The Multistrand Simulator: Stochastic Simulation of the Kinetics of Multiple Interacting DNA Strands

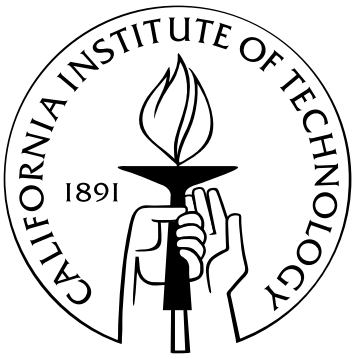
Thesis by

Joseph Malcolm Schaeffer

In Partial Fulfillment of the Requirements

for the Degree of

Master of Science



California Institute of Technology

Pasadena, California

2012

(Submitted February 4, 2012)

Acknowledgements

Thanks to my advisor Erik Winfree, for his enthusiasm, expertise and encouragement. The models presented here are due in a large part to helpful discussions with Niles Pierce, Robert Dirks, Justin Bois and Victor Beck. Two undergraduates, Chris Berlind and Jashua Loving, did summer research projects based on Multistrand and in the process helped build and shape the simulator. There are many people who have used Multistrand and provided very helpful feedback for improving the simulator, especially Josh Bishop, Nadine Dabby, Jonathan Othmer, and Niranjan Srinivas. Thanks also to the many past and current members of the DNA and Natural Algorithms group for providing a stimulating environment in which to work.

There are many medical professionals to which I owe my good health while writing this thesis, especially Dr. Jeanette Butler, Dr. Mariel Tourani, Cathy Evaristo, and the staff of the Caltech Health Center, especially Alice, Divina and Jeannie.

I want to acknowledge all my family and friends for their support. A journey is made all the richer for having good company, and I would not have made it nearly as far without all the encouragement.

Finally, I must thank my wife Lorian, who has been with me every step of this journey and has shared all the high points and low points with her endless love and support.

Abstract

DNA nanotechnology is an emerging field which utilizes the unique structural properties of nucleic acids in order to build nanoscale devices, such as logic gates, motors, walkers, and algorithmic structures. These devices are built out of DNA strands whose sequences have been carefully designed in order to control their secondary structure - the hydrogen bonding state of the bases within the strand (called "base-pairing"). This base-pairing is used to not only control the physical structure of the device, but also to enable specific interactions between different components of the system, such as allowing a DNA walker to take steps along a prefabricated track. Predicting the structure and interactions of a DNA device requires good modeling of both the thermodynamics and the kinetics of the DNA strands within the system. Thermodynamic models can be used to make equilibrium predictions for these systems, allowing us to look at questions like "Is the walker-track interaction a well-formed and stable molecular structure?", while kinetics models allow us to predict the non-equilibrium dynamics, such as "How quickly will the walker take a step?". While the thermodynamics of multiple interacting DNA strands is a well-studied model which allows for both analysis and design of DNA devices, previous work on the kinetics models only explored the kinetics of how a single strand folds on itself.

The kinetics of a set of DNA strands can be modeled as a continuous time Markov process through the state space of all secondary structures. Due to the exponential size of this state space it is computationally intractable to obtain an analytic solution for most problem sizes of interest. Thus the primary means of exploring the kinetics of a DNA system is by simulating trajectories through the state space and aggregating data over many such trajectories. We developed the the **Multistrand** kinetics simulator, which extends the previous work by including the multiple strand version of the thermodynamics model (a core component for calculating parameters of the kinetics model), extending the thermodynamics model to include terms accounting for the fixed volume simulations, and

by adding new kinetic moves that allow interactions between distinct strands. Furthermore, we prove that our modified thermodynamic and kinetic models are exactly equivalent to the canonical thermodynamics model when the simulation is run for sufficiently long time to reach equilibrium.

The kinetic simulator was implemented in C++ for time critical components and in Python for input and output routines as well as post-processing of trajectory data. A key contribution of this work was the development of data structures and algorithms that take advantage of local properties of secondary structures. These algorithms enable the efficient reuse of the basic objects that form the system, such that only a very small part of the state's neighborhood information needs to be recalculated with every step. Another key addition was the implementation of algorithms to handle the new kinetic steps that occur between different DNA strands, without increasing the time complexity of the overall simulation. These improvements led to a reduction in worst case time complexity of a single step being just quadratic in the input size (the number of bases in the simulated system), rather than cubic, and also led to additional improvements in the average case time complexity.

What data does the simulation produce? At the very simplest, the simulation produces a full kinetic trajectory through the state space - the exact states it passed through, and the time at which it reached them. A small system might produce trajectories that pass through hundreds of thousands of states, and that number increases rapidly as the system gets larger! Going back to our original question, the type of information a researcher hopes to get out of the data could be very simple: "How quickly does the walker take a step?", with the implied question of whether it's worth it to actually order the particular DNA strands composing the walker, or go back to the drawing board and redesign the device. One way to acquire that type of information is to look at the first time in the trajectory where we reached the "walker took a step" state, and record that information for a large number of simulated trajectories in order to obtain a useful answer. We designed and implemented new simulation modes that allow the full trajectory data to be condensed as it's generated into only the pieces the user cares about for their particular question. This analysis tool also required the development of flexible ways to talk about states that occur in trajectory data; if someone wants data on when the walker took a step, we have to be able to express that in terms of the Markov process states which meet that condition.

Contents

List of Figures

Chapter 1

Introduction

DNA nanotechnology is an emerging field which utilizes the unique structural properties of nucleic acids in order to build nanoscale devices, such as logic gates [18], motors [4, 1], walkers [19, 1, 20], and algorithmic structures [14, 23]. These devices are built out of DNA strands whose sequences have been carefully designed in order to control their secondary structure - the hydrogen bonding state of the bases within the strand (called "base-pairing"). This base-pairing is used to not only control the physical structure of the device, but also to enable specific interactions between different components of the system, such as allowing a DNA walker to take steps along a prefabricated track. Predicting the structure and interactions of a DNA device requires good modeling of both the thermodynamics and the kinetics of the DNA strands within the system. Thermodynamic models can be used to make equilibrium predictions for these systems, allowing us to look at questions like "Is the walker-track interaction a well-formed and stable molecular structure?", while kinetics models allow us to predict the non-equilibrium dynamics, such as "How quickly will the walker take a step?". While the thermodynamics of multiple interacting DNA strands is a well-studied model [5] which allows for both analysis and design of DNA devices [24, 6], previous work on secondary structure kinetics models only explored the kinetics of how a single strand folds on itself [7].

The kinetics of a set of DNA strands can be modeled as a continuous time Markov process through the state space of all secondary structures. Due to the exponential size of this state space it is computationally intractable to obtain an analytic solution for most problem sizes of interest. Thus the primary means of exploring the kinetics of a DNA system is by simulating trajectories through the state space and aggregating data over many such trajectories. We present here the **Multistrand** kinetics simulator, which extends the previous

work [7] by using the multiple strand thermodynamics model [5] (a core component for calculating transition rates in the kinetics model), adding new terms to the thermodynamics model to account for stochastic modeling considerations, and by adding new kinetic moves that allow bimolecular interactions between strands. Furthermore, we prove that this new kinetics and thermodynamics model is consistent with the prior work on multiple strand thermodynamics models [5].

The **Multistrand** simulator is based on the Gillespie algorithm [8] for generating statistically correct trajectories of a stochastic Markov process. We developed data structures and algorithms that take advantage of local properties of secondary structures. These algorithms enable the efficient reuse of the basic objects that form the system, such that only a very small part of the state's neighborhood information needs to be recalculated with every step. A key addition was the implementation of algorithms to handle the new kinetic steps that occur between different DNA strands, without increasing the time complexity of the overall simulation. These improvements lead to a reduction in worst case time complexity of a single step and also led to additional improvements in the average case time complexity.

What data does the simulation produce? At the very simplest, the simulation produces a full kinetic trajectory through the state space - the exact states it passed through, and the time at which it reached them. A small system might produce trajectories that pass through hundreds of thousands of states, and that number increases rapidly as the system gets larger. Going back to our original question, the type of information a researcher hopes to get out of the data could be very simple: "How quickly does the walker take a step?", with the implied question of whether it's worth it to actually purchase the particular DNA strands composing the walker to perform an experiment, or go back to the drawing board and redesign the device. One way to acquire that type of information is to look at the first time in the trajectory where we reached the "walker took a step" state, and record that information for a large number of simulated trajectories in order to obtain a useful answer. We designed and implemented new simulation modes that allow the full trajectory data to be condensed as it's generated into only the pieces the user cares about for their particular question. This analysis tool also required the development of flexible ways to talk about states that occur in trajectory data; if someone wants data on when the walker took a step, we have to be able to express that in terms of the Markov process states which meet that condition.