An RNA kinetics ansatz derived from an efficient prediction of RNA pathways

Computational Approaches to RNA structure and function, Benasque 2022

Nono S.C. Merleau, V. Opuu, V. Messow & M. Smerlak

August 15, 2022

Max Planck Institute for Mathematics in the sciences
Structure of Evolution Group.

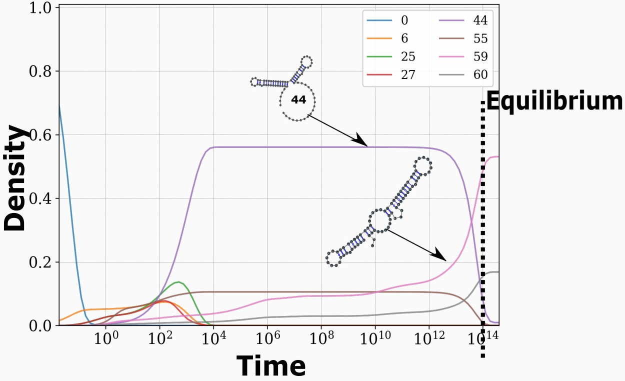
Presentation plan

1. RNA folding: dynamic aspects
2. Traditional methods
3. RAFFT as an alternative solution.
4. The derived kinetics ansatz
5. Test cases
6. Conclusion

RNA folding: dynamic aspects

RNA folding: dynamic aspects

1. The RNA folding is remarkably complex: non-canonical interactions, constant formation or dissolving of base pairs.
2. Thermodynamics gives the distribution at equilibrium, which may not be relevant in biological time scales.
3. Model the folding dynamics with a CTMC [*Ronny L., Flamm C, Hofacker I. and Stadler P., (2008) EPJ B*].



Traditional methods

Traditional methods

Basin-based method

- Sample the equilibrium distribution (`RNAsubopt`).
- Coarse grain the ensemble of structures into a small number of connected basins (`barrier`).
- Estimate transition rates between basins with using transition states (`barrier`).
- Arrhenius formulation: $k_{i\to j} = k_0 \exp(-\beta \Delta G^{\ddagger}_{i\to j})$

Transition rate models

- Base stack transition [*Wenbing et al.*]
- Base pair transition [*Simona et al.*]
- Helix stem transition [*Hervé et al.*]

Limitations

- Enumerate the whole structural space
- Rate model *vs.* CPU time

Master equation

$$\frac{\mathrm{d}p_i(t)}{\mathrm{d}t} = \sum_{j\in\Omega} k_{j\to i} p_j(t) - k_{i\to j} p_i(t) \tag{1}$$

RAFFT as a sampling tool for meta stable structures

Prediction method for RNA pathways

- One-hot encoding

$$A \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, U \rightarrow \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, C \rightarrow \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, G \rightarrow \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \tag{2}$$

This encoding gives us a $(4 \times L)$-matrix we call X, where each row corresponds to a nucleotide as shown below:

$$X = \begin{pmatrix} X^A \\ X^C \\ X^G \\ X^U \end{pmatrix} = \begin{pmatrix} X^A(1) & X^A(2) & \dots & X^A(L) \\ X^C(1) & X^C(2) & \dots & X^C(L) \\ X^G(1) & X^G(2) & \dots & X^G(L) \\ X^U(1) & X^U(2) & \dots & X^U(L) \end{pmatrix} \tag{3}$$

Prediction method for RNA pathways

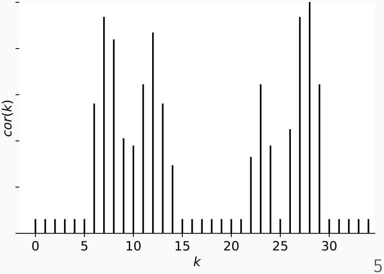
- One-hot encoding
  Next, we create a second copy $\bar{S} = (\bar{S_L} \ldots \bar{S_1})$ for which we reversed the sequence order. Then, each nucleotide of $\bar{S}$ is replaced by one of the following unit vectors:

$$\bar{A} \rightarrow \begin{pmatrix} 0 \\ 0 \\ 0 \\ w_{\text{AU}} \end{pmatrix}, \bar{U} \rightarrow \begin{pmatrix} w_{\text{AU}} \\ w_{\text{GU}} \\ 0 \\ 0 \end{pmatrix}, \bar{C} \rightarrow \begin{pmatrix} 0 \\ 0 \\ w_{\text{GC}} \\ 0 \end{pmatrix}, \bar{G} \rightarrow \begin{pmatrix} 0 \\ w_{\text{GC}} \\ 0 \\ w_{\text{GU}} \end{pmatrix} \qquad (4)$$

- Compute the correlation $cor(X, \bar{X})$

$$\text{cor}(k) = \sum\nolimits_{\alpha \in \{A,U,C,G\}} c_{X^\alpha, \bar{X}^\alpha}(k)$$

$$c_{X^\alpha, \bar{X}^\alpha}(k) = \sum_{\substack{1 \leq i \leq L \\ 1 \leq i+k \leq L}} \frac{X^\alpha(i)\bar{X}^\alpha(i+k)}{\min(k, 2L-k)}$$



- Using the FFT makes it more efficient

Prediction method for RNA pathways: recursion

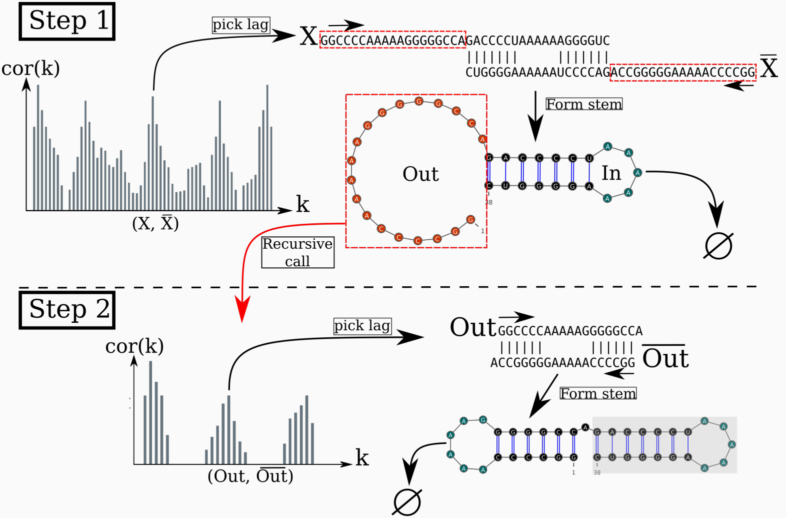


Figure 1: RAFFT heuristic

Prediction method for RNA pathways: output and kinetics ansatz

- Exploiting the folding graph produced
- Stem transition without barrier energies
- Metropolis formulation:

$$k_{i\rightarrow j} = \begin{cases} k_0 \times \min(1, \exp(-\beta\Delta(\Delta G_{i\rightarrow j}))), & \text{if } \sigma_i \in \mathcal{M}(\sigma_j) \\ 0, & \text{else} \end{cases}$$

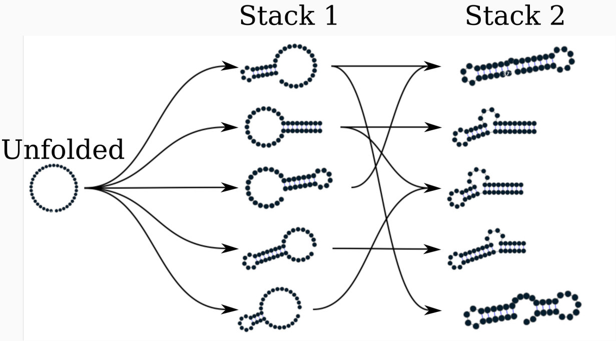
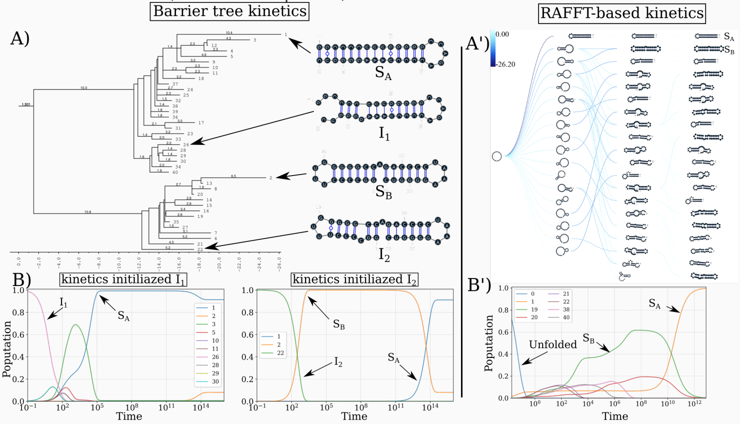


Figure 2: RAFFT fast folding graph

Test cases (1): toy bi-stable sequence

Traditional kinetics using `Treekin` *vs.* `RAFFT`

1. Generate 20*k* suboptimal structures (*vs.* 20/46).
2. Coarse-grained into 40 basins (`barrier`).
3. Compute transition rates between basins using the Arrhenius formulation (*vs.* Metropolis).



Barrier tree kinetics

RAFFT-based kinetics

A) A') B) B')

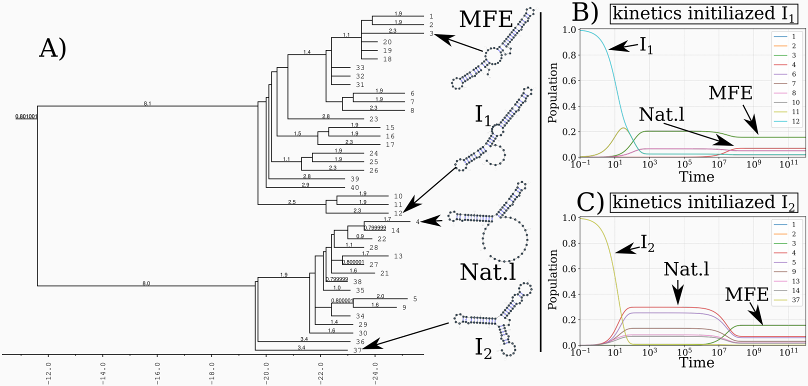
kinetics initiliazed I_1

kinetics initiliazed I_2

Test cases (2): coronavirus frameshifting stimulation element (CFSE)

Traditional kinetics using `Treekin`

1. 1.5 millions structures sampled using `RNAsubopt`
2. Coarse-grained into 40 basins (`barrier`).



Test cases (2): coronavirus frameshifting stimulation element (CFSE)

kinetics ansatz using **RAFFT fast folding graph**

1. 20 concurrent trajectories
2. 60 distinct secondary structures

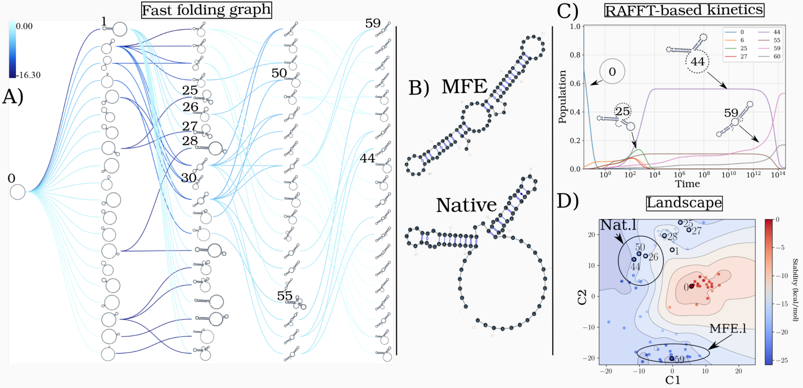


Figure 3: Kinetics using **RAFFT**

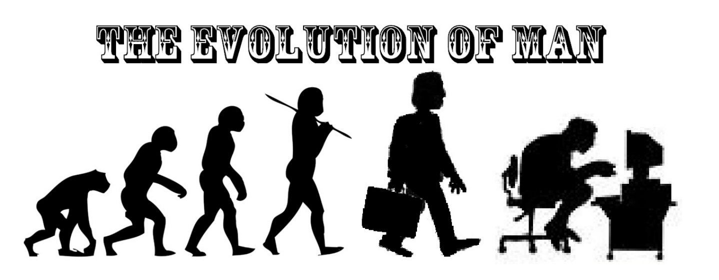
Conclusion and perspectives

Take home

1. We suggest a simple heuristic to predict RNA pathways using an efficient stem sampling method.
2. Qualitatively reproduce the dynamics of simple test cases but using fewer structures from the produced FFG.
3. The FFG reveals important metastable structures.
4. The folding graph spans over the free energy landscape to the closest minima.
5. We also use a coarse-grained model where only helices can be formed and unformed.

What next?

- RNA-RNA interaction
- RNA design that accounts for RNA-Y interactions
- Continuity/plasticity in evolution



Thanks for your attention