Weekly Report

Zhuoran Qiao

August 14, 2018

1 Introduction

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

2 Made initial tests on effect of folding rate on $p_{unbound}$ during transcription.

2 Progress

2.1 Framework

Over decades various algorithms or programs have been developed to predict RNA folding pathways; however present methods 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 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

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

2.2 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 2.3). 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 procedure are as follows (Algorithm 1):

2.3 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

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 / Transcription rate
  for left boundary ∈ {0, dL, 2dL, ..., Current length - dL} **do** ▷ Get all new foldons
    $D^{foldon}_{\text{left boundary, Current length}} \leftarrow$ 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** $| \exp\left(\hat{\mathbf{T}}\right)$ ▷ Master equation
  reserve top N populated strands in ActivePool ▷ Selection
  renormalize $\langle$ActivePool.**population** $|$
end while

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.