Understanding secondary structure effects on DNA hybridization by visualization tools in large CTMCs

Chenwei Zhang

1 Introduction and related work

Nucleic acids, including deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) play important roles in the continuity of life. DNA exists in almost every organism and carries genetic information and instructions on protein synthesis. RNA molecules involving three major types such as transfer RNA (tRNA), messenger RNA (mRNA), and ribosomal RNA (rRNA) are mainly functionalized to convert information stored in DNA into proteins. In the past few decades, DNA and RNA nanotechnologies have emerged and been thriving as their abilities to customize molecular self-assembly. Thermodynamics of nucleic acids has been extensively studied. On the other hand, it still leaves many blind spots to understand kinetics due to its sophisticated intrinsic characteristics. Therefore, it is imperative to seek out efficient ways to comprehend and address the mechanism of nucleic acid reaction kinetics.

Comprehensively understanding DNA hybridization process is important for DNA- and RNA-based biosensor technology. However, few works have been done to analyze the influence of hairpins formation on reaction rates and pathways in DNA hybridization. An illustration in Figure 1 shows an example of this reaction process. Although there is an existing work using *oxDNA* which is a coarse-grained model simulating nucleic acid nanotechnology [1] to simulate hairpin structures in the duplex formation and study their effects on forward and backward reaction rates [2], it is still not very clear what happening and how hairpins contributing during entire transitions of the helix association and dissociation from initial to final states. Here, I propose a more straightforward way to reveal this process by implementing visualization tools to help domain experts to appreciate mechanisms behind the kinetics and help them to design deliberate molecular structures upon DNA nanotechnology. *RiboSketch*, a drawing program to image DNA and RNA secondary structures, can be regarded as a good reference in this project [3].

Continuous-time Markov Chains (CTMCs) are commonly utilized to simulate nucleic acid kinetics, such as *Kinfold* [4], *Kfold* [5], and *Multistrand* [6] owing to their capabilities to infer equilibrium and non-equilibrium dynamics. A CTMC model of a reaction consists of a state space that includes all possible secondary structures, transitions each of which refers to an elementary step (a single base pair forming or breaking), and transition probabilities that are determined by latent kinetic parameters. A CTMC with reasonable space size can offer direct computation of its dynamics with matrix equation. Whereas one big challenge in real-data CTMCs is that the CTMCs of interest normally have a prohibitively large state space size therefore, it is infeasible to enumerate all the states to compute measures over paths. As a result, appropriate sampling approaches have to be applied to simulate statistically correct trajectories, i.e. sequences of observed states (secondary structures, indicating the hydrogen bonding state of the bases) along with the holding time for each transition [6].

The naivest sampling method is known as Gillespie sampling algorithm [7] for which trajectories are generated based on the transition rate and holding time of each state. Due to the inefficiency of Gillespie sampling, however, it requires to simulate an extremely large number of samples to acquire reliable estimates. Alternatively, it necessitates other advanced sampling approaches, such as weighted ensemble sampling (WES) [8], forward flux sampling (FFS) [9, 10], and kinetic path sampling (KPS) [11, 12]. These methods show good performance on CTMCs having relatively small state space sizes and high state occupancies. Instead, for the CTMCs of interest in this study, the state space size is quite large while the occupancy is especially

low. It may need to adapt these sampling strategies to realize the purpose.

In this project, I aim to implement a meaningful graphing tool to visualize DNA hybridization kinetics, particularly helix association and dissociation processes and further to understand the effects on hairpins forming and breaking on the course of DNA double-helical strands hybridization and melting. To achieve this goal, I propose a 3D visualization tool with a time slider to display the secondary structure transformation from initial to final states, accompanying by hairpin structure forming and breaking. I am planning to implement Python and Julia to deal with datasets and make plots. And I am also resorting to the software package *DISCOTRESS* [13, 14] to accomplish the various sampling methods, such as Gillespie, WES, FFS, and KPS.

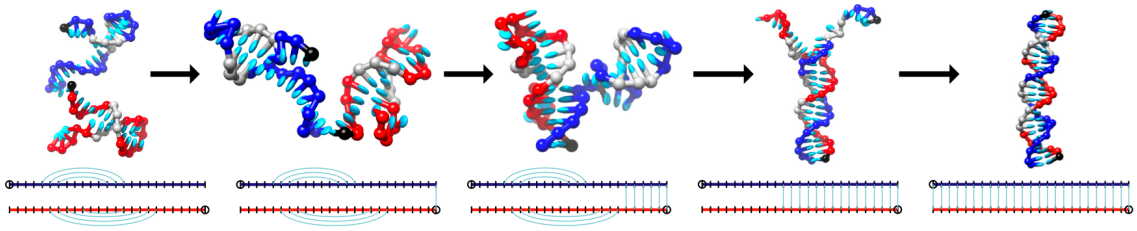

Figure 1: A typical helix association pathway. Schematic illustrations below the figure indicate hairpin and duplex base pairs present [2].

2 Background on CTMCs

A CTMC is defined as a tuple $C = (S, \boldsymbol{K}, \pi_0, F)$, where S refers to the set of states, $\boldsymbol{K} : S \times S \to \mathbb{R}_{\geq 0}$ refers to the rate matrix subject to $\boldsymbol{K}(s, s) = 0$ for any $s \in S$, $\pi_0 : S \to [0, 1]$ refers to the initial state distribution subject to $\sum_{s \in S} \pi_0(s) = 1$, and $F \subset S$ refers to the set of target (absorbing) states. We define $I \subset S$ as the set of initial states, that is $I = \{s \in S \,|\, \pi_0(s) > 0\}$. A transition from states s to s' is possible only if $\boldsymbol{K}(s, s') > 0$. The exit rate matrix, $\boldsymbol{E} : S \times S \to \mathbb{R}_{\geq 0}$, is a diagonal matrix defined as $\boldsymbol{E} = \sum_{s' \in S} \boldsymbol{K}(s, s')$ in which each entry represents an exit rate. And the holding time of a state s is exponentially distributed with respect to its corresponding exit rate $\boldsymbol{E}(s, s)$. We also define the generating matrix, $\boldsymbol{Q} : S \times S \to \mathbb{R}$, as $\boldsymbol{Q} = \boldsymbol{K} - \boldsymbol{E}$.

A trajectory is denoted by (s_0, t_0), (s_1, t_1), ..., (s_n, t_n) with n transitions over a CTMC model, where states s_i and holding time t_i subject to $\boldsymbol{K}(s_i, s_{i+1}) > 0$ and $t_i \in \mathbb{R}_{\geq 0}$ for $i \geq 0$. A path is defined as s_0, s_1, ..., s_n with n transitions over a CTMC model, where states s_i subject to $\boldsymbol{K}(s_i, s_{i+1}) > 0$.

3 Datasets

1. Modified CTMCs from toy-DNA dataset [15].
2. Generating CTMCs from the existing dataset [2].
3. Synthetic CTMCs.
4. If applicable, CTMCs from *Multistrand*.

4 Contributions

In this project, I am planning to :

1. Augment CTMCs from existing experimental data to allow the hairpin forming in the process of helix association and dissociation. And generate CTMCs from the existing dataset [2].

2. Implement different sampling methods, including Gillespie sampling, WES, FFS, and KPS to obtain statistically correct states and trajectories in large CTMCs.
3. Find a meaningful way to index each state and design efficient visualization tools to understand the effect of hairpin on DNA hybridization.
4. If possible, get insights from Deep Graph Embeddings paper and then find common ground to combine my project with this approach [16].
5. If time permits, implement the pathway elaboration algorithm [17] and combine it with visualization tools to analyze the relationship between truncated states and hairpin secondary structures in helix association and dissociation.

References

[1] Erik Poppleton, Roger Romero, Aatmik Mallya, Lorenzo Rovigatti, and Petr Šulc. Oxdna. org: a public webserver for coarse-grained simulations of dna and rna nanostructures. *Nucleic acids research*, 49(W1):W491–W498, 2021.

[2] John S Schreck, Thomas E Ouldridge, Flavio Romano, Petr Šulc, Liam P Shaw, Ard A Louis, and Jonathan PK Doye. Dna hairpins destabilize duplexes primarily by promoting melting rather than by inhibiting hybridization. *Nucleic acids research*, 43(13):6181–6190, 2015.

[3] Jacob S Lu, Eckart Bindewald, Wojciech K Kasprzak, and Bruce A Shapiro. Ribosketch: versatile visualization of multi-stranded rna and dna secondary structure. *Bioinformatics*, 34(24):4297–4299, 2018.

[4] Christoph Flamm, Walter Fontana, Ivo L Hofacker, and Peter Schuster. Rna folding at elementary step resolution. *Rna*, 6(3):325–338, 2000.

[5] Eric C Dykeman. An implementation of the gillespie algorithm for rna kinetics with logarithmic time update. *Nucleic acids research*, 43(12):5708–5715, 2015.

[6] J.M. Schaeffer. The Multistrand simulator: Stochastic simulation of the kinetics of multiple interacting DNA strands. Master's thesis, California Institute of Technology, 2012.

[7] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[8] G. Huber and S. Kim. Weighted-ensemble brownian dynamics simulations for protein association reactions. *J. Biophys.*, 70, 1996.

[9] R. Allen, P. Warren, and P. ten Wolde. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.*, 94, 2005.

[10] R. Allen, D. Frenkel, and P. ten Wolde. Simulating rare events in equilibrium or nonequilibrium stochastic systems. *J. Chem. Phys.*, 124, 2006.

[11] M. Athenes and V. Bulatov. Weighted-ensemble brownian dynamics simulations for protein association reactions. *Phys. Rev. Lett.*, 113, 2014.

[12] M. Athenes, S. Kaur, G. Adjanor, T. Vanacker, and T. Jourdan. Weighted-ensemble brownian dynamics simulations for protein association reactions. *Phys. Rev. Mater.*, 3, 2019.

[13] D. Sharpe and D. Wales. Efficient and exact sampling of transition path ensembles on markovian networks. *J. Chem. Phys*, 024121(153), 2020.

[14] D. Sharpe and D. Wales. Nearly reducible finite markov chains: theory and algorithms. *J. Chem. Phys*, in press.

[15] S. Zolaktaf, F. Dannenberg, X. Rudelis, A. Condon, J.M. Schaeffer, M. Schmidt, C. Thachuk, and E. Winfree. Inferring parameters for an elementary step model of DNA structure kinetics with locally context-dependent Arrhenius rates. In *DNA Computing and Molecular Programming*, pages 172–187. Springer International Publishing, 2017.

[16] Egbert Castro, Andrew Benz, Alexander Tong, Guy Wolf, and Smita Krishnaswamy. Uncovering the folding landscape of rna secondary structure using deep graph embeddings. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4519–4528. IEEE, 2020.

[17] S. Zolaktaf, F. Dannenberg, M. Schmidt, A. Condon, and E. Winfree. The pathway elaboration method for mean first passage time estimation in large continuous-time Markov chains with applications to nucleic acid kinetics, 2021. Unpublished paper.