

Work Report

Zhuoran Qiao

August 15, 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 and multiple coarse grained energy landscape sampling methods built on Markov state model. 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: $D_{A,B}^{foldon} = \text{MFE}(\text{sequence}[A,B])$. Note that $'.'$ is a trivial example of foldon. Our assumption 1 can be rewritten as

$$D_{A,B} = D_{A,i_1}^{foldon} \oplus D_{i_1,i_2}^{foldon} \oplus \dots \oplus D_{i_n,B}^{foldon} \quad (1)$$

Where \oplus represents a link operation. Note that all structural information of $D_{A,B}$ is encoded by the sequential representation $[A, i_1, \dots, i_n, B]$; 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: $\forall k, D_{i_k,i_{k+1}}^{foldon} \oplus D_{i_{k+1},i_{k+2}}^{foldon} \neq D_{i_k,i_{k+2}}^{foldon}$. Then the sufficient and necessary condition for structural rearrangement is

$$\begin{aligned}
\langle D_{A,B}^u | \hat{\mathbf{T}} | D_{A,B}^v \rangle &\neq 0 \text{ if and only if } \exists i, j \text{ satisfies} \\
&i, j \in D_{A,B}^u \cdot \text{IFR}, i, j \in D_{A,B}^v \cdot \text{IFR}; \\
&D_{A,i}^u = D_{A,i}^v, D_{j,B}^u = D_{j,B}^v; \\
&D_{i,j}^u = D_{i,j}^{\text{foldon}} \text{ or } D_{i,j}^v = D_{i,j}^{\text{foldon}}. \\
\text{Then } \langle D_{A,B}^u | \hat{\mathbf{T}} | D_{A,B}^v \rangle &= \langle D_{i,j}^u | \hat{\mathbf{T}} | D_{i,j}^v \rangle.
\end{aligned}$$

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. Multiple methods to rigorously calculate the maximum likelihood during RNA folding have been reported; 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 $k_{uv} = \langle D_{A,B}^u | \hat{\mathbf{T}} | D_{A,B}^v \rangle$ is calculated by Arrhenius approximation $k_{uv} = k_0 \exp \left[-\frac{1}{RT} (\Delta G_u^{\text{Stack}} + \Delta G_v^{\text{Loop}}) \right]$. '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 different population is updated. New candidate strands $D_{0,L+\Delta L}^{\text{Candidate}}$ with length $L + \Delta L$ are generated by a recombination process: for every old strand $D_{0,L}^{\text{Strand}}$, all indices in its IFR is identified as possible rearrangement site, then its child strands are generated by linking partial domains $D_{0,\text{Site}}^{\text{Strand}}$ with a foldon $D_{\text{Site},L+\Delta L}^{\text{foldon}}$ that terminated at $L + \Delta L$.

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

After structural 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 Results

The only remaining free parameter to be determined is k_0/k_T , the ratio of pre-exponential factor in Arrhenius rate formulation and transcription 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 calculate 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 is very frequent at early stage of transcription,

Algorithm 1 Co-transcriptional folding elongation procedure

```

1: Initialize ActivePool
2: while sequence length > current length do
3:   OldPool  $\leftarrow$  ActivePool
4:   renew ActivePool
5:   Current length  $\leftarrow$  Current length +  $dL$ 
6:    $dt \leftarrow dL/k_T$  ▷ Transcription time
7:   for left boundary  $\in \{0, dL, 2dL, \dots, \text{Current length} - dL\}$  do ▷ Get all new foldons
8:      $D_{\text{left boundary}, \text{Current length}}^{\text{foldon}} \leftarrow \text{numpy.mfe}(\text{sequence}[\text{left boundary}, \text{Current length}])$ 
9:   end for
10:  for Strand  $\in$  OldPool do ▷ Recombination
11:    for Site  $\in$  Strand.IFR do
12:       $D_{0, \text{Current length}}^{\text{Candidate}} \leftarrow D_{0, \text{Site}}^{\text{Strand}} \oplus D_{\text{Site}, \text{Current length}}^{\text{foldon}}$ 
13:      if  $D_{0, \text{Current length}}^{\text{Candidate}} \in \text{ActivePool}$  then
14:        update  $D_{0, \text{Current length}}^{\text{Candidate}}$ .IFR
15:      else
16:        add  $D_{0, \text{Current length}}^{\text{Candidate}}$  to ActivePool
17:      end if
18:      if site = Current length -  $dL$  then
19:         $\langle \text{ActivePool.population} | D_{0, \text{Current length}}^{\text{Candidate}} \rangle \leftarrow \langle \text{OldPool.population} | D_{0, \text{Site}}^{\text{Strand}} \rangle$ 
20:      end if
21:    end for
22:  end for
23:  for  $D_{0, \text{Current length}}^u \neq D_{0, \text{Current length}}^v \in \text{ActivePool}$  do ▷ Calculate new rate matrix
24:    calculate  $D_{\text{rearrange}}^u, D_{\text{rearrange}}^v$  ▷ Find all helices involved in rearrangement
25:     $\langle D_{\text{rearrange}}^u | \hat{\mathbf{T}} | D_{\text{rearrange}}^v \rangle \leftarrow k_0 \exp\left(-\frac{1}{RT}(\Delta G_u^{\text{Stack}} + \Delta G_v^{\text{Loop}})\right)$ 
26:  end for
27:   $\langle \text{ActivePool.population} | \leftarrow \langle \text{ActivePool.population} | \exp\left(t \times \hat{\mathbf{T}}\right)$  ▷ Master equation
28:  reserve top  $N$  populated strands in ActivePool ▷ Selection
29:  renormalize  $\langle \text{ActivePool.population} |$ 
30: end while

```

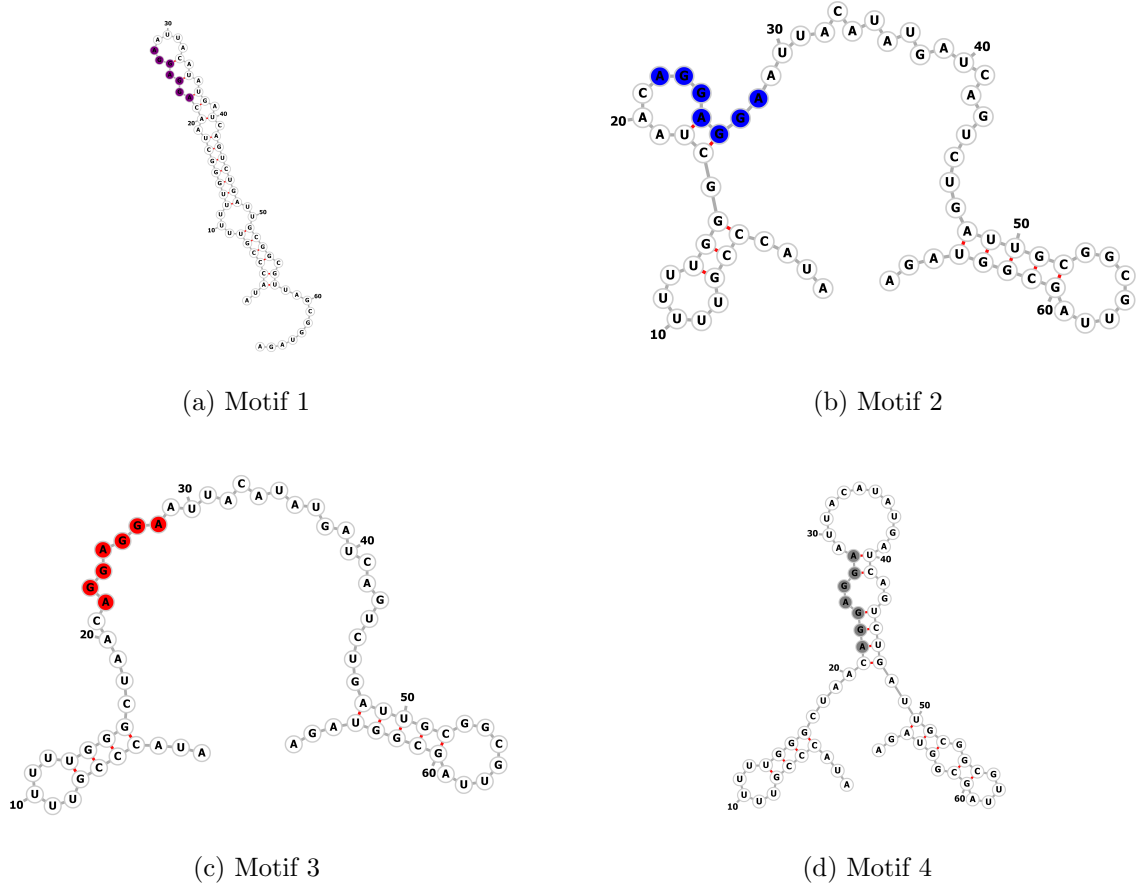


Figure 1: Exemplary ss containing folding motifs within folA-WT Shine-Dalgarno sequence

indicating the sensitivity of local structures on long-range contacts, and dependence of motif predominance on the limited timescale for folding.

$p_{unbound}$ analysis We calculated $p_{unbound}$ with respect to transcription time and k_0/k_T (Figure 3-10). Deviation of asymptotic behavior from equilibrium value (calculated by nupack.ppairs) is possibly due to the limited set of foldons (only used mfe to obtain current results).

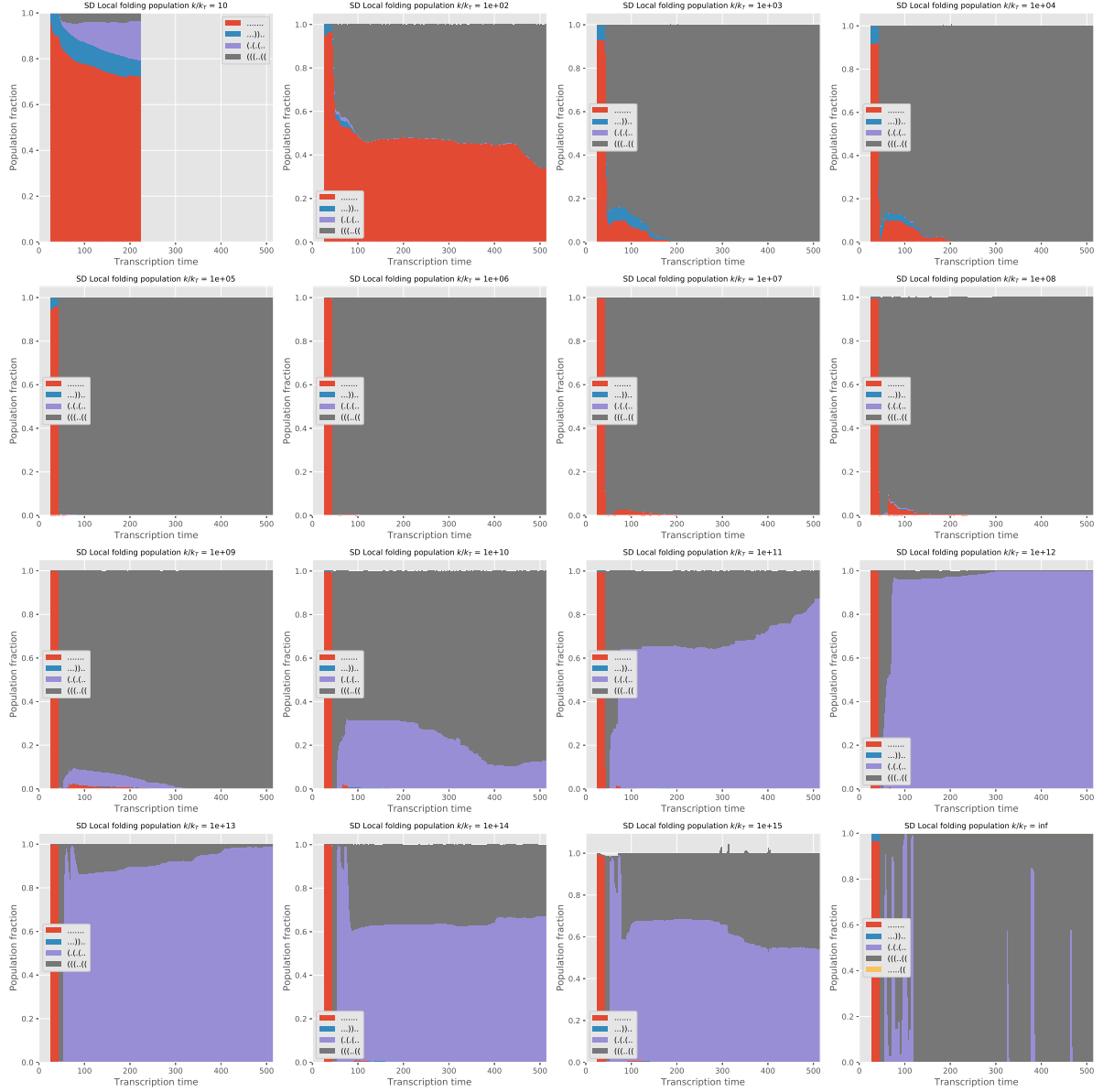


Figure 2: Population dynamics of four structural motifs during co-transcriptional folding.

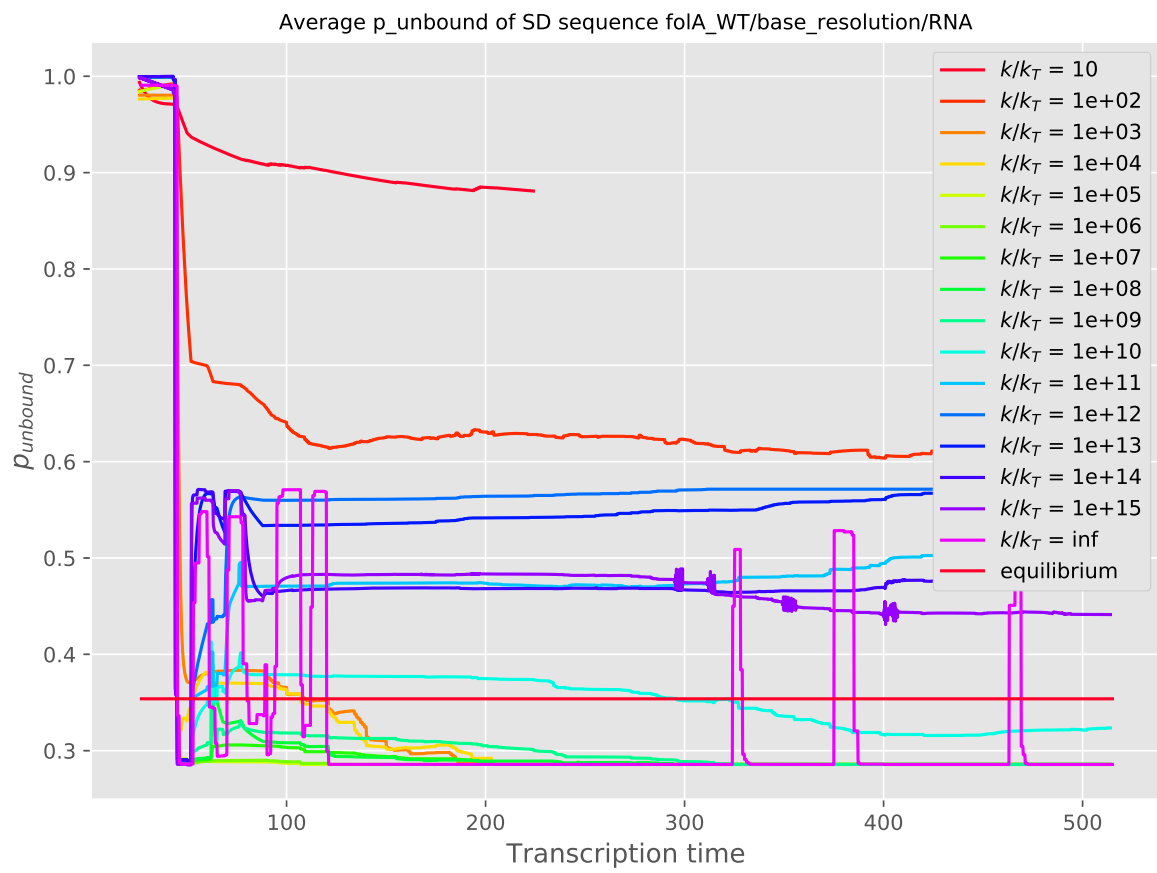


Figure 3

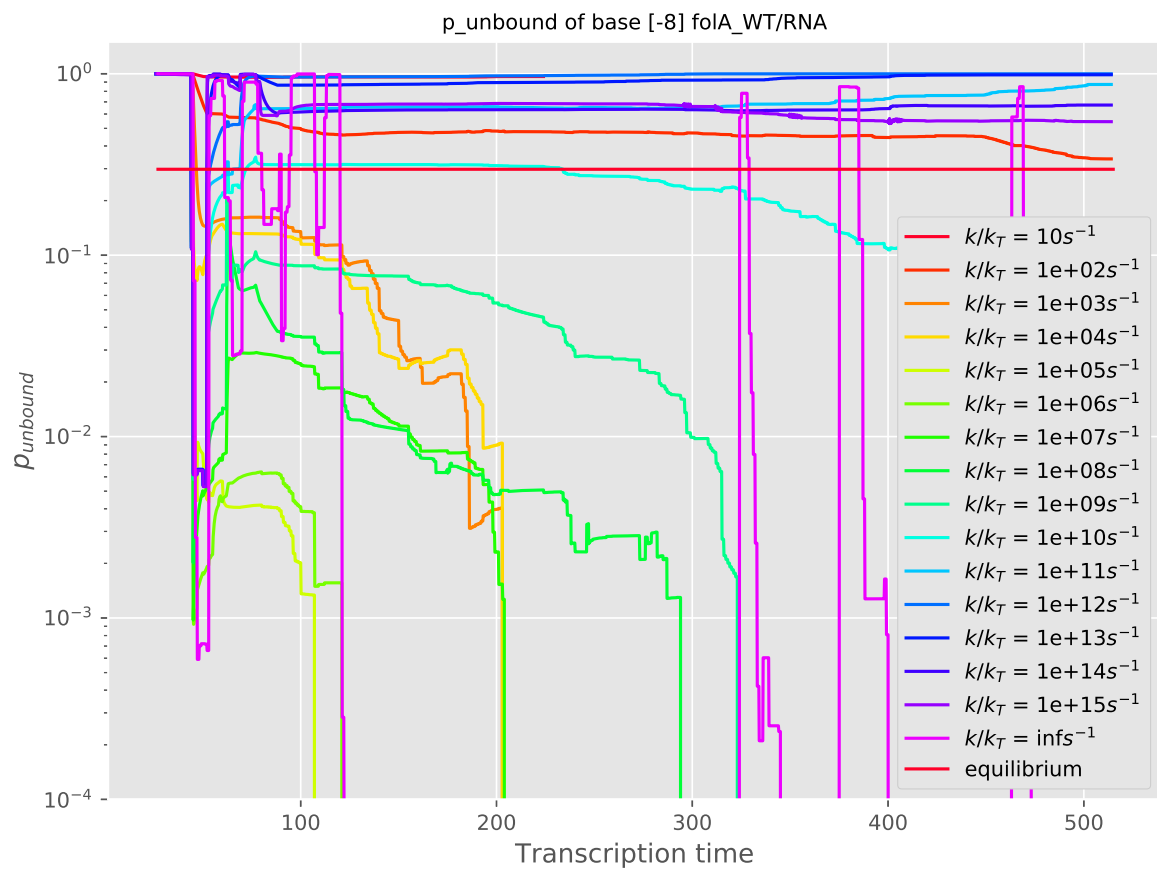


Figure 4

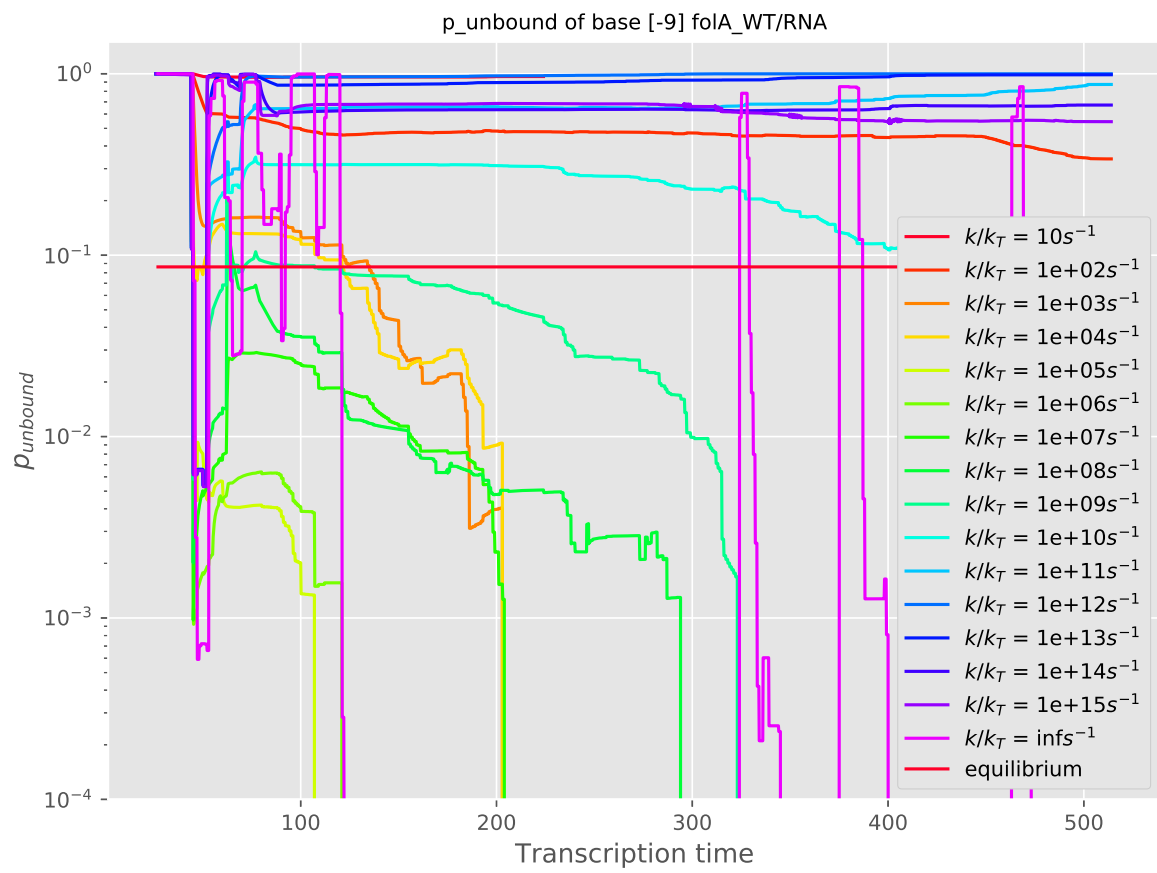


Figure 5

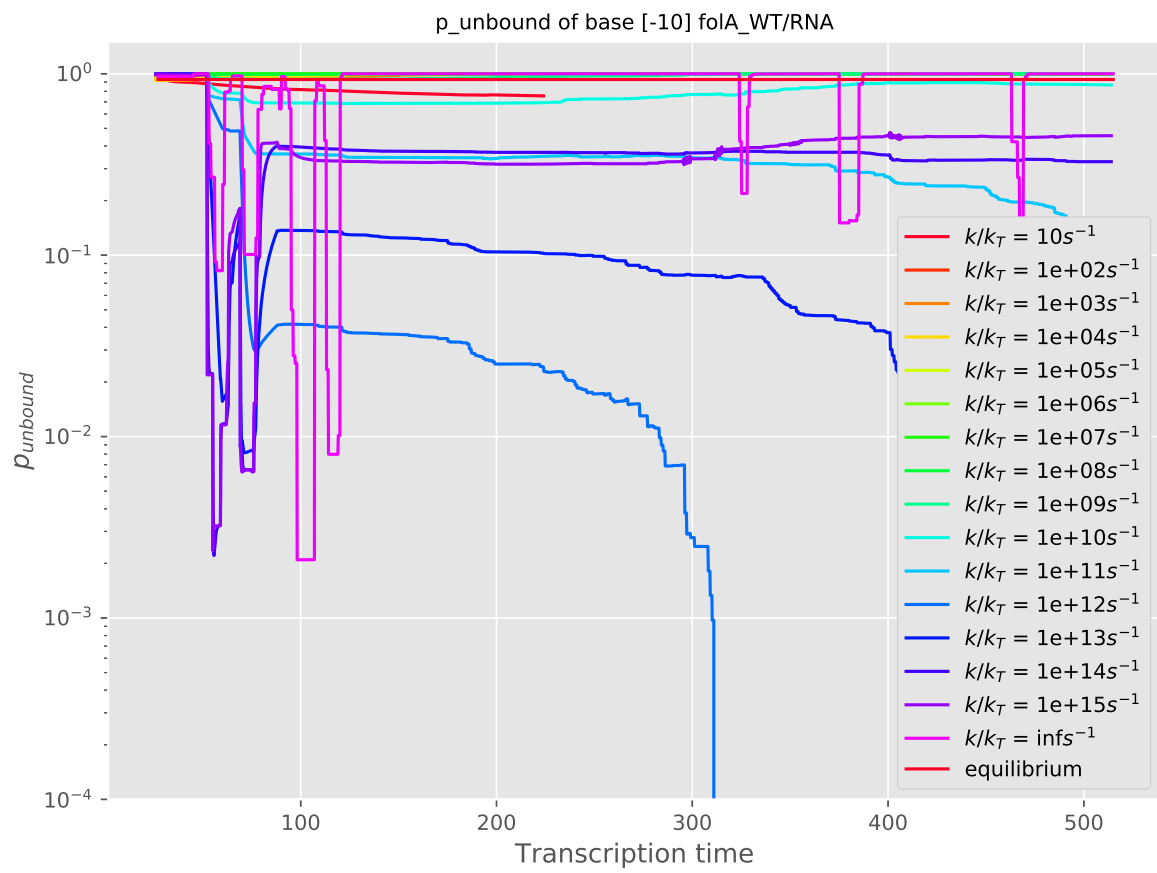


Figure 6

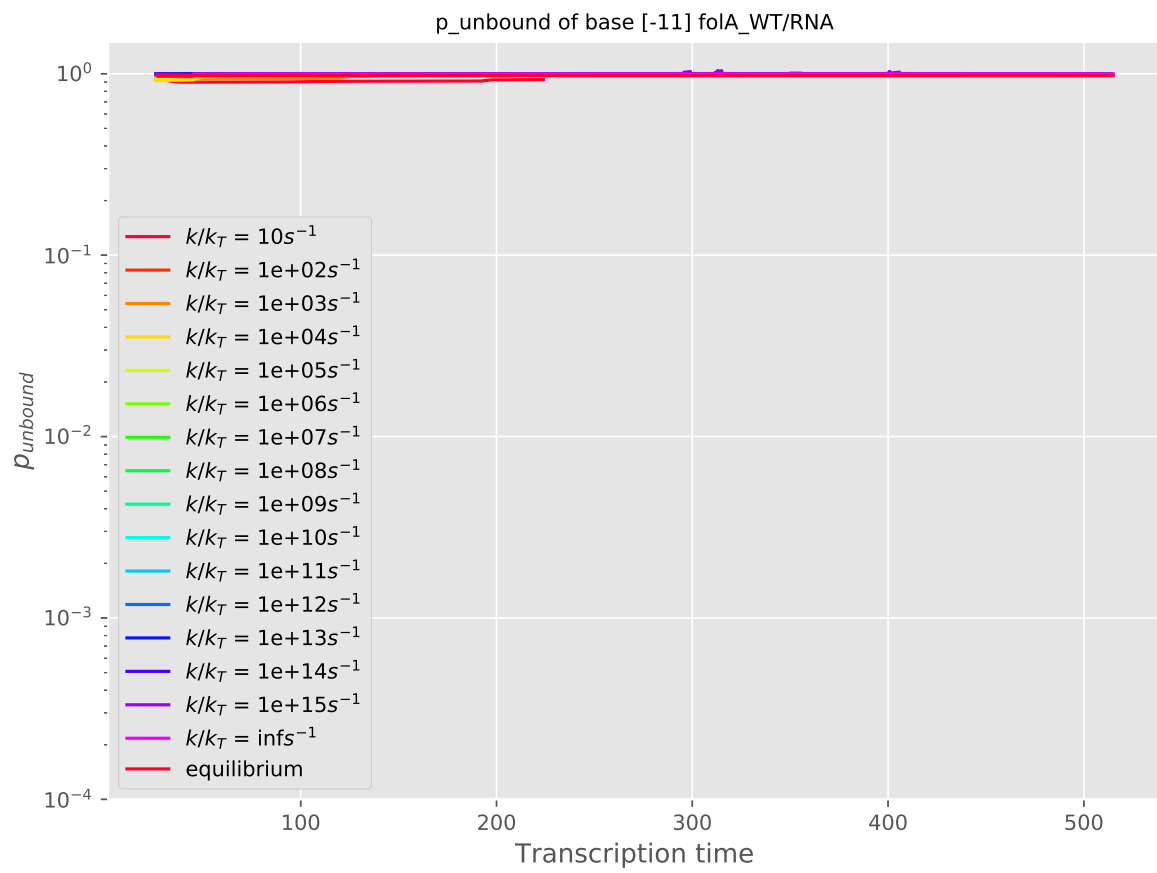


Figure 7

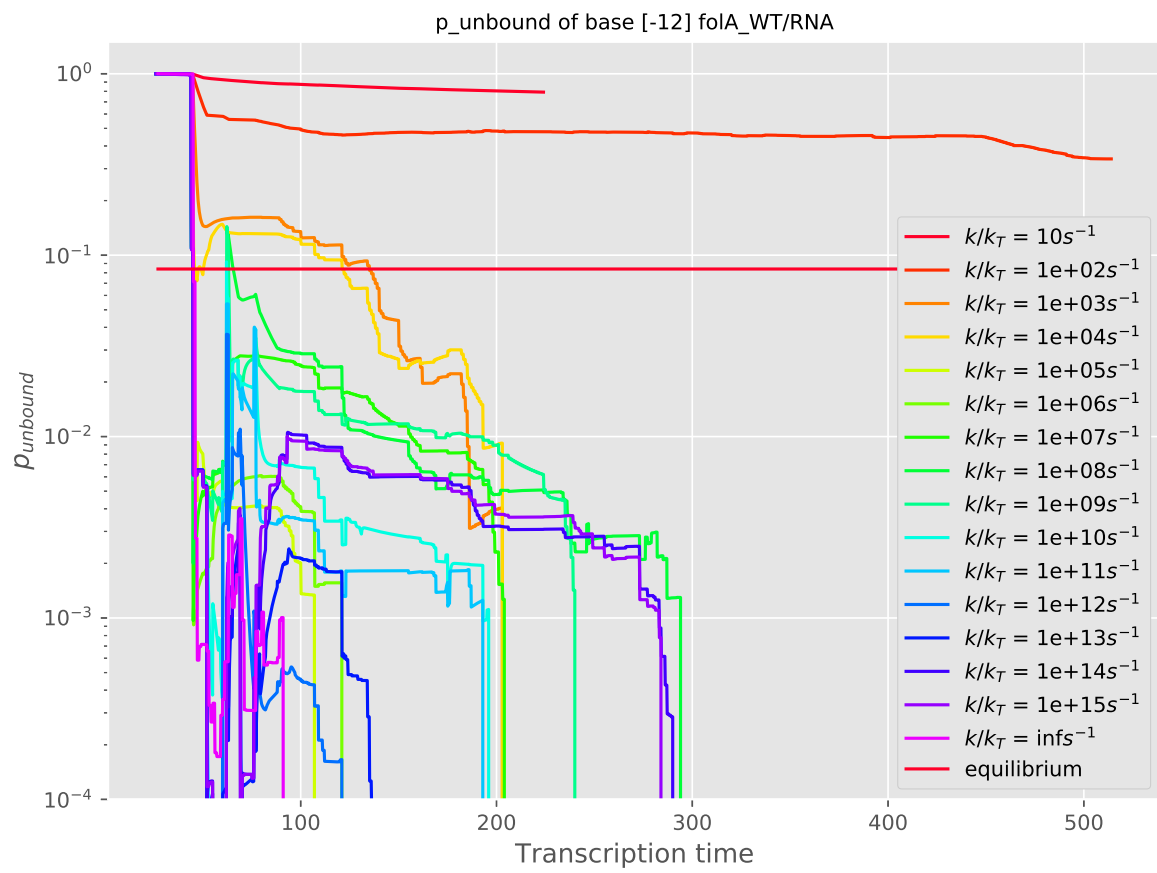


Figure 8

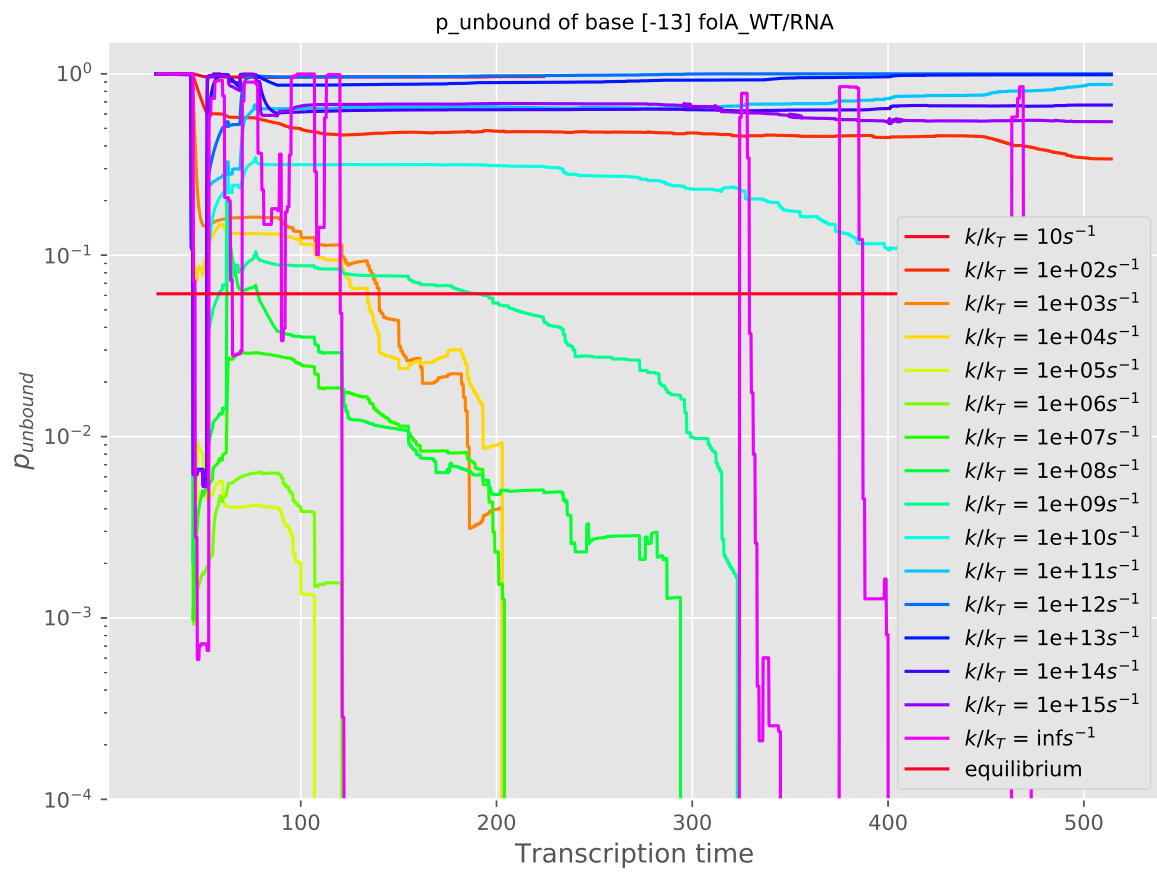


Figure 9

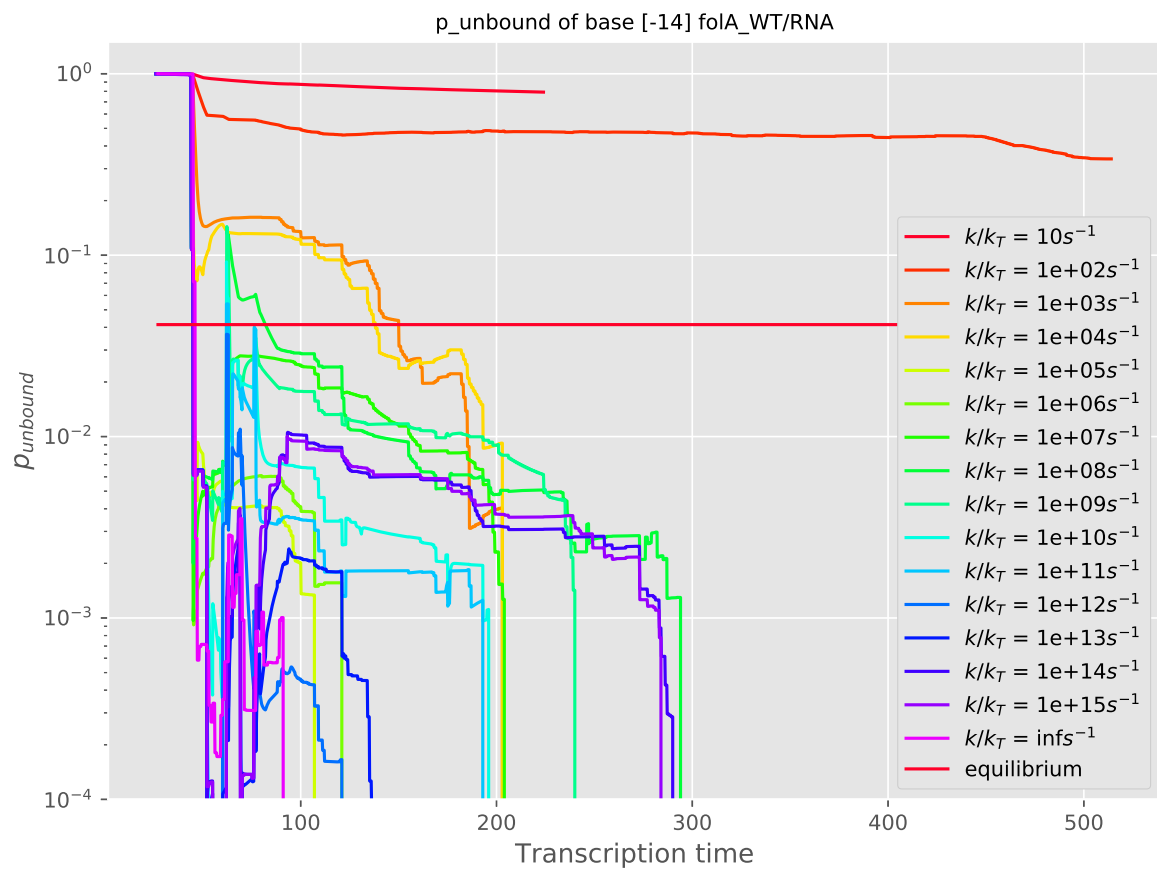


Figure 10