Visualization tools for understanding secondary structure effects on DNA reaction kinetics

Chenwei Zhang

Introduction and related work

Nucleic acids, namely 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 been developed that are capable of sensing and responding to changes in their environments, self-assembling into complex structures, and simulating computational models such as logic circuits or artificial neural networks. Thermodynamics of nucleic acids has been extensively studied, but the mechanisms that influence DNA reaction kinetics are less well understood. There are multiple simulation tools available for this purpose, including *Multistrand* [1, 2], *Kinfold* [3], *Kfold* [4], and *oxDNA* [5]. But there is also a need for better visualization techniques that can make the output of the simulation tools more comprehensible to domain experts. This project will focus on implementing powerful graphing tools to visualize nucleic acid reaction kinetics.

Multistrand [1, 2], *Kinfold* [3], and *Kfold* [4] use continuous-time Markov Chains (CTMCs) to model nucleic acid kinetics. A CTMC model of a reaction consists of a set of states corresponding to secondary structures, plus transitions, and transition rates between states. A secondary structure describes the set of base pairs formed via hydrogen bonding between Watson-Crick complementary bases, and each state has an associated free energy that is determined by latent thermodynamic parameters. Each elementary step corresponds to a single base pair forming or breaking, and the elementary step rates are determined by the state free energies as well as latent kinetic parameters. Roughly, transitions that lead to lower-energy states are more likely than those that increase the free energy. A CTMC with reasonable space size can offer direct computation of its dynamics with matrix equations. However, CTMCs that model typical DNA or RNA reactions of interest often 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, such as the Gillespie sampling algorithm [6] have to be applied to simulate statistically correct trajectories, i.e., sequences of observed states (secondary structures) along with the time for each transition [1].

One example of an important DNA reaction is toehold-mediated three-way strand displacement, see Figure 1. An invader strand initially binds to unpaired bases of a substrate strand, and then, through a process called branch migration, forms more base pairs with the substrate, displacing the incumbent strand in the process. The rate of the process depends in part on whether the invader is a perfect Watson-Crick complement to the region of the substrate to which it binds, or whether there is a mismatched base, and furthermore on the position of such a mismatch [7].

There are really two different challenges with respect to showcasing nucleic acid reactions based on their CTMCs or sampled trajectories. One is visualizing energy landscapes. The other is visualizing trajectories through those landscapes. One of the most basic visualizations just uses "dot-parenthesis" notation to represent a secondary structure and a sequence of such strings to represent a trajectory. This method does not situate the trajectory in the overall energy landscape. Visualizations of energy landscapes use mappings of the high-dimensional state space to 2D or 3D. One example, for toehold-mediated three-way

strand displacement, is shown in Figure 2: the two dimensions show the number of base pairs between the substrate and incumbent, and the substrate and invader [7]. However, this visualization is very specific to the underlying reaction. Flamm et al. use barrier trees to visualize landscapes. This approach is distinctive but does not provide a way to visualize trajectories through such landscapes [8, 9]. Castro et al. [10] use deep graph embedding techniques to uncover the energy landscape of DNA secondary structure. However, this approach does not address how to visualize trajectories through those landscapes and the output of deep graph embedding approaches may not be at all intuitive to researchers in DNA nanotechnology. *Kinefold* generates a movie showing reaction kinetics but is limited to displaying one trajectory, and does not show the overall energy landscape [11]. Accordingly, there is a need to have more visualization tools to visualize kinetic trajectories. Apart from sequence-level analysis, Berleant et al. [12] provide a domain-level based coarse-graining tool to display kinetic pathways, which is considered as an important reference for this project. The book *Visualization Analysis and Design* [13] related to the principles and implementations of visualization is also a good reference for this project and I will spend some time to get through it during the project.

There are many other interesting types of DNA reactions, in addition to toehold-mediated strand displacement shown in Figure 1. While my work will be guided by common examples, my goal is to develop visualization tools that will also work more generally, and will help domain experts appreciate the mechanisms behind reaction kinetics.

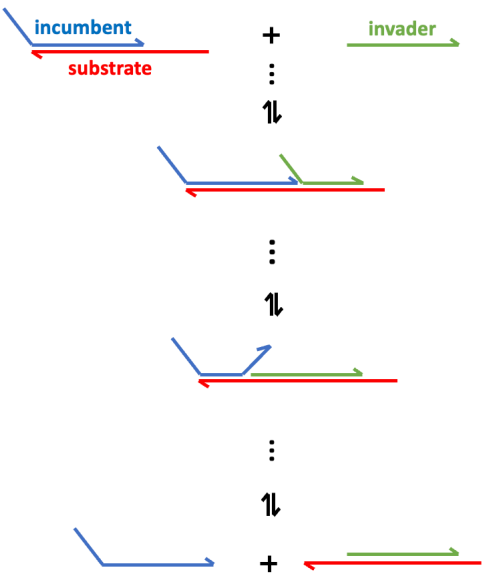

Figure 1: A toehold-mediated three-way strand displacement reaction and its mechanism. An invader strand (green line) replaces one of the strands in a duplex, i.e. incumbent strand (blue line), and forms a new double-helix strand with the substrate strand (red line). This figure shows a coarse-grained representation of the reaction, i.e. it groups many elementary steps into one coarse-grained step.

Goals

To summarize, in this project I am planning to :

1. Review the literature. Here I list some related literature that I will dig deeper into when I start the

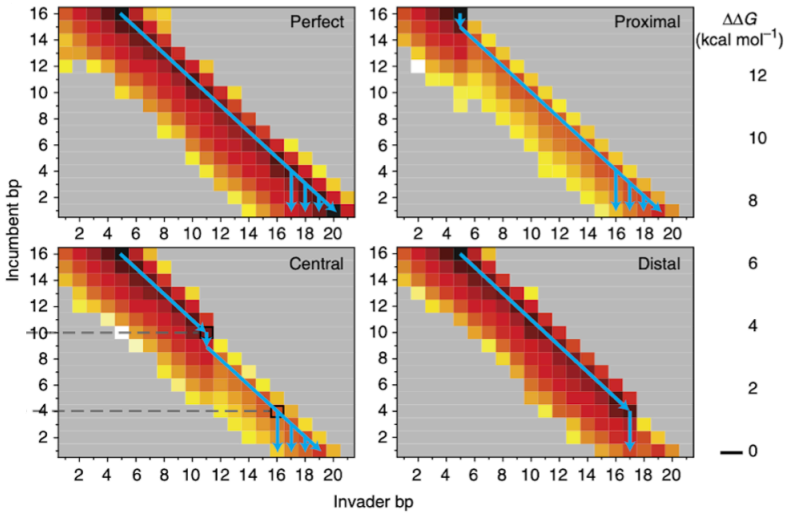

Figure 2: Results of simulation of toehold-mediated three-way strand displacement without (top left panel) and with different kinds of mismatches (the rest of three panels). For these plots, coordinates represent the numbers of base pairs formed between the substrate and incumbent, and the substrate and invader. This figure is retrieved from [7].

project.

(a) 2D or 3D embedding papers [10, 14, 15, 16].

(b) Case study papers [12, 17, 18, 19, 20].

(c) Barrier trees papers [8, 9].

(d) The book *Visualization Analysis and Design* [13].

2. Develop visualization tools using Python and Julia.
Here I list three possible options for designing visualization tools. Eventually I will choose one of the most interest to pursue further study.

   (a) 2D or 3D embeddings on a coarse-grained landscape. I am interested in generalizing this approach to other types of reaction including, but not limited to, those in [21].

   (b) Base pairing probabilities over time. 2D plots showing equilibrium base pairing probabilities are familiar to domain experts. The idea is to show these over time (add a slider).

   (c) Possibly barrier trees. I may try to modify this visualization approach to show a pathway representation.

3. Evaluate the visualization tools using one or two case studies.

   (a) Visualize kinetic pathways on two examples from the papers: Schreck et al. [22] and Zhang et al. [17].

   (b) Use *Multistrand* to generate trajectories (or CTMCs, using pathway elaboration) if applicable, or augment the DNA23 collection of CTMCs to include helix association and dissociation with intermediate hairpin formation.

Challenges

There are several challenges in this work.

1. Find meaningful ways to index each state.
2. How to show multiple trajectories through the landscapes.
3. How to assess which visualization approach works better? One obvious way is domain experts' evaluation. Other assessing ways may resort to some insights from the visualization book [13].

Timeline

- 1-Jun to 30-Jun: Review past works.
- 1-Jun to 15-Jul: Identify strengths and weaknesses of different possible visualization methods, and implement one of the more promising options.
- 16-Jul to 31-Aug: Evaluate the implemented tool(s) using case studies.
- 1-Sep to 30-Sep: Improve methods and write final report.

In conclusion, in this project I aim to design and implement general visualization tools for nucleic acid kinetics to help domain experts to make the output from the simulation approaches more comprehensible.

References

[1] J.M. Schaeffer. *Stochastic simulation of the kinetics of multiple interacting nucleic acid strands*. PhD thesis, California Institute of Technology, 2013.

[2] Joseph Malcolm Schaeffer, Chris Thachuk, and Erik Winfree. Stochastic simulation of the kinetics of multiple interacting nucleic acid strands. In *International Workshop on DNA-Based Computers*, pages 194–211. Springer, 2015.

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

[4] Eric C Dykeman. An implementation of the Gillespie algorithm for RNA kinetics with logarithmic time update. *Nucleic Acids Research*, 43(12):5708–5715, 2015.

[5] 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.

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

[7] R.R. Machinek, T.E. Ouldridge, N.E. Haley, J. Bath, and A.J. Turberfield. Programmable energy landscapes for kinetic control of DNA strand displacement. *Nature Communications*, 5, 2014.

[8] Christoph Flamm, Ivo L Hofacker, Peter F Stadler, and Michael T Wolfinger. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, 2002.

[9] Mag. rer. nat. Stefan Badelt. *Control of RNA function by conformational design*. PhD thesis, Universitt Wien, 2016.

[10] 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.

[11] Alain Xayaphoummine, T Bucher, and Herve Isambert. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, 33(suppl_2):W605–W610, 2005.

[12] Joseph Berleant, Christopher Berlind, Stefan Badelt, Frits Dannenberg, Joseph Schaeffer, and Erik Winfree. Automated sequence-level analysis of kinetics and thermodynamics for domain-level DNA strand-displacement systems. *Journal of the Royal Society Interface*, 15(149):20180107, 2018.

[13] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

[14] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019.

[15] Benjamin WB Shires and Chris J Pickard. Visualizing energy landscapes through manifold learning. *Physical Review X*, 11(4):041026, 2021.

[16] Rony Lorenz, Christoph Flamm, and Ivo L Hofacker. 2D projections of RNA folding landscapes. In *German Conference on Bioinformatics 2009*. Gesellschaft für Informatik eV, 2009.

[17] David Yu Zhang, Andrew J Turberfield, Bernard Yurke, and Erik Winfree. Engineering entropy-driven reactions and networks catalyzed by DNA. *Science*, 318(5853):1121–1125, 2007.

[18] Kyle E Watters, Eric J Strobel, M Yu Angela, John T Lis, and Julius B Lucks. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nature Structural & Molecular Biology*, 23(12):1124–1131, 2016.

[19] M Yu Angela, Paul M Gasper, Luyi Cheng, Lien B Lai, Simi Kaur, Venkat Gopalan, Alan A Chen, and Julius B Lucks. Computationally reconstructing cotranscriptional RNA folding from experimental data reveals rearrangement of non-native folding intermediates. *Molecular Cell*, 81(4):870–883, 2021.

[20] Matthew R Lakin, Simon Youssef, Filippo Polo, Stephen Emmott, and Andrew Phillips. Visual DSD: a design and analysis tool for DNA strand displacement systems. *Bioinformatics*, 27(22):3211–3213, 2011.

[21] 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.

[22] 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.