Work Report

Zhuoran Qiao

August 17, 2018

1 Introduction

1 Developed a genetic algorithm based approach to simulate kinetics of co-transcriptional folding.

2 Tested effect of folding rate variation on folding population dynamics and $p_{unbound}$.

2 Progress

2.1 Framework

Various algorithms or programs have been developed to predict RNA folding pathway utilizing force-field based simulations[4] and multiple sampling methods based on Monte Carlo trajectories[3][2] or coarse graining of energy landscape built on Markov state model[6][5]. Those present methods have succeeded in revealing multiscale dynamic events during RNA folding, however are either designed for only predicting annealing dynamics or limited to RNA segments with length up to hundred bases. To quantitatively predict folding dynamics coupled with transcription, we developed a genetic algorithm and chemical Master equation based approach, which is capable of capturing kilobase level kinetics. Our method is built on two following assumptions:

1 All populated RNA secondary structures (SS) are linkage of locally optimal or sub-optimal structures at various folding sites;

2 Global structural rearrangement of a partial RNA segment is permitted only if it's folding to the optimal SS on that segment.

Formally, we denote a domain $D_{A,\,B}$ as a segment between base A and B that all contacts on that segment are local. For simplicity, we denote **foldon** as domains with optimal secondary structures: $D^{foldon}_{A,\,B}$ = MFE(sequence[A,B]). Note that $'.'$ is a trival example of foldon. Our assumption 1 can be rewritten as

$$D_{A,\,B} = D^{foldon}_{A,i_1} \oplus D^{foldon}_{i_1,i_2} \oplus ... \oplus D^{foldon}_{i_n,B} \tag{1}$$

Where $\oplus$ represents a link operation. Note that all structural information of $D_{A,\,B}$ is encoded by the sequential representation $[A,\, i_1,\, ...,\, i_n,\, B]$; as a foldon is also a linkage of smaller foldons, there could be multiple way to represent $D_{A,\,B}$. Here we introduce **Irreducible Foldon Representation** (IFR) to be the sequential representations for which linkage of every adjacent foldons is not another foldon: $\forall k,\, D^{foldon}_{i_k,i_{k+1}} \oplus D^{foldon}_{i_{k+1},i_{k+2}} \neq D^{foldon}_{i_k,i_{k+2}}$. Then the sufficient and necessary condition for structural rearrangement is

$\langle D^u_{A,B}|\hat{\mathbf{T}}|D^v_{A,B}\rangle \neq 0$ if and only if $\exists\, i,\, j$ satisfies
$i,\, j \in D^u_{A,B}$.IFR, $i,\, j \in D^v_{A,B}$.IFR;
$D^u_{A,i} = D^v_{A,i}$, $D^u_{j,B} = D^v_{j,B}$;
$D^u_{i,j} = D^{foldon}_{i,j}$ or $D^v_{i,j} = D^{foldon}_{i,j}$.
Then $\langle D^u_{A,B}|\hat{\mathbf{T}}|D^v_{A,B}\rangle = \langle D^u_{i,j}|\hat{\mathbf{T}}|D^v_{i,j}\rangle$.

2.2 Folding pathway identification & Rate calculation

Given two domains between which rearrangement is allowed, the task is to compute forward and backward rate constant linking each other. Methods to rigorously calculate the maximum likelihood pathway between arbitrary RNA structrues have been reported[1]; here we proposed a computationally feasible approach: the forward free energy barrier is estimated by sum up all free energy associated with old stacks unzipping and new loop forming; then rate constant $k_{uv} = \langle D^u_{A,B}|\hat{\mathbf{T}}|D^v_{A,B}\rangle$ is calculated by Arrhenius approximation $k_{uv} = k_0 \exp\left[-\frac{1}{RT}(\Delta G^{Stack}_u + \Delta G^{Loop}_v)\right]$. 'New' and 'old' helices are identified by comparing elementary domains (defined as domains that cannot be decomposed to smaller valid domains) between reactant and product domains; identical elementary domains are excluded.

2.3 Algorithm procedure

During every iterative elongation step, an active species pool of strands with unique SS and diffrent population is updated. New candidate strands $D^{Candidate}_{0,L+\Delta L}$ with length $L+\Delta L$ are generated by a recombination process: for every old strand $D^{Strand}_{0,\mathrm{L}}$, all indices in its IFR is identified as possible rearrangement site, then its child strands is generated by linking partial domains $D^{Strand}_{0,\mathrm{Site}}$ with a foldon $D^{foldon}_{\mathrm{Site},L+\Delta L}$ that terminated at $L+\Delta L$.

We assume that elongation will not change the inital population distribution of secondary structures: child strands with the exact parental SS on $[0, L]$ ($D^{child}_{0,L+\Delta L} = D^{strand}_{0,L} \oplus D^{foldon}_{L,L+\Delta L}$) will also inherit the population of their parents.

After structual generation the rate matrix among all candidate strands within the new active species pool is calculated (see part 2.2). Then the population distribution of strands after elongation is computed by propagate the chemical master equation.

For the sake of computational efficiency, we introduce a cutoff N as the size limit of the active species pool. After each elongation step, we impose a selection sweep on all active strands; species with top N fitness is reserved. In the current edition, we simply used population as the fitness function. Population of remaining strands within the active pool is renormalized after selection.

Pseudocodes of the whole simulation procedure are as follows (Algorithm 1):

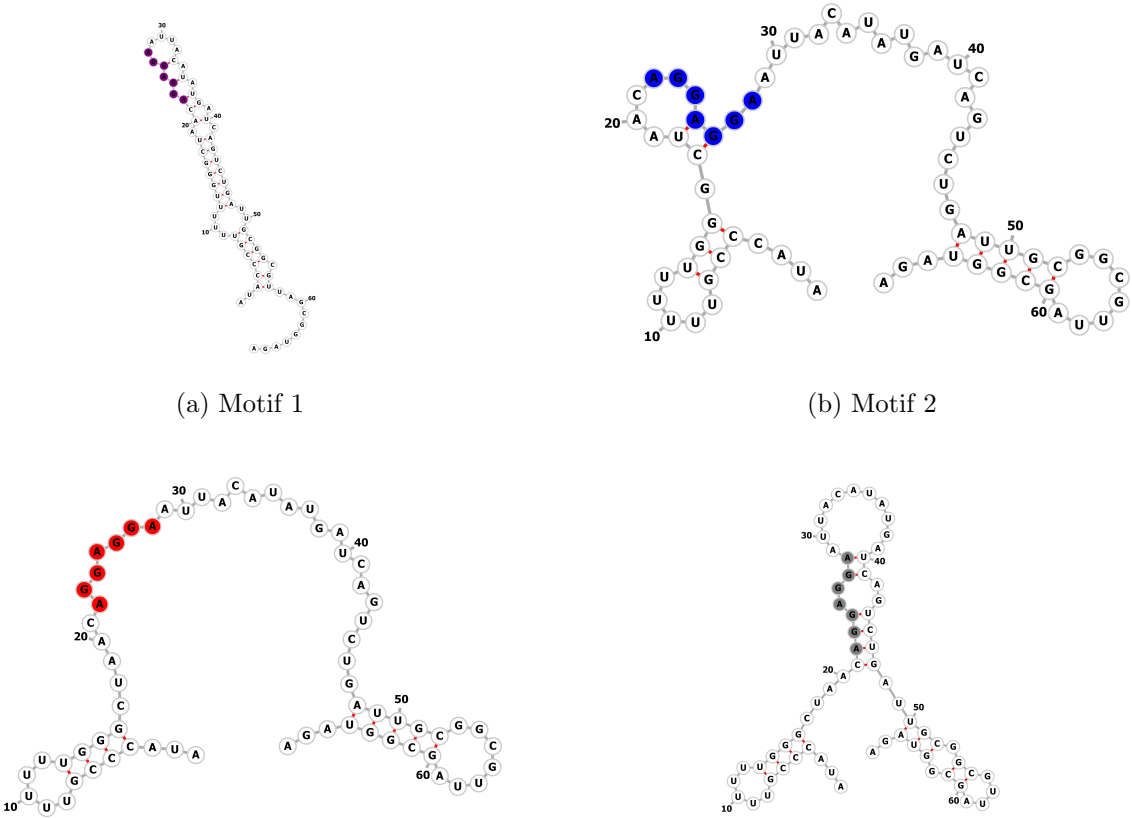
2.4 Test results

The only remaining free parameter to be determined is k_0/k_T, the ratio of pre-exponential factor in Arrhenius rate formulation for folding and trancription rate ($nt \cdot s^{-1}$). I tuned k_0/k_T from 10^1 to 10^{15} and obtained the data for $k_0/k_T = \infty$ by calculating stationary distribution ($\frac{1}{Q}\exp(-G_i)$) after every elongation step for strand i in active pool.

Population analysis For folA-WT four predominant local folding motifs within SD sequence are identified. Figure 1 shows exemplary secondary structures containing these motifs; figure 2 shows evolution of these structure motifs during co-transcriptional folding with different k_0/k_T. Identical motifs are marked by the same color as in figure 1. Surprisingly we noticed that when $k_0/k_T = \infty$, exchange between dominant motifs was very frequent at early stage of transcription,

Algorithm 1 Co-transcriptional folding elongation procedure

Initalize ActivePool
while sequence length > current length **do**
  OldPool ← ActivePool
  renew ActivePool
  Current length ← Current length + dL
  dt ← dL/k_T ▷ Transcription time
  for left boundary ∈ {0, dL, 2dL, ..., Current length - dL} **do** ▷ Get all new foldons
    $D^{foldon}_{\text{left boundary, Current length}}$ ← numpy.mfe(sequence[left boundary, Current length])
  end for
  for Strand ∈ OldPool **do** ▷ Recombination
    for Site ∈ Strand.IFR **do**
      $D^{Candidate}_{0,\text{Current length}} \leftarrow D^{Strand}_{0,\text{ Site}} \oplus D^{foldon}_{\text{Site, Current length}}$
      if $D^{Candidate}_{0,\text{Current length}} \in$ ActivePool **then**
        update $D^{Candidate}_{0,\text{Current length}}$.IFR
      else
        add $D^{Candidate}_{0,\text{Current length}}$ to ActivePool
      end if
      if site = Current length − dL **then**
        $\langle$ActivePool.**population** $|D^{Candidate}_{0,\text{Current length}}\rangle \leftarrow \langle$OldPool.**population** $|D^{Strand}_{0,\text{ Site}}\rangle$
      end if
    end for
  end for
  for $D^{\text{u}}_{0,\text{Current length}} \neq D^{\text{v}}_{0,\text{Current length}} \in$ ActivePool **do** ▷ Calculate new rate matrix
    calculate $D^{u}_{\text{rearrange}}$, $D^{v}_{\text{rearrange}}$ ▷ Find all helices involved in rearrangement
    $\langle D^{u}_{\text{rearrange}}|\hat{\mathbf{T}}|D^{v}_{\text{rearrange}}\rangle \leftarrow k_0 \exp\left(-\frac{1}{RT}(\Delta G^{Stack}_{u} + \Delta G^{Loop}_{v})\right)$
  end for
  $\langle$ActivePool.**population** $| \leftarrow \langle$ActivePool.**population** $| \exp\left(t \times \hat{\mathbf{T}}\right)$ ▷ Master equation
  reserve top N populated strands in ActivePool ▷ Selection
  renormalize $\langle$ActivePool.**population** $|$
end while



(a) Motif 1

(b) Motif 2

(c) Motif 3

(d) Motif 4

Figure 1: Exemplary ss containing folding motifs within folA-WT Shine-Dalgarno sequence

indicating the sensitivity of local structures on long-range contacts. We also noticed that motif predominance after transcription strongly depended on folding rate, reiterating

$p_{unbound}$ **analysis** We calculated $p_{unbound}$ with respect to transcription time and k_0/k_T (Figure 3-11). Deviation of asymptotic behavior from equilibrium value (calculated by nupack.ppairs) is possiblly due to the limited set of foldons (only used mfe to obtain current results).

2.5 Model optimization

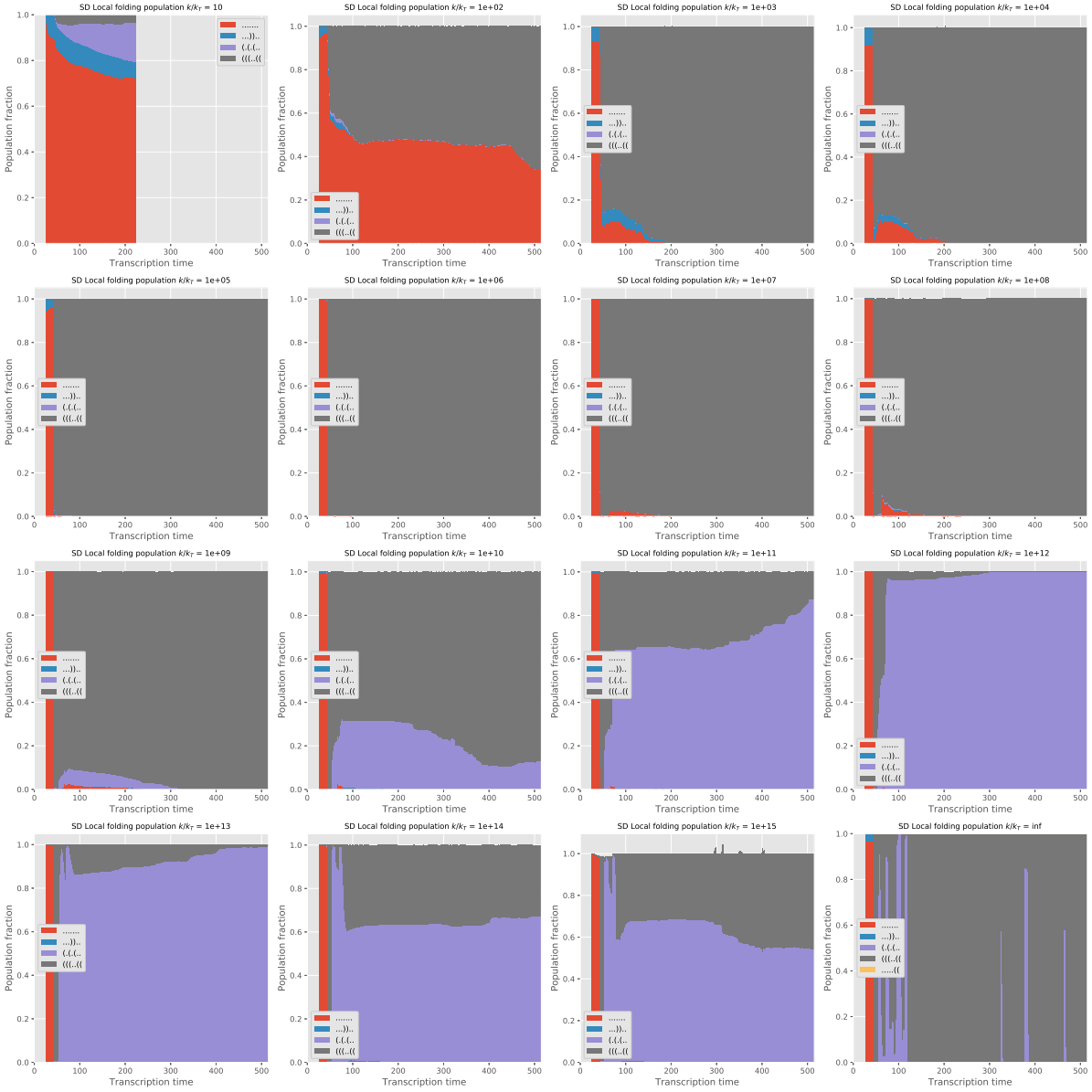


Figure 2: Population dynamics of four structrual motifs during co-transcriptional folding.

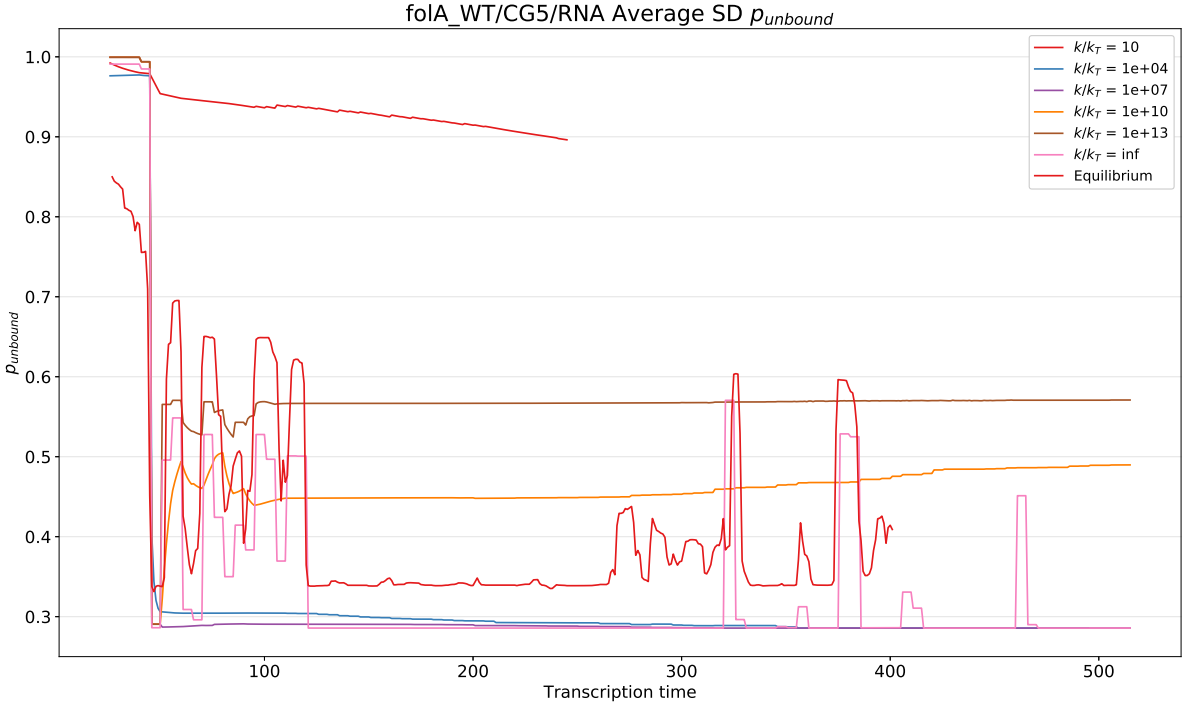


Figure 3

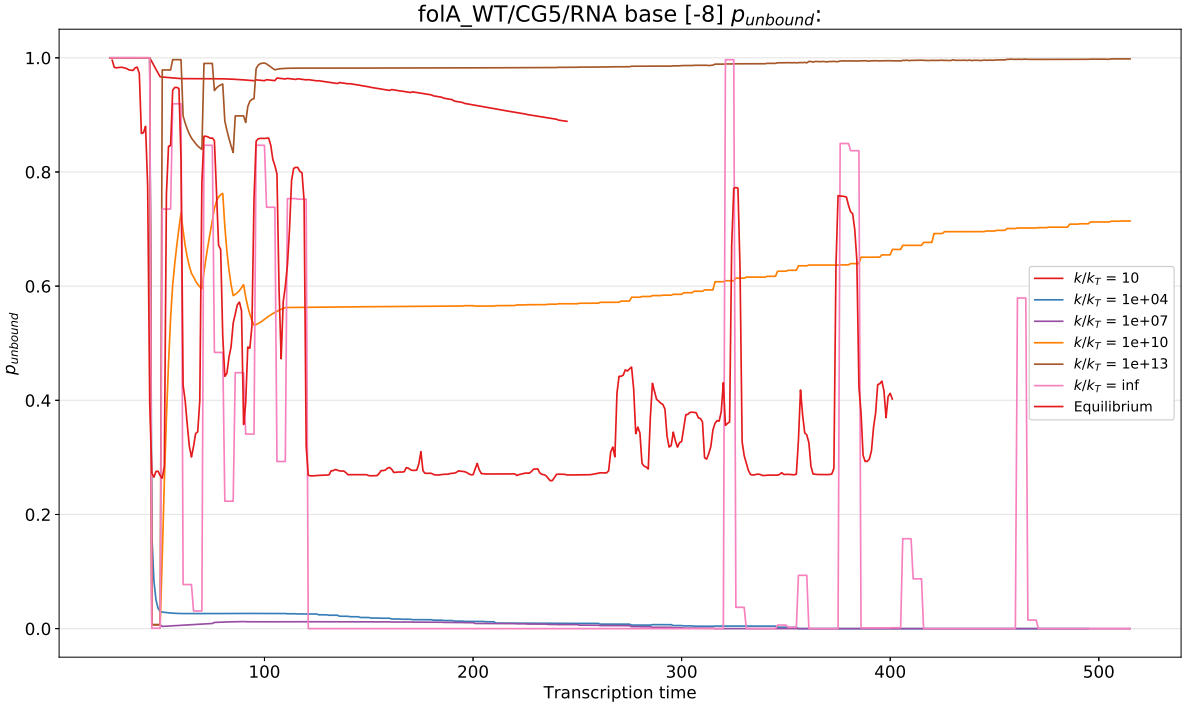


Figure 4

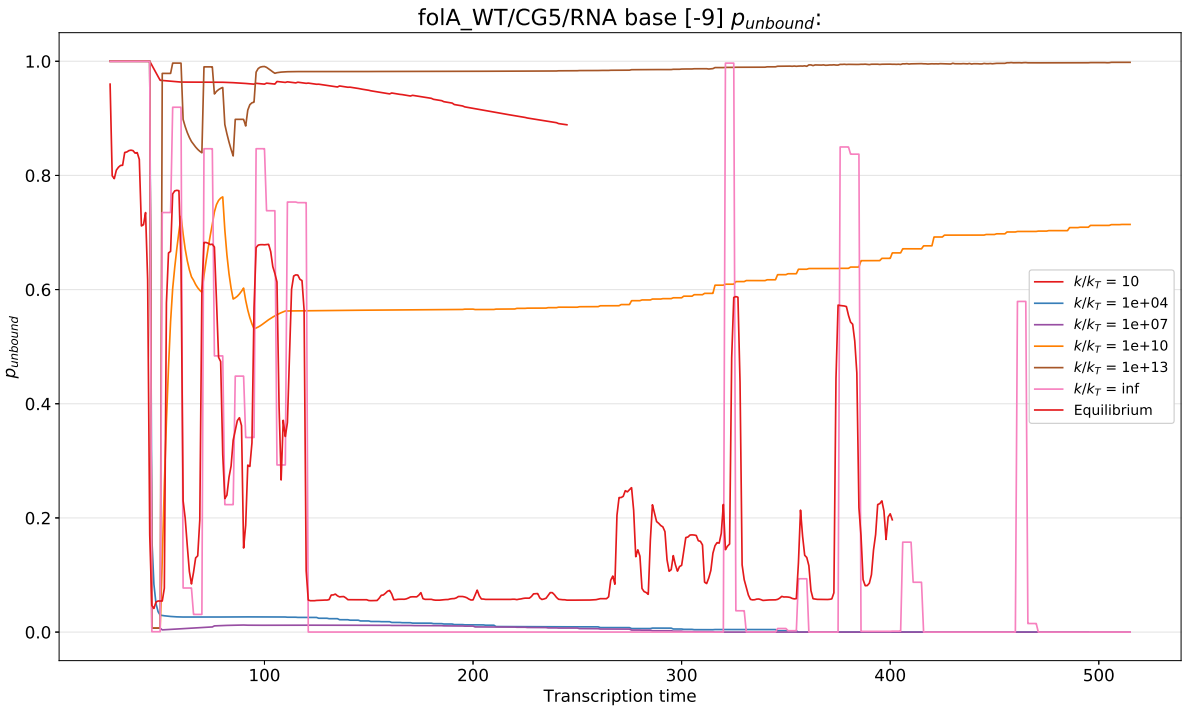


Figure 5

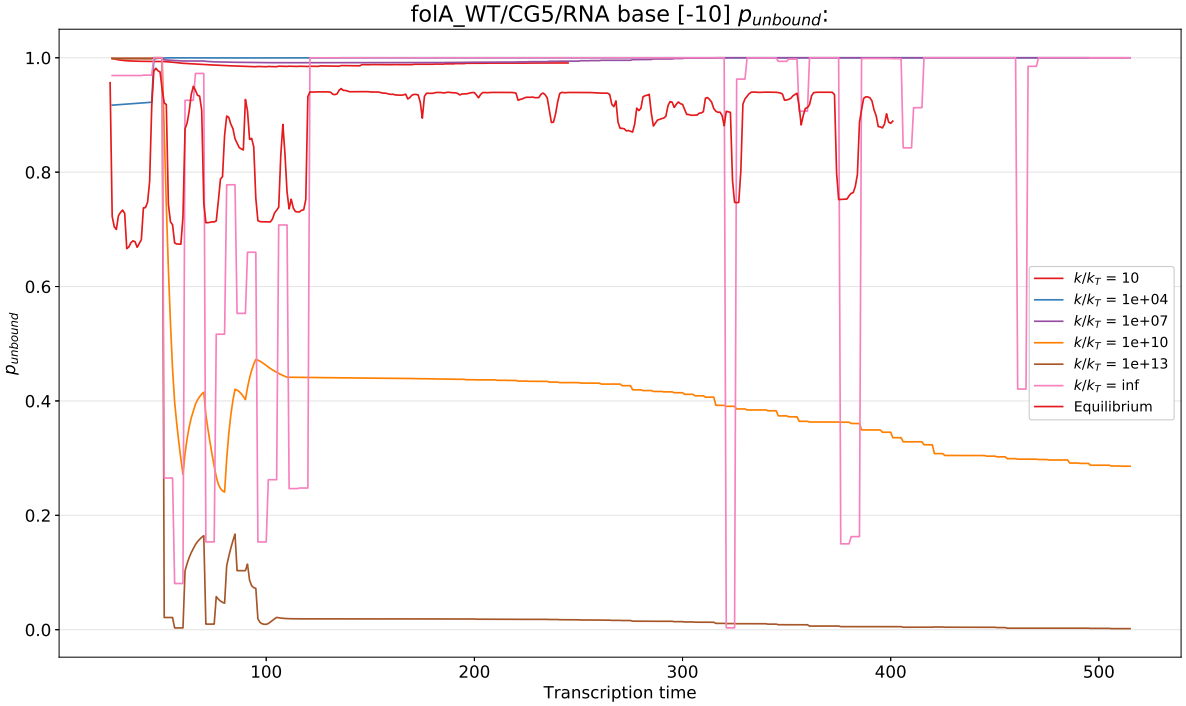


Figure 6

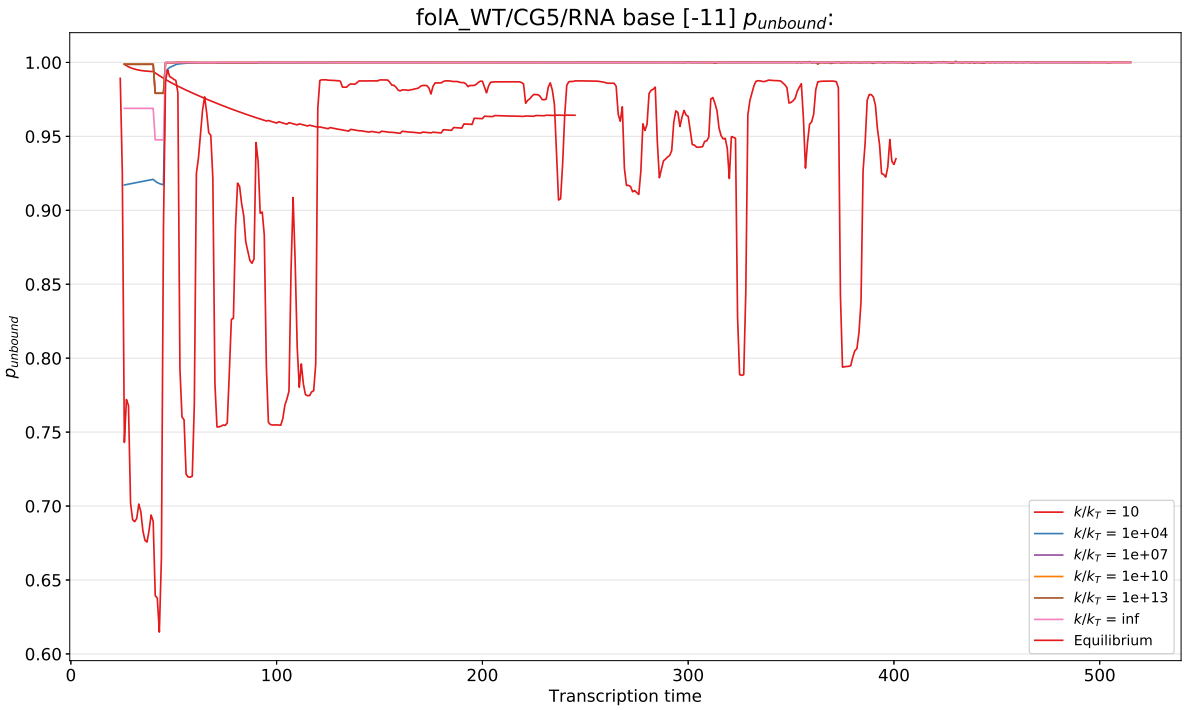


Figure 7

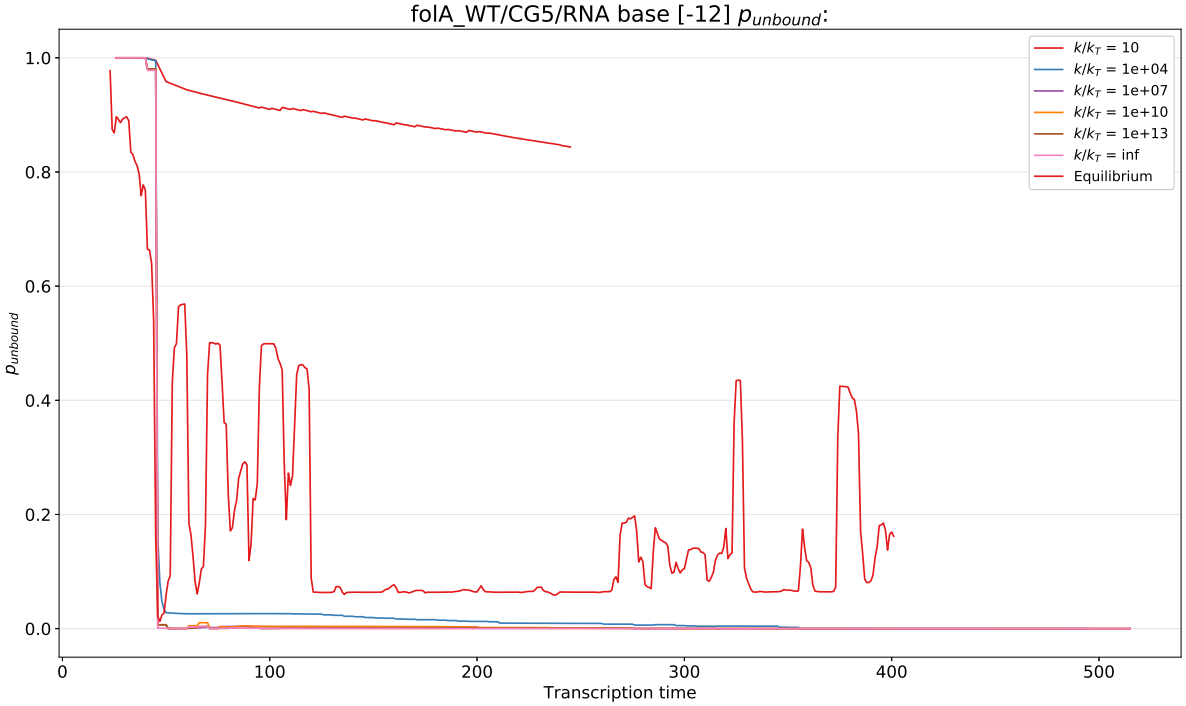


Figure 8

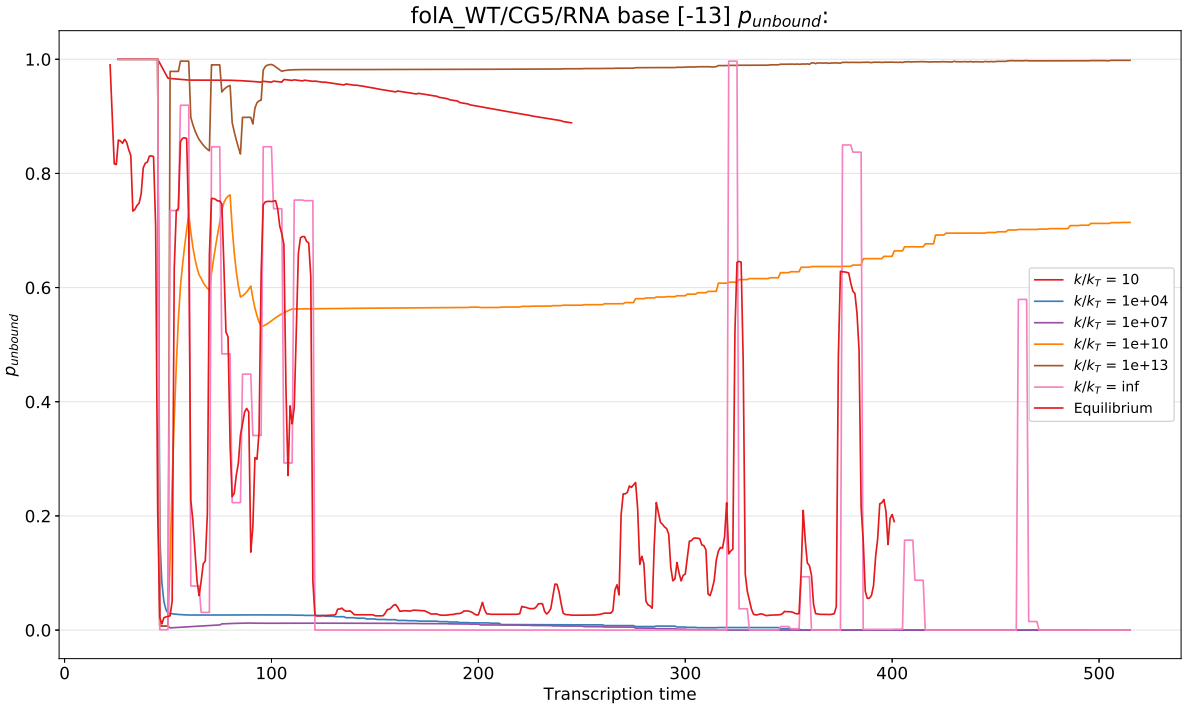


Figure 9

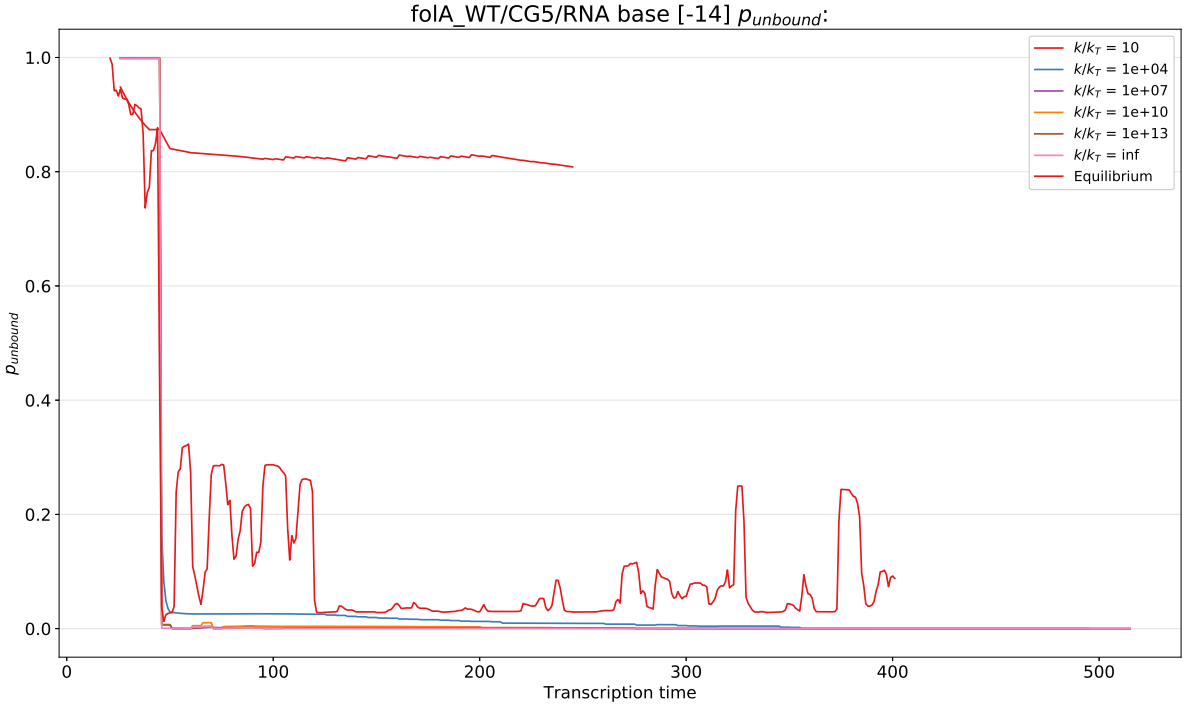


Figure 10

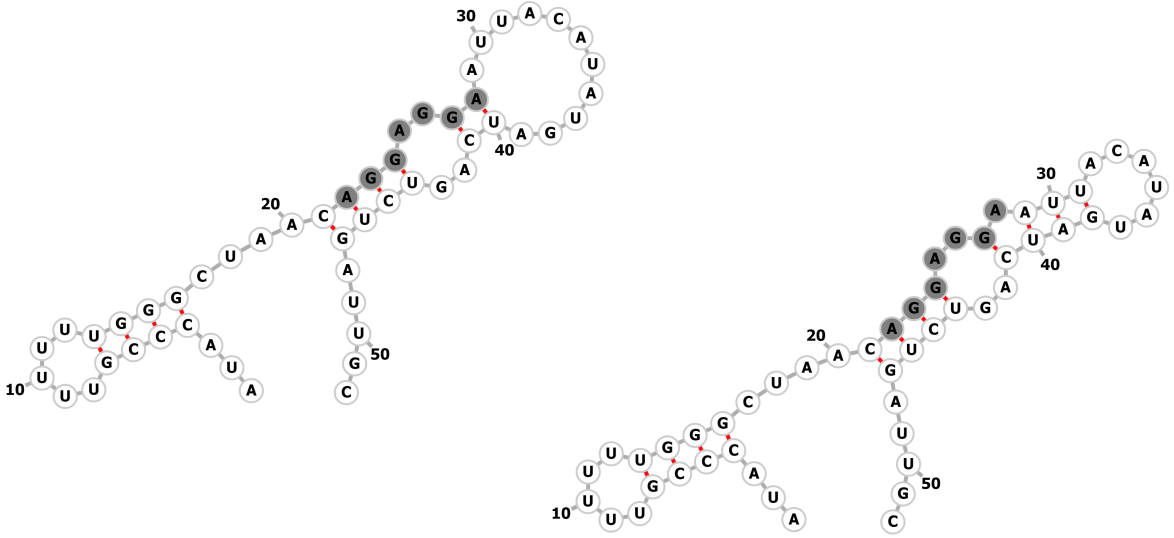


Figure 11: Tentative kinetically important sub-optimum structures (dG = 0.7 kcal/mol).

References

[1] A H Bayegan and P Clote. Minimum length RNA folding trajectories. Technical report.

[2] Peter Clote and Amir H Bayegan. RNA folding kinetics using Monte Carlo and Gillespie algorithms *. Technical report.

[3] Alexander P Gultyaev, F H D Van Batenburg, and Cornelis W A Pleij. The Computer Simulation of RNA Folding Pathways Using a Genetic Algorithm. Technical report, 1995.

[4] Raviprasad Aduri, Brian T. Psciuk, Pirro Saro, Hariprakash Taniga, , H. Bernhard Schlegel, and Jr.* John SantaLucia. AMBER Force Field Parameters for the Naturally Occurring Modified Nucleosides in RNA. 2007.

[5] Ting-ting Sun, Chenhan Zhao, and Shi-Jie Chen. Predicting Cotranscriptional Folding Kinetics For Riboswitch. 2018.

[6] Michael T Wolfinger, W Andreas Svrcek-Seiler, Christoph Flamm, Ivo L Hofacker, and Peter F Stadler. Exact Folding Dynamics of RNA Secondary Structures Dynamics of RNA Folding 2. Technical report.