and p_{N_i}) in the respective clusters.

The kinetic cluster theory is advantageous over the rate matrix master equation not only in its ability to treat longer sequences, but also in its ability to give the rates and pathways. The most probable (dominant) pathway for the transition between two clusters $N \rightarrow U$ in Eq. 12 is the transition between conformations $\mathrm{U}_i \rightarrow N_i$ with the maximum value of $\mathrm{P}_{U_i}\mathrm{k}_{U_i \rightarrow N_i}$ in Eq. 12. In general, a folding reaction may involve multiple parallel dominant pathways and each pathway often involves multiple intermediate states (clusters). The slowest inter-cluster transitions in the pathways give the rate-limiting steps of the folding reaction. Moreover, we can estimate the folding rate from the pathways. For a multi-step pathway $C_0 \rightarrow C_1 \rightarrow C_2 ... \rightarrow C_n$ (C_i is the i-th cluster along the pathway), the folding rate (k_n) for $C_0 \rightarrow C_n$ is can be conveniently (and crudely) estimated from the following equation (14, 16):

$$k_n = k_1 r_{n-1} \prod_{p=1}^{n-2} \frac{r_p}{1 - r'_{p+1} r_p} \tag{15}$$

where k_1 is the rate for $C_0 \rightarrow C_1$, r_i and r'_i are the probabilities for the forward $C_i \rightarrow C_{i+1}$ and the backward (rebounce) $C_i \rightarrow C_{i-1}$ reactions, respectively:

$$r_i = \frac{k_{C_i \rightarrow C_{i+1}}}{k_{C_i \rightarrow C_{i+1}} + k_{C_i \rightarrow C_{i-1}}}; \;\; r'_i = 1 - r_i. \tag{16}$$

We note that the rigorous master equation based on the original rate matrix can also give the rates and the rate-limiting steps (17). For the short sequences that allows for master equation calculations, we find that the original rate matrix and the approximate formula (Eq. 16 give the same results for the kinetic rates and the rate-limiting steps.

References

[1] Gerland, U., R. Bundschuh and T. Hwa. 2003. Mechanically probing the folding pathway of single RNA molecules. *Biphys. J.* **84**: 2831-2840.

[2] Cao, S. and S. -J. Chen. 2008. Predicting ribosomal frameshifting efficiency. *Phys. Biol.* **5**: 016002.

[3] Green, L., C. -H. Kim, C. Bustamante and I. Tinoco, Jr. 2008. Characterization of the Mechanical Unfolding of RNA Pseudoknots. *J. Mol. Biol.* **375**: 511-528.

[4] Bustamante, C., J. F. Marko, E. D. Siggia and S. B. Smith. 1994. Entropic elasticity of lambda-phage DNA. *Science.* **265**: 1599-1600.

[5] Cao, S. and S. -J. Chen. 2005. Predicting RNA folding thermodynamics with a reduced chain representation model.*RNA.* **11**: 1884-1897.

[6] Cao, S. and S. -J. Chen. 2006. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.* **34**: 2634-2652.

[7] Cao, S. and S. -J. Chen. 2006. Free energy landscapes of RNA/RNA complexes: with applications to snRNA complexes in spliceosomes. *J. Mol. Biol.* **357**: 292-312.

[8] Li, P. T. X., D. Collin, S. B. Smith, C. Bustamante and I. Tinoco, Jr. 2006. Probing the mechanical folding kinetics of TAR RNA by hopping, force-jump, and force-ramp methods. *Biophys. J.* **90**: 250-260.

[9] Serra, M. J. and D. H. Turner. 1995. Predicting thermodynamic properties of RNA. *Methods Enzymol.* **259**: 242-261.

[10] SantaLucia, Jr., J. 1998. A unified view of polymer, dumbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA* **95**: 1460-1465.

[11] Tan, Z.-J. and S. -J. Chen. 2006. Nucleic acid helix stability: Effects of salt concentration, cation valence and size, and chain length *Biophys. J.* **90**: 1175-1190.

[12] Zhang, W. B. and S. -J. Chen. 2006. Exploring the complex folding kinetics of RNA hairpins: II. Effect of sequence, length, and misfolded state. *Biophys. J.* **90**: 778-787.

[13] Konishi, Y., T. Ooi and H. A. Scheraga. 1982. Regeneration of ribonuclease A from the reduced protein. Rate-limiting steps. *Biochemistry.* 21: 4734-4740.

[14] Zhang, W. B. and S. -J. Chen. 2003. Analyzing the biopolymer folding rates and pathways using kinetic cluster method. *J. Chem, Phys.* **119**: 8716-8729.

[15] Zhang, W. B. and S. -J. Chen. 2006. Exploring the complex folding kinetics of RNA hairpins: I. General folding kinetics analysis. *Biophys. J.* **90**: 765-777.

[16] Cao, S. and S. -J. Chen. 2007. Biphasic RNA folding kinetics and telomerase activity. *J. Mol. Biol.* **367**: 909-927.

[17] Zhang, W. and S. -J. Chen. 2003. Master equation approach to finding the rate-limiting steps in biopolymer folding *J. Chem. Phys.* **118**: 3413-3420.

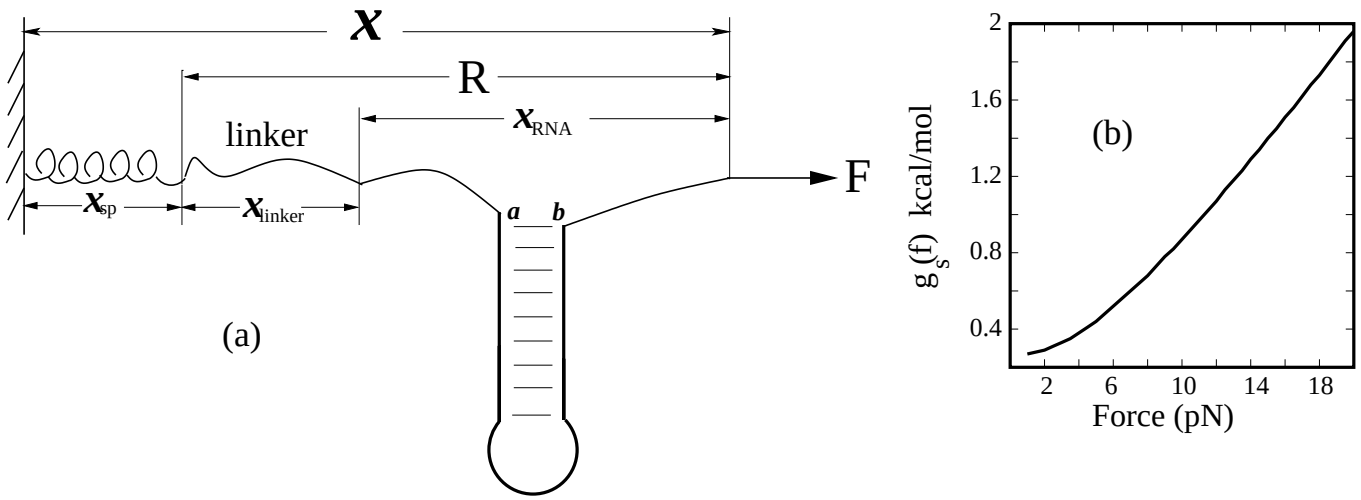

Figure 1: (a) A sketch of the system. The potential of optical trap is approximated as a linear spring with spring constant λ. (b) The elastic free energy ($g_s(f)$) per nucleotide for different force.

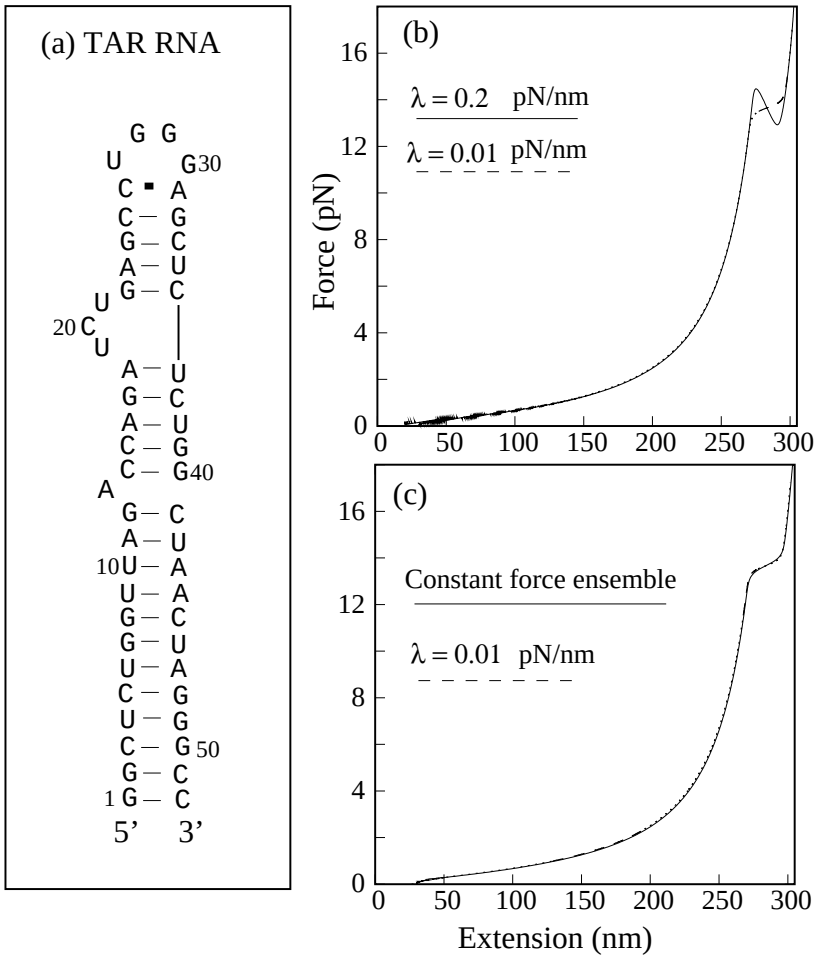

Figure 2: (a) The predicted secondary structure for TAR RNA. (b) The force extension curve (FEC) for two different spring constants $\lambda = 0.2$pN/nm(solid line) and 0.01 pN/nm (dashed line) at 0.1M KCl. (c) The calculated FEC for TAR RNA for constant force (solid line) ensemble and constant distance ensemble (dashed line) with spring constant $\lambda = 0.01$pN. We obtain a critical force about 13 pN, which is in good agreement with the experimental value (13.4 pN) (8).

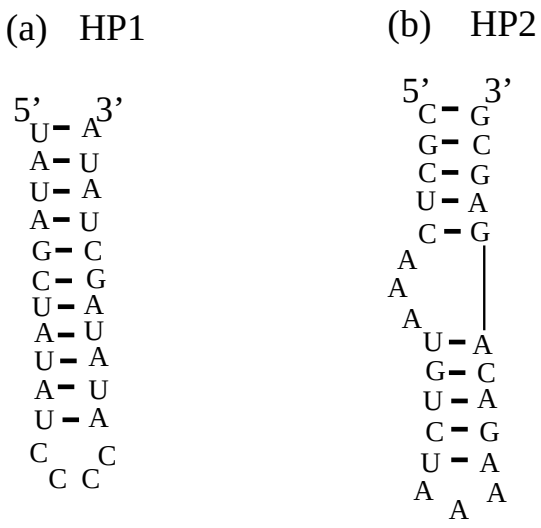

Figure 3: The predicted secondary structures for two hairpins (a) HP1 and (b) HP2.

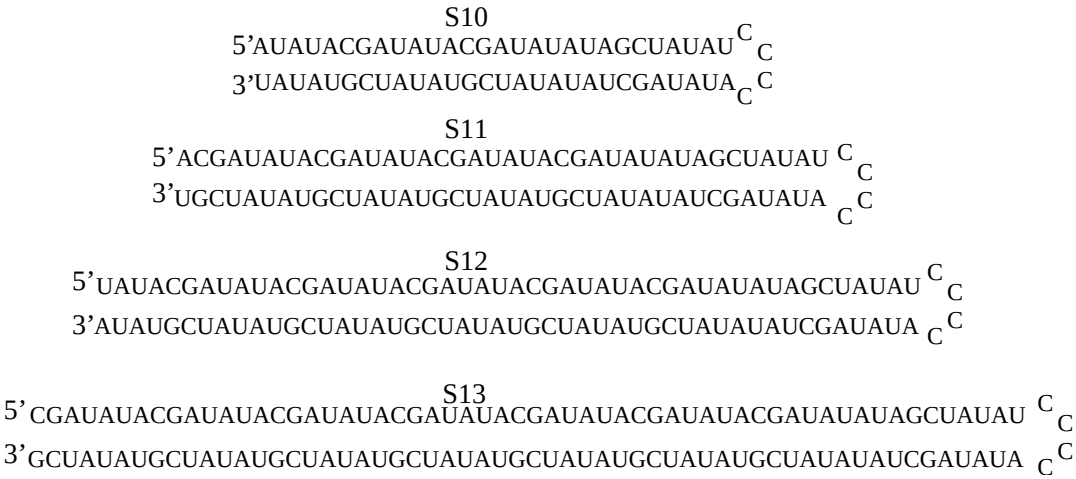

Figure 4: The secondary structures for four larger hairpins S10, S11, S12 and S13 with the lengths 60, 80, 100 and 120 nts, respectively.

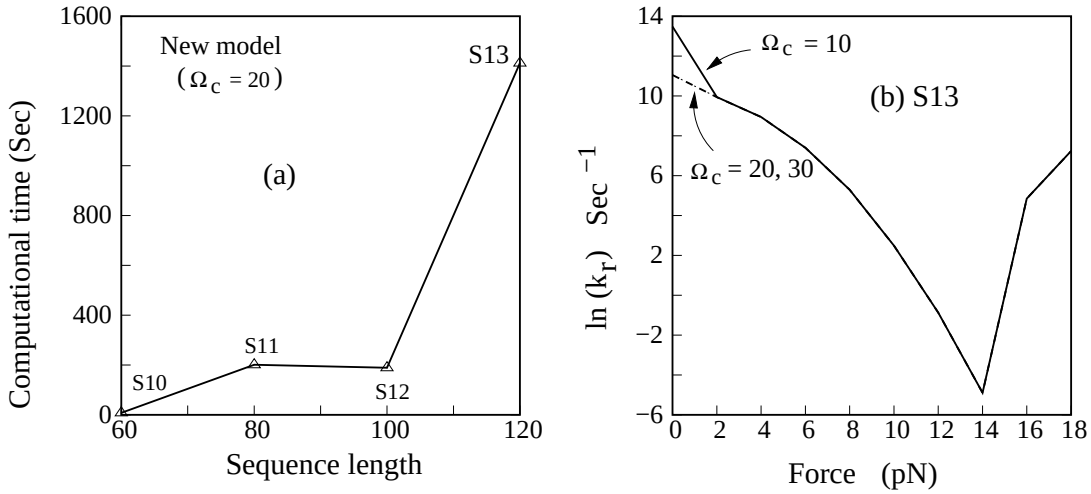

Figure 5: (a) The computational time for four larger RNA hairpins (S10, S11, S12 and S13) using the new kinetic method. In the calculation, Ω_c is set to be 20. For the 120-nt hairpin S13, it takes less than half an hour to compute the folding rate. Thus, our new method is computationally efficient and useful for larger RNA hairpins. (b) The force dependence of the relaxation rate with different Ω_c. The temperature is 25 °C. We find that the folding rates converge for $\Omega_c \geq 20$.

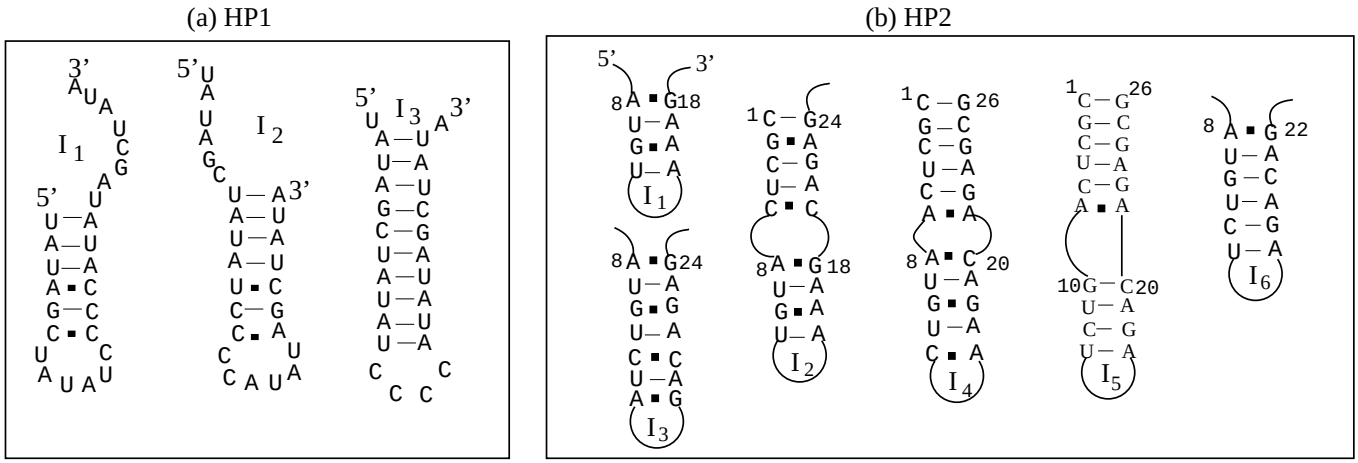

Figure 6: (a) The secondary structures of the native-like (I_3) and misfolded (I_1 and I_2) intermediates for HP1. (b) The secondary structures of the native-like (I_5 and I_6) and nonnative (I_1, I_2, I_3, I_4) intermediates for HP2.

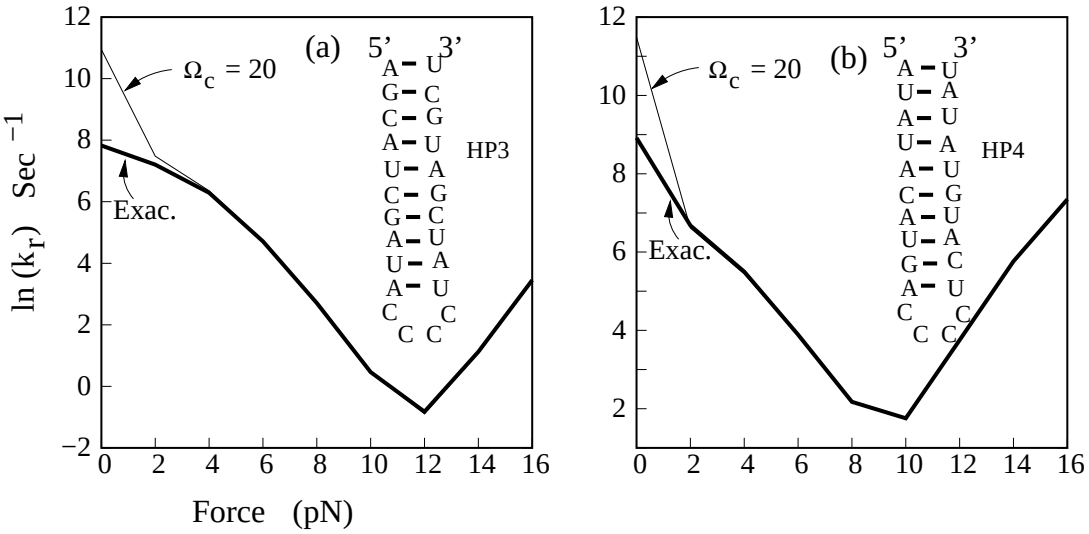

Figure 7: The force dependence of the relaxation rate at $T = 25°C$. We compare the results from the complete ensemble model and from the new kinetic model for two sequences (a) HP3 and (b) HP4. We find that the folding rates from the new kinetic model with $\Omega_c = 20$ agree well with the results from the complete ensemble model for force ≥2 pN.

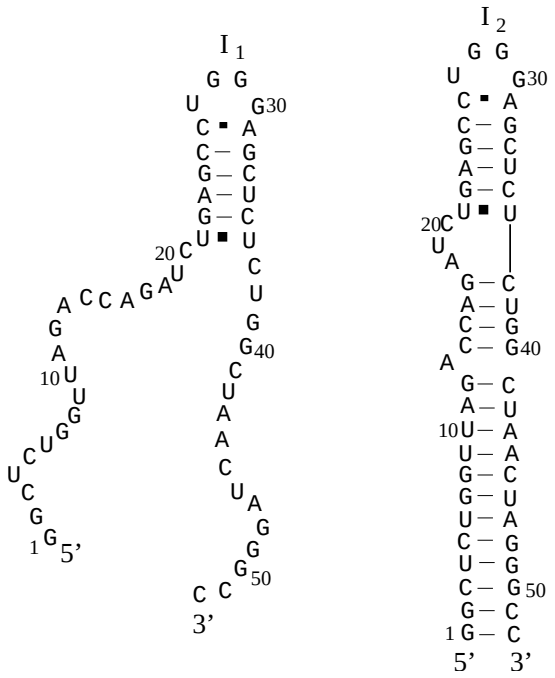

Figure 8: The secondary structures for two intermediates I_1 and I_2 in TAR RNA folding process at f = 8 pN and T = 25 °C.