Non-equilibrium RNA Folding as an Evolutionary Basis of E. Coli 5'-UTR Codon Bias

Zhuoran Qiao

October 18, 2018

1 Introduction

1 Developed GenoFold, a novel approach to simulate kilobases level co-transcriptional folding kinetics.

2 Quantitatively investigated the dependence of transcient Shine-Dalgarno accessibility on mRNA level variation within a synonymous mutant library.

2 Method development

2.1 Framework

Various algorithms or programs have been developed to predict RNA folding pathway utilizing force-field based simulations[1] and multiple sampling methods based on Monte Carlo trajectories[2][3] or coarse graining of energy landscape built on Markov state model[**?**][4]. 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 GenoFold, a genetic algorithm and Markov state model (MSM) based approach, which is capable of capturing kilobase level kinetics. This method is composed of a recombination step to generate configuration states in the MSM from historical information, a propagation step to calculate configuration probability distribution evolution, and a selection step to reduce the total size of configuration states set. Schematic illustration of this method is depicted (Figure 1).

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 different 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.

The folding configuration space could be discretized to above-mentioned SS sets to build a Markov state model. 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 (sub)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

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

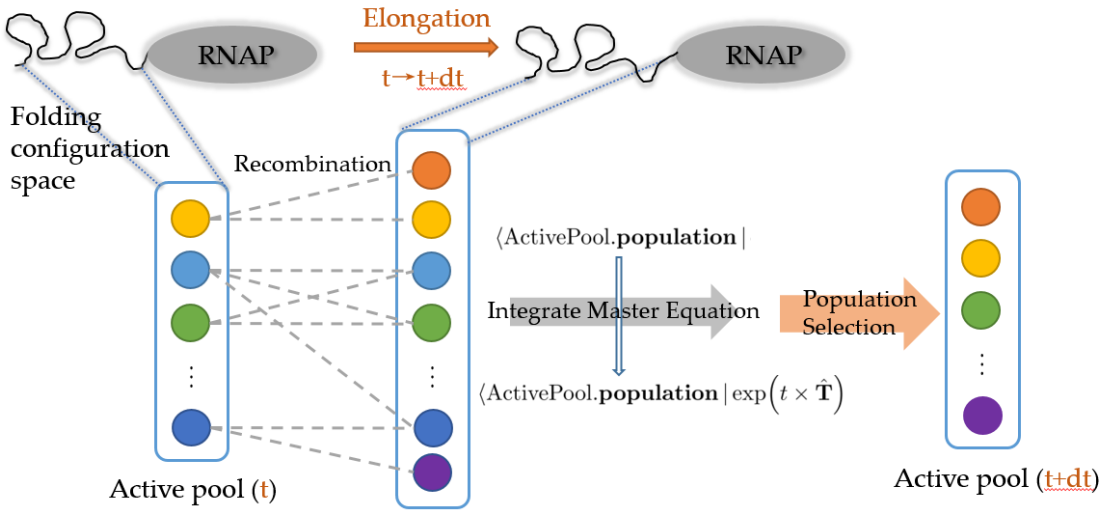


Figure 1: Schematics of the GenoFold workflow.

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

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

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 Construct the rate matrix of MSM

We then proposed following ansatz to estimate the generator (rate) matrix in the MSM once the set of discretized folding configurations is determined. 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 structures have been reported[**?**]. Here, we proposed a computationally feasible approach: the forward free energy barrier is estimated by summing up all free energy associated with old stacks unzipping and new helices 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. We note that this prediction could be improved by considering different rearrangement tendencies of stem and internal loops [4].

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.

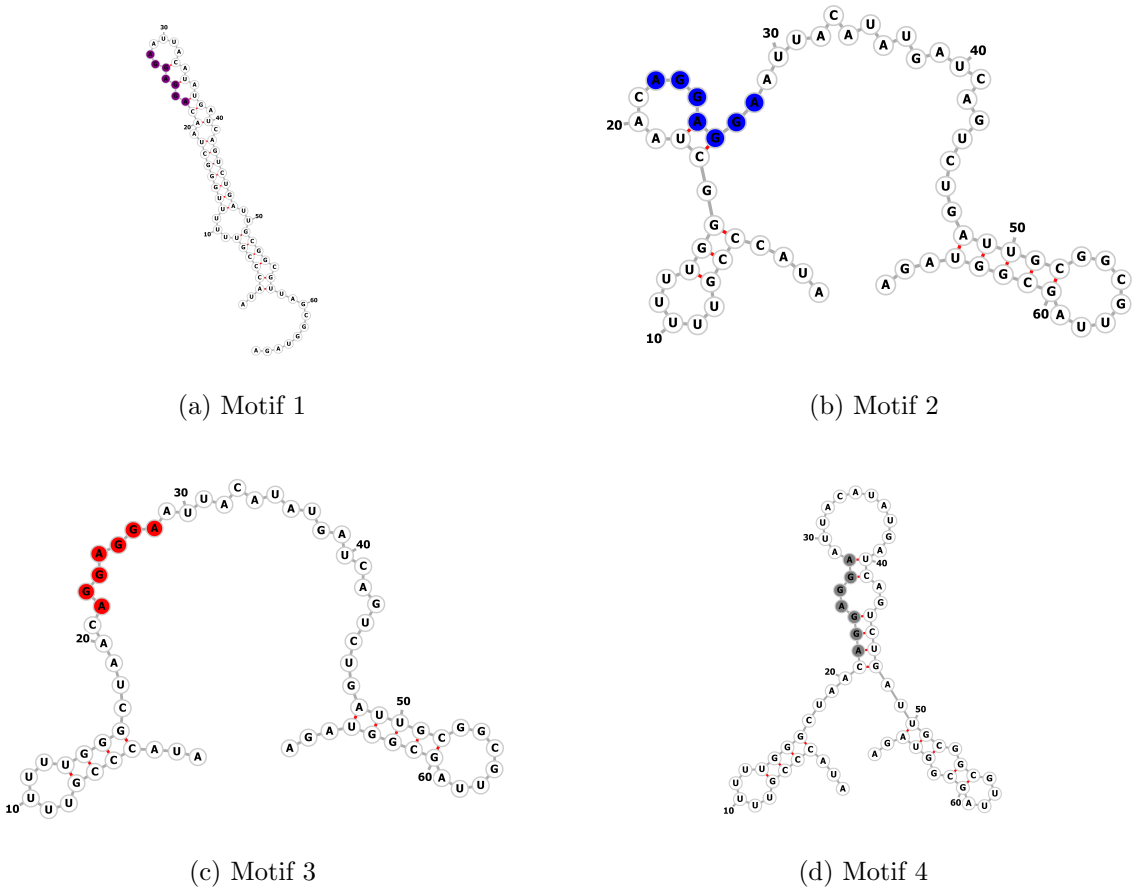
After structual generation the rate matrix among all candidate strands within the new active species pool is calculated (see part 2.2). Then the population distribution of strands after elongation is computed by propagate the chemical master equation.

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 GenoFold simulation procedure are shown in Algorithm 1.

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 $D^{\mathrm{u}}_{0,\text{Current length}} \neq D^{\mathrm{v}}_{0,\text{Current length}} \in$ 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 2: Exemplary ss containing folding motifs within folA-WT Shine-Dalgarno sequence

2.4 Test and optimization of GenoFold

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 2 shows exemplary secondary structures containing these motifs; figure 3 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 2. Surprisingly we noticed that when $k_0/k_T = \infty$, exchange between energetially favorable motifs was very frequent at early stage of transcription, indicating the sensitivity of local structures on long-range contacts. We also noticed that motif predominance after transcription strongly depended on folding rate, reiterating the importance of time scale in the folding problem.

$p_{unbound}$ **analysis.** We calculated $p_{unbound}$ with respect to transcription time and k_0/k_T (Figure 4). As a data test we used nupack.ppairs to calculate equilibrium $p_{unbound}$ for all truncated sequences. Deviation of asymptotic behavior of model from equilibrium value is possiblly due to the limited set of foldons (only used mfe to obtain current results).

Calibrate pre-exponential factor from TCSPC data. We then resorted to the time scale of hairpin refolding to calibrate the absolute value of k_0. Specifically, we observed the relaxation corresponding to folding of the motif involving S-D sequence when $\frac{k_0}{k_T}$ is between 10^2 and 10^3

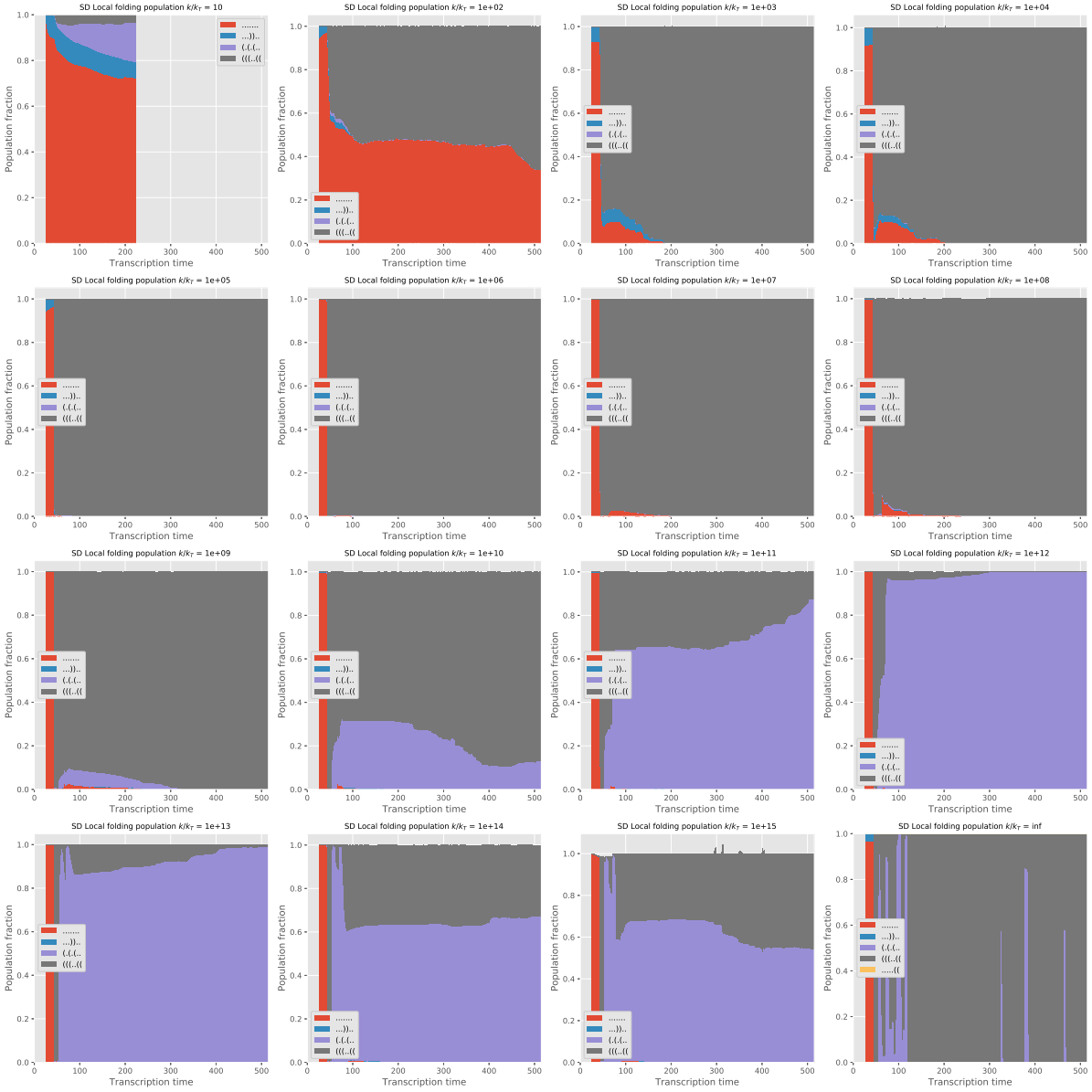
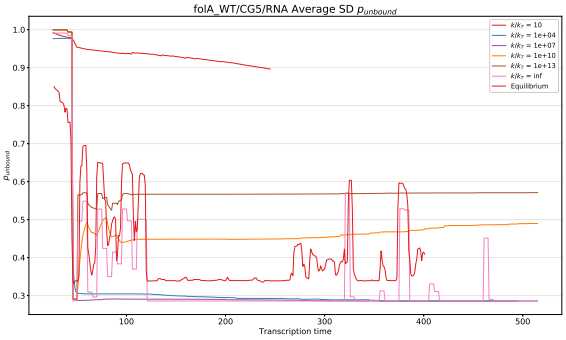
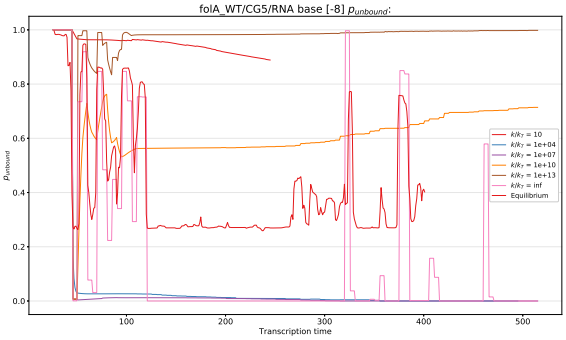


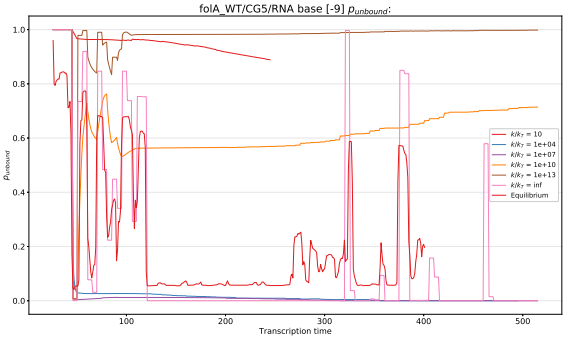
Figure 3: Population dynamics of four S-D structrual motifs during co-transcriptional folding.



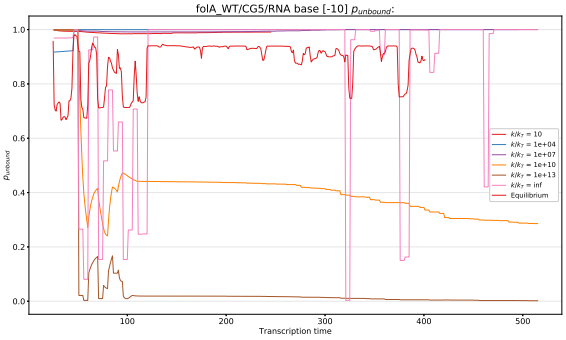
(a) folA$_{\mathrm{WT}}$ average $p_{unbound}$ results.



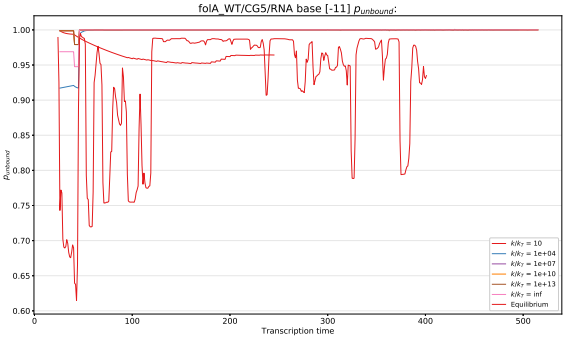
(b) folA$_{\mathrm{WT}}$ base[-8] $p_{unbound}$ results.



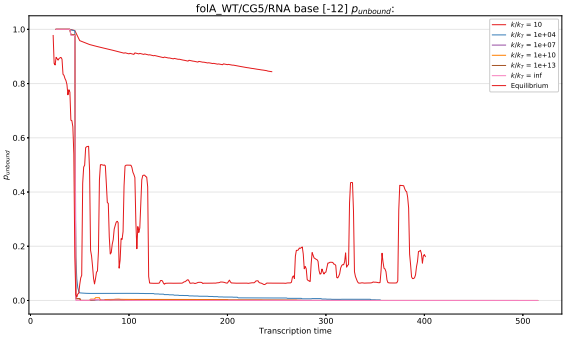
(c) Same as above, base -9.



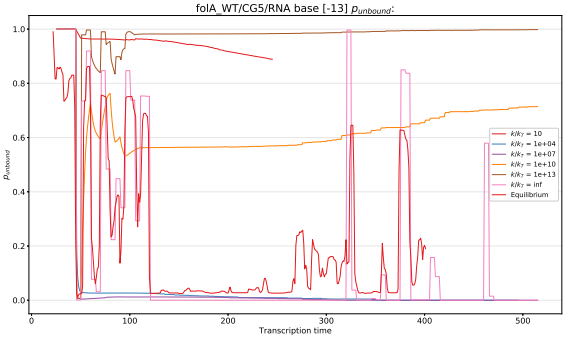
(d) Same as above, base -10.



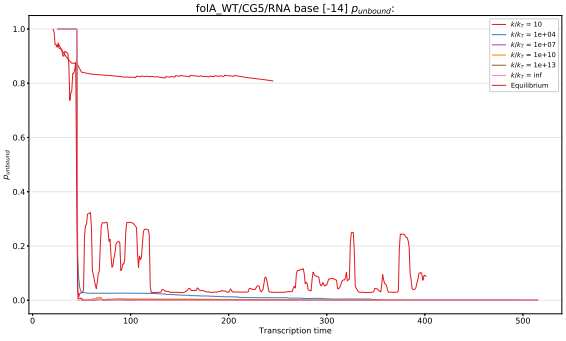
(e) Same as above, base -11.



(f) Same as above, base -12.



(g) Same as above, base -13.



(h) Same as above, base -14.

Figure 4: Shine-Dalgarno average and single-base $p_{unbound}$ during co-transcription folding of folA$_{\mathrm{WT}}$ mRNA (after model optimization).

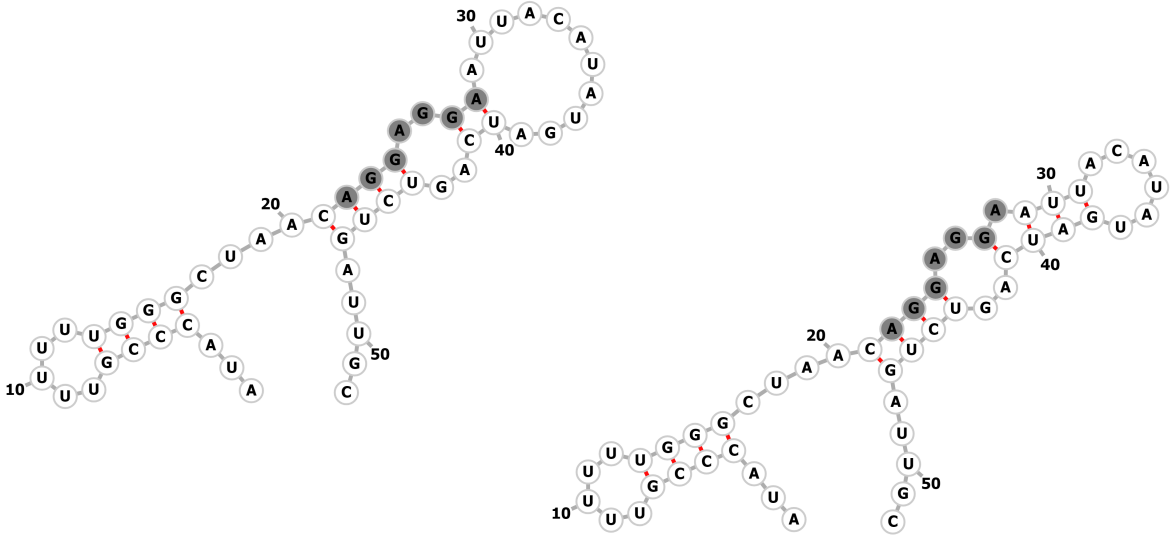


Figure 5: Tentative kinetically important sub-optimum structures (dG = 0.7 kcal/mol).

(Figure 3); From TCSPC measurement data in [], we note that a similar 20 nt hairpin takes about 70 ns for refolding, commensurable to a scheme where $k_T = 10^9\ s^{-1}$. Consequently, we estimated the value of pre-exponential factor to be $k_0 = 10^{11}$.

Model optimization. We observed that our model tends to overesitimate the flucuation of average $p_{unbound}$ when $k = \infty$, and some kinetic patterns were lost (for bases -11, -12 and -14). Specifically we found that for base -8 and base -9, our model gave identical results while for equilibrium nupack.ppairs calculation there was a overall upshift of $p_{unbound}$.

Then we examined sub-optimal structures using nupack.subopt for sequence truncated at 150 nt, and found two tentative kinetically important structures between which free energy difference is 0.7 kcal (Figure 5). Coexistence of those sub-optimal structures instead of only the minimum free energy structure will possibly result in similar population dynamics as above but different $p_{unbound}$ at single base resolution. We also noticed that for mininum free energy motifs, base -11, -12 and -14 shared the same pairing patterns, meaning that $p_{unbound}$ on these bases has no response to motif population dynamics.

We optimized foldon collection sets by incorporate sub-optimal structures with a energy gap of 1cal/mol by nupack.subopt algorithm and reconducted the computation for folA_{WT}, setting ActivePool cutoff = 50 (Figure 7). It's worth noting that after optimization predicted equilibrium $p_{unbound}$ values agreed much better with ensemble averaged $p_{unbound}$ (computed by nupack.ppairs, not shown here), and base-specificity of $p_{unbound}$ can be correctly reproduced.

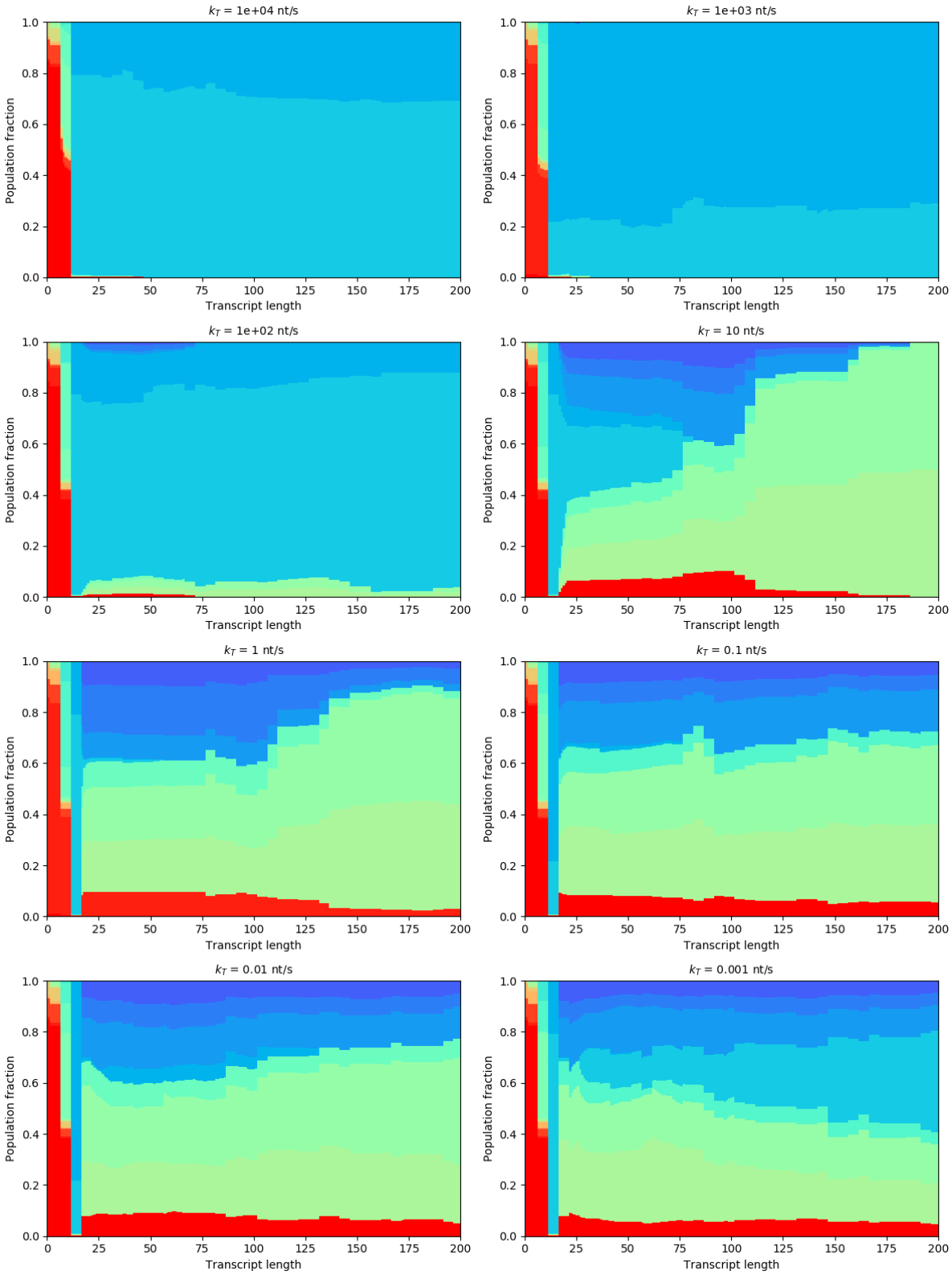
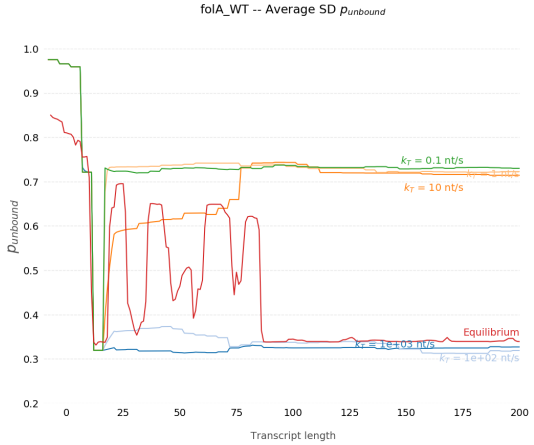
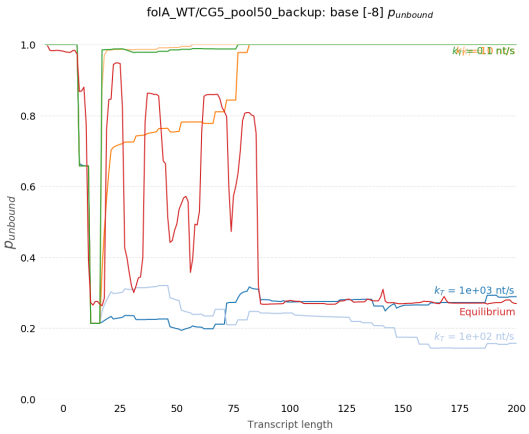


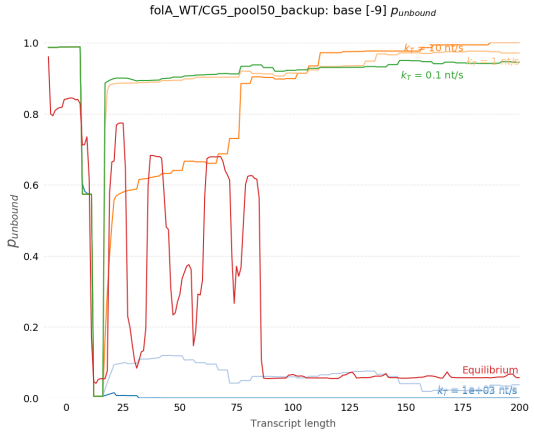
Figure 6: Population dynamics of folA$_{\mathrm{WT}}$ S-D structrual motifs during co-transcriptional folding under varied k_T (after calibration and model optimization).



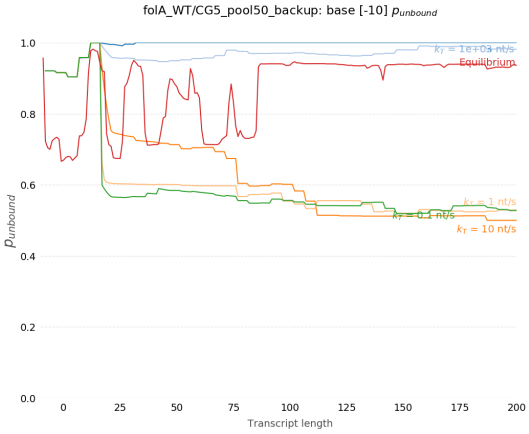
(a) Revised folA$_{\mathrm{WT}}$ average $p_{unbound}$ results.



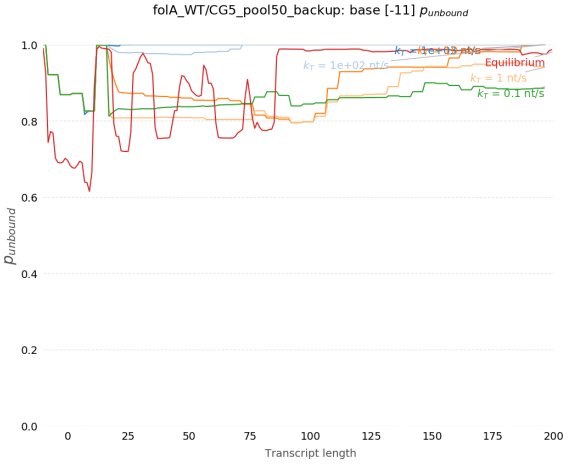
(b) Revised folA$_{\mathrm{WT}}$ base[-8] $p_{unbound}$ results.



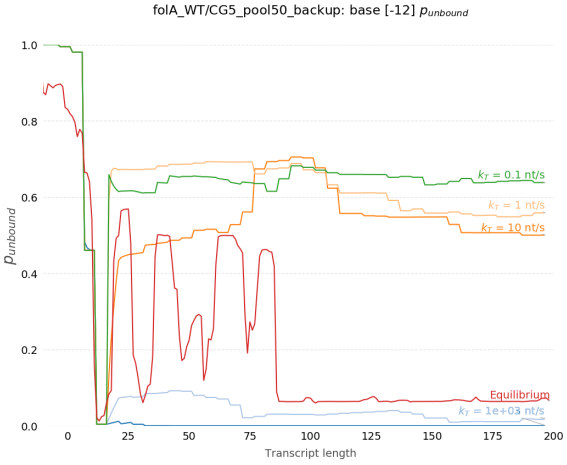
(c) Same as above, base -9.



(d) Same as above, base -10.



(e) Same as above, base -11.



(f) Same as above, base -12.

Figure 7: Shine-Dalgarno average and single-base $p_{unbound}$ during co-transcription folding of folA$_{\mathrm{WT}}$ mRNA (after model optimization).

3 Results and discussion

3.1 Transcription rate dependence on co-transcriptional mRNA folding kinetics.

We then tuned the RNA polymerase transcription rate away from physiological condition, from $k_T = 1\,nt \times 10^{-7}\,s^{-1}$ to $k_T = 1\,nt \times 10^{7}\,s^{-1}$ and compared the behavior of $p_{unbound}$ during co-transcription folding (Figure 7, a). It's worth noting that before the transcription of base 20, $p_{unbound}$ value of all kinetic schemes remained close to equilibrium asymptotic results, which was analogous to a scheme of infinitely slow transcription. Once downstream region is transcribed, however, the $p_{unbound}$ value started to diverge.

This divergence reflected an altered Shine-Dalgarno folding stability after the formation of downstream structures, allostericly shifting from a minima in local free energy surface to a sub-optimum which contributed to globally more stable secondary structures. The kinetic dependence was thus triggered by the barrier crossing dynamics between the local folding free energy minima to another favorable structure in the configuration space of longer mRNA intermediate segments.

3.2 mRNA level - transcient $p_{unbound}$ correlation within a synonymous mutants library.

We repeated GenoFold folding prediction for all synonymous sequences in [], which contains synonymous codon substitutions for gene folA and Adk in both N-terminal and downstream region. As a initial test we calculated the correlation between transcient averge Shine-Dalgarno $p_{unbound}$ when base 100 is transcribed and measured mRNA levels among the synonymous mutants library. Comparing to the equilibrium model (Figure 8, left) which predicted no $p_{unbound}$ dependence of mRNA level variation for sequences with synonymous substitution downstream in the coding region, the non-equilibrium model (Figure 8, right) can much better discern the sequence dependence of transcient S-D $p_{unbound}$ and desmonstrated a significant correlation between transcient S-D $p_unbound\,(base[100])$ and the log ratio of mRNA levels.

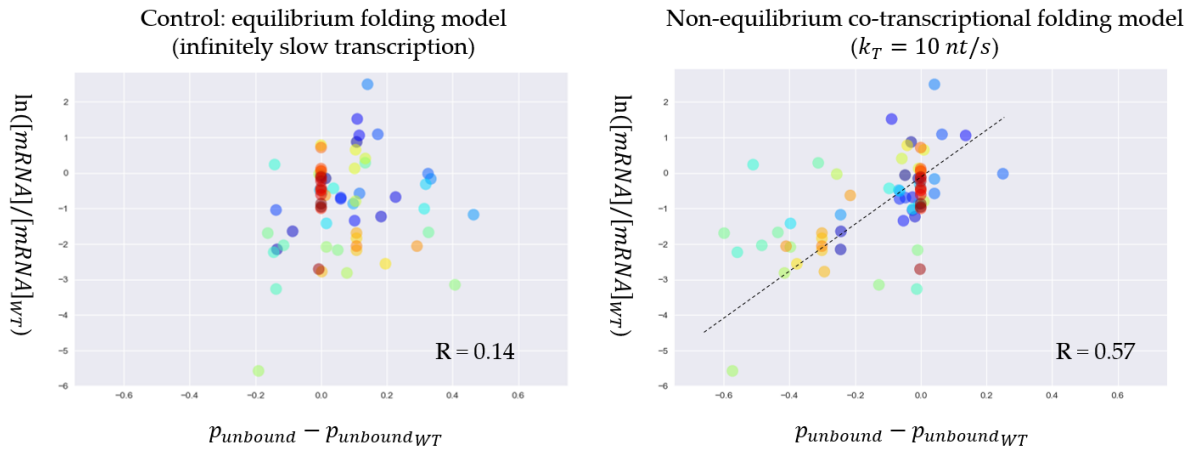


Figure 8: Correlation between transient average Shine-Dalgarno $p_{unbound}$ when base 100 is transcribed and measured mRNA levels.

Then we intended to capture the characteristic time scale of ribosome binding during transcription. We varied the snapshot transcription time taken to calculate transient $p_{unbound}$ and trancription rate k_T, then summarized the result of measured mRNA level - transient S-D $p_{unbound}$ correlation coefficient (pearson R) at given snapshot transcription times and k_T (Fig-

ure 9). Strikingly we noticed that significant correlation could be observed only within diagonal elements, indicating kinetic schemes with commensurable snapshot and transcription time scale, reflecting a coupling between ribosome binding. This is consistent with the fact that mRNA levels are largely insensitive to codon substitutions when folA synonymous codon variants are cloned under the T7 rather than pBAD/E. Coli RNAP system. In addition to the discussion in [], it offered an complementary hypothesis that variation of ribosome binding affinities among synonymous codon mutants quantitatively diffrent between fast and slow transcription scheme, resulting in altered measured protein levels.

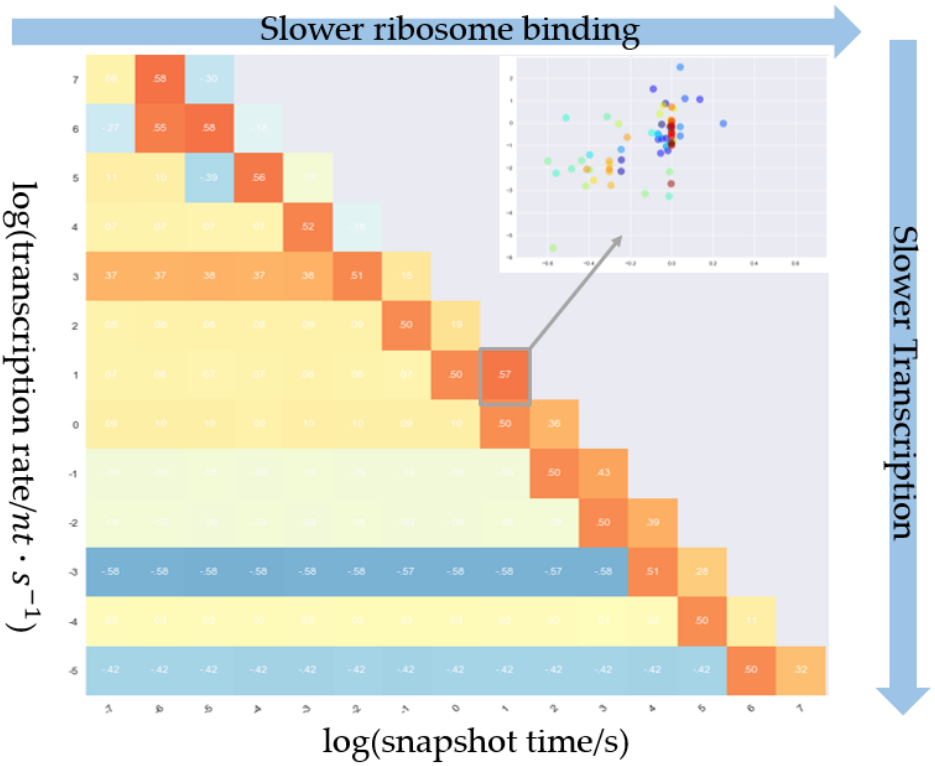


Figure 9: Correlation coefficients between S-D averaged transcient $p_{unbound}$ and mRNA level variation correlation in various transcription-ribosome binding kinetic schemes.

More detailedly we analyzed mRNA level - transient S-D averaged $p_{unbound}$ correlation after each base is transcribed with a bin of 5 bases (Figure 10). Significant correlation could be observed in N-terminal region over all kinetic schemes, which is possibly because of the similar refolding kinetics of N-terminal region. In contrast, only physiological kinetic scheme ($k_T = 10nt/s$) showed a strong correlation after the transcript is elongated to downstream region. On the contrary, equilibrium model showed no significant correlation downstream, reiterating that transcient S-D sequence accessibility variation is not an artifact of equilibrium folding stability.

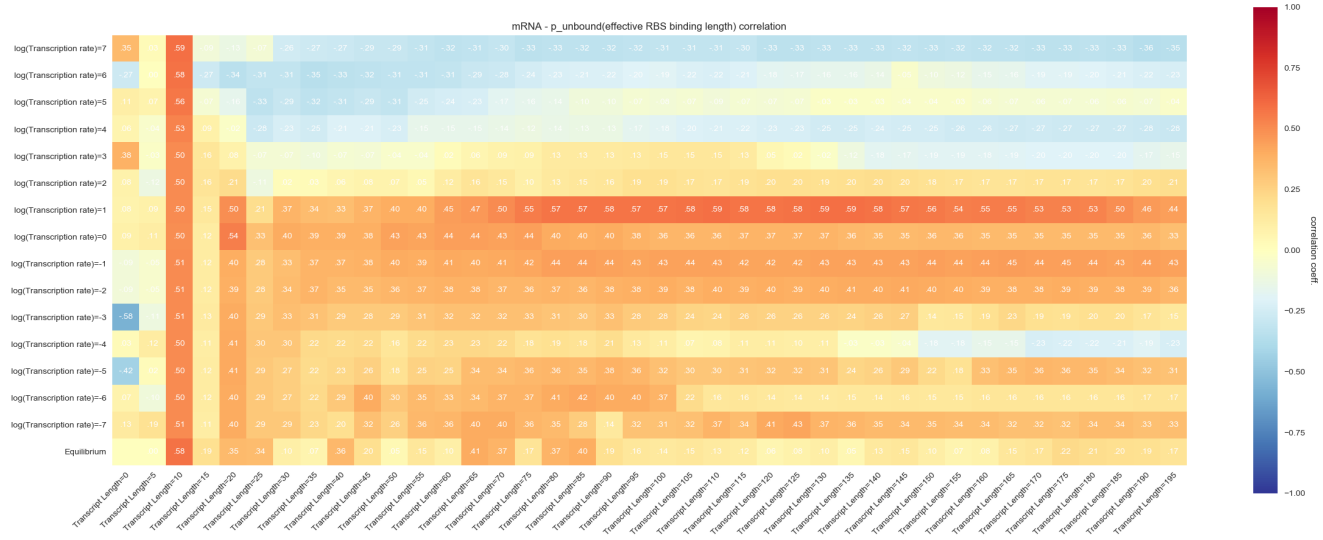


Figure 10: Correlation between transient S-D averaged $p_{unbound}$ - mRNA level when given bases is transcribed under different log ratio of k_T. Correlation is most significant when down-stream region is released under physiological transcription rate.

4 Future Plan

1 Obtain more accurate S-D accessibility indicator by calculating local domain unfolding free energy from transcient secondary structure distribution.

2 Develop kinetic models to address the mechanism underlying translation initiation rate – mRNA level correlation, including Rho-dependent transcription termination and RNA degradation.

3 Use genome and species-wide analysis to investigate the effect of transcription - translation initiation coupling on evolutionary selection pressure on codon usage.

References

[1] Raviprasad Aduri, Brian T. Psciuk, Pirro Saro, Hariprakash Taniga, H. Bernhard Schlegel, and John SantaLucia. AMBER force field parameters for the naturally occurring modified nucleosides in RNA. *Journal of Chemical Theory and Computation*, 3(4):1464–1475, 2007.

[2] 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 1, 1995.

[3] Peter Clote and Amir H. Bayegan. RNA folding kinetics using Monte Carlo and Gillespie algorithms? Technical report, 2017.

[4] Tingting Sun, Chenhan Zhao, and Shi Jie Chen. Predicting Cotranscriptional Folding Kinetics for Riboswitch. *Journal of Physical Chemistry B*, 2018.

[5] Springer Series in Biophysics 13. Technical report.