Simulate co-transcriptional folding kinetics by genetic algorithm

Zhuoran Qiao

August 8, 2018

We developed a genetic algorithm based approach to simulate kinetics of co-transcriptional folding. Our method is built on two following assumptions:

1 Most of RNA secondary structures (SS) are linkage of locally optimal or sub-optimal structures (foldons) 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 direct product... irreducible foldon representation... green's function...

1 Algorithm procedure

During every elongation step, an active species pool of strands with unique SS and diffrent population is updated. New candidate strands with length of $L + \Delta L$ are generated by a recombination process: for every old strand, all indices in its IFR is identified as possible rearrangement site, then its child strands is generated by linking partial segments with a foldon 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 $[0, L]$ 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 3). Then the population distribution of strands after elongation is computed by chemical master equation, and one iterative elongation step is finished.

Pseudocodes of the whole procedure are as follows:

Algorithm 1 Co-transcriptional folding elongation procedure

Initalize active pool
while sequence length > current length **do**
    old pool ← active pool
    renew active pool
    current length+ = dL
    dt ← dL / transcription rate
    for left boundary ∈ {0, dL, 2dL, ..., current length - dL} **do** ▷ Get all new foldons
        D^{foldon}(left boundary, current length) ← numpy.mfe(sequence[left boundary, current length])
    end for
    for strand ∈ old pool **do** ▷ Recombination
        for rearrangement site ∈ strand.IFR **do**
            Candidate ←
$D^{strand}(0,\ \text{rearrangement site}) \oplus D^{foldon}(\text{rearrangement site},\ \text{current length})$
        end for
    end for
end while

Algorithm 2 Euclid's algorithm

procedure EUCLID(a, b) ▷ The g.c.d. of a and b
    $r \leftarrow a \bmod b$
    while $r \neq 0$ **do** ▷ We have the answer if r is 0
        $a \leftarrow b$
        $b \leftarrow r$
        $r \leftarrow a \bmod b$
    end while
    for `<some condition>` **do**
        `<do stuff>`
    end for
    return b ▷ The gcd is b
end procedure

2 Secondary structure generation

Our goal is to figure out a general approach to calculate transition path statistics (average transit time and transit time distribution) for free energy landscape with discrete states. The Markov rate matrix $\mathbf{W}$ element between state σ_i and state σ_j is given by

$$\mathbf{W}_{ij} = \langle \sigma_i | \mathbf{W} | \sigma_j \rangle = k_0 \exp\left(-\frac{F_j - F_i}{2} \right) \tag{1}$$

$$\begin{aligned} P(t) &= \sum_{\varphi} P(t|\varphi) P(\varphi) \\ &= \sum_{\varphi} P(t|\vec{H}[\varphi]) P(\varphi) \\ &= \sum_{\vec{H}} P(t|\vec{H}) P(\vec{H}) \\ &= \sum_{\vec{H},l} P(t|\vec{H}) P(\vec{H}|l) P(l) \end{aligned} \tag{2}$$

3 Folding pathway identification & Rate calculation

a. Using a genetic-algorithm based method and NUPACK to predict populated RNA configurations during cotranscriptional folding (in progress);

b. Using master equation method to simulate evolution of folding configurations and SD sequence accessibility.