Investigating DNA Junction Structure and Dynamics using a Coarse-grained Implicit Ion Model for DNA

by

Abraham Kipnis

Class of 2019

An honors thesis submitted to the faculty of Wesleyan University in partial fulfillment of the requirements for the Degree of Bachelor of Arts with Departmental Honors in Physics

Middletown, Connecticut April, 2019
Abstract

Four-way Holliday junctions are cruciform-shaped DNA structures which play vital roles in biological processes. In this thesis, we validate the ability of an explicit ion coarse-grained model for DNA (3SPN.2) to accurately simulate Holliday junction dynamics above and below DNA junction melting temperatures. We analyze a variety of junction behaviors, including ion binding, junction conformations, junction melting, and branch migration, and compare our results with expected results from scientific literature. We discuss four different methods to determine the structure of the junction, evaluate the drawbacks and advantages of these different methods by comparing them with each other and with data from previous studies and show that results qualitatively reflect our expectations of DNA junction structure at equilibrium. Then we use one of these methods to show that DNA junction dynamics as produced by the 3SPN.2 explicit ion model demonstrate the expected trends from literature. Next, we investigate melting by observing the dissociation of individual bases during our simulations and provide quantitative predictions for the dynamics of junction melting. Our results show that melting initiates at specific locations on a per-strand basis. These simulations helped inspire fluorescence melting experiments which strategically place nucleotide base analogs at several locations along junction strands and validate the primary predictions of the modeling. We offer a comprehensive overview of DNA junction research made possible using the 3SPN.2 explicit ion model for DNA and conclude that this model is a useful tool for exploring Holliday junction structure and dynamics in the presence of ions.
Acknowledgements

I would like to express my utmost gratitude to:

- My supervisor, Francis Starr, for giving me these projects, providing guidance and inspiration, patiently helping me hone my skills and find solutions to problems, and reminding me to not lose the forest for the trees.

- My second supervisor, Ishita Mukerji, for having an eye for details and an intuition that only an experimentalist can have.

- Other members of the Mukerji lab, including Rachel Savage, Julie McDonald, Nick Taylor, and Dacheng Zhao, for their contributions to experimental four-way junction research.

- Members of the Wesleyan physics department: Tsampikos Kottos, Greg Voth, Renee Sher, Fred Ellis, Tom Morgan and Yunseong Nam, for teaching me more than I could have imagined during my time here.

- Other current and ex members of the Starr lab, including Wujie Wang, Xinyu Zhu, Natalie Strassheim, Abhishek Fakiraswami, Wenghang Zhang, Hamed Emamy, Amber Storey, Chloe Thorburn and Nathan Shankman for laying the initial groundwork this project and for sharing expertise, advice, and being my personal heroes.

- The residents of 59 Warwick: Will Barr, Cail Daley, Ryan Adler-Levine, Gabe Weinreb, and Tony Strack for being with me during the summer of 2017.

- My family, for providing love and care the whole way through.
# Contents

1 Introduction ......................................................... 1
   1.1 Homologous Recombination .................................. 2
   1.2 DNA Junction Studies ........................................... 6
   1.3 DNA Junction Melting ........................................... 12
   1.4 DNA Nanotechnology ............................................ 14
   1.5 Motivation and Objectives .................................... 16

2 Model and Simulations ............................................... 17
   2.1 Coarse-grained DNA Model ..................................... 18
   2.2 Molecular Simulations .......................................... 20
   2.3 Hidden Markov models and our implementation .............. 23

3 Equilibrium Structure and Dynamics ................................. 25
   3.1 Radial distribution function .................................... 25
   3.2 Methods to determine junction isoform ....................... 28
      3.2.1 Central base distances .................................. 28
      3.2.2 Consecutive phosphate angles ......................... 31
      3.2.3 RMSD between ideal isoform structures and current snapshot .................. 33
      3.2.4 Markov model state classification .................... 34
   3.3 Transitions between conformations ............................ 38
CONTENTS

4 Melting Dynamics ......................................................... 39
  4.1 Method to determine base dissociation ........................... 40
  4.2 Preferential melting ............................................... 42
  4.3 Isoform transitions during melting process .................... 44

5 Branch Migration Dynamics .............................................. 47
  5.1 Methods to determine junction location ........................ 48
  5.2 Junction migration probability .................................. 48

6 Conclusion .................................................................... 51
  6.1 Evaluation of ionic distributions around the junction .......... 51
  6.2 Evaluation of junction isoform determination methods ........ 52
  6.3 Evaluation of melting data .......................................... 53
  6.4 Future directions ..................................................... 54
List of Figures

1.1 Meiotic recombination .................................................. 2
1.2 Holliday model ............................................................. 3
1.3 Junction conformation and branch migration energy landscape .......... 4
1.4 Holliday junction formation and branch migration .......................... 5
1.5 Holliday junction resolution .......................................... 5
1.6 Products from cleavage of junction by restriction enzymes ................. 6
1.7 Electrophoretic gel mobilities of restriction enzyme products ............. 7
1.8 DNA junction conformer transitions .................................. 8
1.9 FRET experiment to determine conformer populations as a function of \([\text{Mg}^{2+}]\) .......... 9
1.10 FRET experiment to determine conformer populations as a function of other ions .... 10
1.11 FRET experiments of mobile junctions to determine junction migration rates .... 11
1.12 Simulations of B-DNA show increased flexibility at high salt concentrations ........ 12
1.13 Radial distribution function between \(\text{Mg}^{2+}\) ions and DNA junction phosphates ... 13
1.14 Junction melting curve from implicit ion model for DNA ............... 13
1.15 Two-dimensional lattices made from DNA junction subunits ............ 14
1.16 Nanotubes made from DNA junction subunits ............................. 15
2.1 Illustration of coarse-grained representation of B-DNA using 3SPN.2 model ... 18
2.2 Fraction of broken bases from melting simulations .......................... 22
3.1 Sodium radial distribution functions .................................. 26
3.2 Magnesium radial distribution functions .................................. 26
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Schematic of isoform determination using base distances</td>
<td>28</td>
</tr>
<tr>
<td>3.4</td>
<td>Implicit ion model central base distance histograms and isoform population distributions</td>
<td>29</td>
</tr>
<tr>
<td>3.5</td>
<td>Explicit ion central base distance histogram</td>
<td>30</td>
</tr>
<tr>
<td>3.6</td>
<td>Isoform population distributions from base distances criterion using explicit ion model</td>
<td>30</td>
</tr>
<tr>
<td>3.7</td>
<td>Angle between consecutive phosphates</td>
<td>31</td>
</tr>
<tr>
<td>3.8</td>
<td>Explicit ion central phosphate angle histograms</td>
<td>32</td>
</tr>
<tr>
<td>3.9</td>
<td>Isoform population distributions from central phosphate angles criterion</td>
<td>32</td>
</tr>
<tr>
<td>3.10</td>
<td>Isoform population distribution by using RMSD from ideal isoform</td>
<td>33</td>
</tr>
<tr>
<td>3.11</td>
<td>Isoform population distributions from base distances Markov models</td>
<td>35</td>
</tr>
<tr>
<td>3.12</td>
<td>Isoform population distributions from phosphate angles Markov models</td>
<td>36</td>
</tr>
<tr>
<td>3.13</td>
<td>Isoform population distributions from base distances and angles Markov models</td>
<td>37</td>
</tr>
<tr>
<td>3.14</td>
<td>Transition matrices from phosphate angles criterion</td>
<td>38</td>
</tr>
<tr>
<td>4.1</td>
<td>Stages of the melting process</td>
<td>39</td>
</tr>
<tr>
<td>4.2</td>
<td>Example of base pair distances and melting time</td>
<td>41</td>
</tr>
<tr>
<td>4.3</td>
<td>Number of base pairs and $t_m$ from one melting simulation</td>
<td>41</td>
</tr>
<tr>
<td>4.4</td>
<td>Mean base pair dissociation time $t_m$ at 360K</td>
<td>43</td>
</tr>
<tr>
<td>4.5</td>
<td>Experimental fluorescence melting curves</td>
<td>44</td>
</tr>
<tr>
<td>4.6</td>
<td>Theoretical melting temperatures for individual strands that comprise the J3 junction</td>
<td>45</td>
</tr>
<tr>
<td>4.7</td>
<td>Isoform timeseries for a single junction melting simulation</td>
<td>46</td>
</tr>
<tr>
<td>4.8</td>
<td>Isoform population distribution during melting</td>
<td>46</td>
</tr>
<tr>
<td>5.1</td>
<td>Junction location timeseries, [Na$^+$] 100mM [Mg$^{2+}$] 0mM</td>
<td>49</td>
</tr>
<tr>
<td>5.2</td>
<td>Junction location timeseries, [Na$^+$] 100mM [Mg$^{2+}$] 50mM</td>
<td>49</td>
</tr>
<tr>
<td>5.3</td>
<td>Junction migration probability</td>
<td>50</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

“Science, for me, gives a partial explanation for life. In so far as it goes, it is based on fact, experience and experiment.”

Rosalind Franklin, *a letter to Ellis Franklin*, 1940

DNA is an information-encoding polymer responsible for carrying instructions for how organisms live, procreate, and die. For solving the most common structure of DNA, the method by which DNA replicates, Watson, Crick, and Wilkins won the Nobel Prize for Physiology in 1962 with the help of x-ray diffraction data from Rosalind Franklin [1]. DNA is composed of two oppositely oriented strands of single-stranded DNA (ssDNA) in a right-handed helix, with bases on opposing strands connected through hydrogen bonding. Under physiological conditions, double-stranded DNA (dsDNA) exists in “B form” (B-DNA), with one turn of the helix approximately every 10 base pairs, or 34 Å. Most of the time, DNA is organized into condensed networks called chromosomes. In cells, helicases and polymerases aid DNA replication by unzipping the strands and using each as a template for synthesis of a new complementary strand. Imperfect copying and mismatch repair results in approximately one point mutation occurring per 10 billion bases replicated [2], which can alter protein structure and function. A diversity of phenotypes within a population contribute to a species’ ability to withstand environmental perturbations [2].
Another important evolutionary tool that increases genetic diversity is meiotic recombination. During meiosis (the repeated division of cells into four cells with half the number of chromosomes), duplicated homologous chromosomes align at the centers of cell nuclei before segregating and triggering cell division [See Figure 1.1]. Recombination begins between aligned homologous chromosomes at sites called chiasmata when recombination proteins create two “nicks” in similar regions of a single strand on each chromosome [3].

**Figure 1.1:** A biological pathway in which DNA junctions appear. In fungi, plants, and male mammals, meiosis begins with the duplication of the cell’s chromosomes (a). Meiotic recombination is then carried out by the formation and resolution of Holliday junctions (b). To complete meiosis, the cell divides into two cells with two sets of chromosomes each (c), and then division occurs again as each of those cells divides into one cell with one set of chromosomes each (d).

An enzyme complexes with the nicked strands, separating them from their intact antisense strand. This enzyme-strand complex then searches for a similar sequence on the opposing chromosome, and the nicked strands cross over and pair with the antisense strand of the homologous duplex. This cross-over region between duplexes is named the Holliday junction, proposed by Robin Holliday in 1964 [3] as a model for aberrant segregation of genes in fungi [See Figure 1.2 for the recombination schematic from Holliday’s paper].

Holliday’s theory describes the process by which DNA strand exchange occurs, and predicts junctions exist in a parallel conformation with opposing strands oriented in the same direction, as visualized by the
arrows in Figure 1.2. However, structural data from in-vitro experiments show four-way DNA junctions preferentially adopt stacked anti-parallel conformations [see the arrows in Figures 1.8, 1.9 and 1.10 for schematics of DNA junctions with the experimentally determined polarity]. If homologous chromosomes are aligned with the same polarity prior to strand invasion yet the minimum energy structure of the junction as determined experimentally is anti-parallel, then chromosomes must undergo major conformational changes to adopt this structure during recombination. Uncovering the principle behind this disagreement between experimental and biological data is important for understanding how DNA junctions evolve dynamically in-vivo and for manipulating DNA for genetic and nanotechnological applications.

After strand invasion, junctions undergo branch migration, during which two opposite “arms” increase while the other two arms decrease in number of base pairs, moving the intersection point between the two strands. A schematic of this process is visualized in Figure 1.4. Proteins such as the E. Coli RuvAB complex and eukaryotic Rad51 and Rad54 can accelerate branch migration [4], and the action of these proteins depends on the DNA sequence of the two strands at the branch point.

Locally, DNA junctions exist in one of three major conformational classes: open and two stacked isoforms [See Figures 1.8, 1.9 and 1.10 for examples of these isoforms]. The two stacking conformers correspond to the stacking of different pairs of arms and share the open conformation as an intermediate state. This is observable using a multitude of experimental techniques [5–11]. In order for branch migration to occur, the junction must be in an open configuration, in which the arms are in a square-
planar configuration. As Overmars et al. note, “The junction...crosses a kinetic barrier into the transient high-energy open form, which does allow for branch migration. After one or more migration steps the molecule crosses the barrier again, either in a forward direction to yield the second stacked X-conformer or backwards to the original arrangement of the stacked arms of the junction” [12]. A proposed free energy landscape for junction migration and conformer transitions is visible in Figure 1.3.

In order for chromosomes to separate, DNA junctions must be resolved into two separate duplexes. Depending on the directionality of this resolution, each resulting chromosome can have small or large regions of genetic material incorporated from the other, corresponding to two types of recombination intermediates outlined in Figure 1.5. As stated by McKinney et al. in 2003, “local DNA sequences can determine the outcome of genetic recombination via the bias in the stacking conformation” [10]. These recombinants are an important pathway through which organisms regulate the exchange of genetic material between chromosomes, increasing genetic diversity in offspring and species’ populations. How DNA sequence affects junction migration, binding specificity of proteins, and conformational specificity is an active area of research. The following chapter serves as an introduction to major discoveries about DNA junctions and the current state of the field.

Figure 1.3: Junction conformer transition and branch migration energy landscape proposed by McKinney et al. [10]. The red and green lines are the free energies along the branch migration and conformer transition reaction coordinates. The lines intersect at the open conformation, which is believed to be an intermediate for both the stacked conformer transitions and for branch migration.
Figure 1.1.2: Homologous recombination begins with the nicking of two homologous chromosomes, which can lead to the formation of a crossover region, known as a Holliday junction. Nicks are introduced (b) to homologous strands, allowing strand invasion into the opposite chromosome (c-d). Ligation of the crossing strand to their new partners produces a heterologous duplex, containing material from both chromosomes (f). (Voet, Voet et al. 2008)

Figure 1.4: This image depicts the junction formation process from the creation of nicks (a) in each of the homologous chromosomes, to strand invasion (b, c, and d), to branch migration (e and f). After branch migration, junction resolution occurs, depicted in Figure 1.5 below. Adapted from Voet et al. [13]

Figure 1.5: Following junction migration, the junction must be resolved into two continuous strands. Resolution of DNA junctions can occur in two different ways (j). In one way, a junction-resolving protein introduces incisions along the continuous strands (the vertical cut), and in the other way, the incisions are made along the invading strands (the horizontal cut). The product of these two different resolution processes is visible in (l). When the incision is made along the continuous strands, the resulting chromosomal products have traded arms, and the chromosomes might be more genetically different than when the incisions are made along the invading strands. Adapted from Voet et al. [13]
1.2 DNA Junction Studies

DNA junctions form during other recombination events such as damage repair, replication, and chromosomal rearrangement. As the dynamics of these processes span a range of time and length scales, scientists use various techniques to study them in-vitro, including optical and electron microscopy, gel electrophoresis, Förster resonance energy transfer (FRET), nuclear magnetic resonance (NMR), chemical or enzyme probing, and molecular modeling. Below we review experiments and their results as they relate to the research carried out in this thesis.

In 1983, Seeman et al. [14] suggested the sequence of the strands that comprise the junction could be engineered to ensure junction formation and limit or enhance the resulting junction’s ability for branch migration. This group developed algorithms to find sequences to form DNA junctions which do not permit branch point migration. These “immobile” junction sequences are used to study the thermodynamics and dynamic properties of DNA junctions in the absence of branch migration, whereas “mobile” junction sequences are used in studies of junction migration.

Duckett et al. and Diekmann and Lilley [5, 6] ran experiments using four DNA fragments which anneal into stable junctions with one restriction enzyme (HindII, EcoRI, Xba, or BamHI) target sequence along each arm. This led to standard nomenclature of DNA junction arms as H, R, X, and B. The experiments they carried out studied DNA junction fragment motion through electrophoretic polyacrylamide gels after junctions were cleaved by the restriction enzymes, both in the presence and absence of magnesium cations. Fully intact junctions were observed to have slow mobility through gels, whereas junctions whose arms had been cleaved by restriction enzymes had faster motion. Figure 1.6 shows two different products that result from cleaving of junction arms by restriction enzymes.

![Figure 1.6](image_url)

**Figure 1.6:** The locations of the restriction enzyme target sequences are visible on the left. Six possible cleavage products are possible. These six products have the same number of base pairs and molecular weight, but differ in their mobility through gels due to their geometries, i.e. linear structures have higher gel mobility than structures which have folded. See Figure 1.7 for all possible restriction products and their relative mobilities. Adapted from Duckett et al. [5].
They then performed the experiment omitting magnesium and observed results consistent with restriction enzymes cutting junctions in a square-planar structure. They conclude that without magnesium, “phosphate repulsion overcomes the free energy available from helix stacking and the junctions adopt a square-planar conformation with fourfold symmetry”. [See Figure 1.8 for a schematic of the transitions from open to stacked conformations]. They also observed an exchange in helix-stacking partners resulting from changing the terminal (closest to the junction) G-C base pair of arm H to a C-G and the terminal A-T of arm X to T-A. This author’s major contributions are that DNA junctions adopt a stacked conformation at high salt concentration and that the preference for helix-stacking partners depends on the sequence at the center of the junction.

The phenomenon of local ion concentrations changing the free energy minimum conformation from square-planar to stacked is a universal observation in DNA junctions. Hinckley et al. [15] note, “ionic strength directly determines the flexibility of dsDNA via the degree of shielding of electrostatic repulsion between phosphate sites”. As Ortiz-Lombardia et al. [16] explain, “in the absence of cations the junction is believed to be extended, with its arms unstacked, and to have a four-fold planar conformation. However, in the presence of metal ions (as occurs in physiological conditions), the arms stack in pairs and form a two-fold nonplanar junction, known as the X-stacked model.” This consensus motivates further research on the structure and dynamics of junctions with mobile and immobile junction sequences and different stacking preferences in and around physiological salt concentrations.

Förster resonance energy transfer (FRET) [17], sometimes described as the “molecular ruler”, is a spectroscopic technique which can resolve distances on the Å scale. This technique measures energy transferred via dipole-dipole interactions between a donor molecule, which absorbs probing light, and an acceptor molecule, which emits measured signals, as a proxy for the distance between donor and
In time resolved single-molecule FRET (or smFRET) experiments from studies which quantify the dynamics of Holliday junction state transitions, molecules with integrated donor and acceptor probes are bound to a substrate, and the acceptor signal for each molecule is recorded as a function of time [19]. Time-resolved signals from probes at the ends of two arms of the junction allow for real-time measurement of the distance between the arm ends. These distances are used to extract quantities such as the angle between the arms or, in experiments of junction sequences capable of branch migration, the location of the junction center. Classification of these data can resolve junction migration and isomerization transition rates.

McKinney et al. [10] used smFRET to analyze transition rates as a function of DNA sequence, types and concentrations of counterions, and temperature, showing that the conformer transition and branch migration processes share the open structure as a common intermediate. In 2004, Joo et al. [19] used time-resolved smFRET to observe a stacking conformer bias in three different DNA junctions as a function of metal ions, observing that the junction arms stack preferentially upon the addition of divalent cations [See Figure 1.9]. Okamoto et al. [11] used a variational Bayes hidden Markov model [See Chapter 2.3 for an introduction to Markov models] method applied to single-molecule FRET trajectories of mobile junctions to classify three junction locations. With their models they investigated branch migration kinetics and reproduced residence times, transition rates, and free energy differences between states.

Experiments from 2016 by Litke et al. [20] depict the difference in FRET efficiency of labels placed at the ends of arms of DNA junctions as a function of ionic concentrations. In Figure 1.10, FRET efficiency increases with increasing salt concentration, plateauing at a saturation value, corresponding to two adjacent arms of the junction stacking. In Figure 1.11, we see a timeseries of FRET energy efficiency as reported by Karymov et al. [21] using mobile DNA junctions with FRET labels at the ends of two of the arms. Since the FRET labels get closer or farther apart when the junction migrates, the FRET efficiency
is directly related to the exact location of the crossover. At low salt concentrations, junction migration is more likely and the residence time at any particular location is short. At higher salt concentrations, migration is less likely and the FRET efficiency plateaus for longer timescales. NMR, x-ray diffraction of crystallized DNA junctions, and molecular dynamics (MD) simulations have been used in addition to FRET to elucidate DNA junction structure and dynamics.

For example, in 1997, Overmars et al. [12] used NMR to observe the dynamics of conversion between the two stacked conformers and found a 71/29 population ratio between them. Scientists have been particularly successful at simulating DNA junctions to compare with experimental data and predict structural and dynamic quantities. MD studies by Wheatley et al., Yu et al., and Wang et al. [7–9] report observations of spontaneous transitions from the open to stacked conformation of DNA junctions at various salt conditions. Wheatley et al. [9] used the all atom AMBER 9 model potential for DNA to study 1DCW, a DNA junction with 10 base pairs per strand. They evaluate the stability of the AMBER 9 model concerning this junction and check agreement between experimental and MD data while predicting structural quantities such as the distribution of water and ions around the junction.

In 2004, Yu et al. [7] carried out all-atom MD simulations at various ionic concentrations to investigate the thermodynamics of junction conformer transitions. The method they employed measures free energy over hundreds of MD simulations. They used two geometric coordinates to describe the global conformation of the junction and observe a local free energy minimum near the tetrahedral form of 30 kcal/mol in simulations lacking Mg$^{2+}$ ions. After adding Mg$^{2+}$ ions, they observe no minimum near the folded...
form and locate one Mg$^{2+}$ ion bound at junction exchange point. The authors posit that “magnesium ions stabilize the stacked-X form and destabilize the open and tetrahedral intermediates”, a conclusion consistent with other literature in the field. They suggest that the disappearance of a local minimum explains that Mg$^{2+}$ ions destabilize the tetrahedral form, which is a possible intermediate in the conformer transition.

Intracellular cation concentrations can regulate stages of the cell cycle, maintain genomic stability, and are an “essential cofactor in almost all enzymatic system involved in DNA processing” [22]. For example, cations aid in stabilizing G-quadruplexes, which are DNA structures important in chromosomal maintenance and restructuring. J. Lee et al. [23] used total internal reflectance fluorescence (TIRF) to observe that increasing K$^+$ ion concentrations correlates with the ratio of folded to unfolded states in G-quadruplexes. The propensity for DNA structures to undergo conformational changes upon addition of cations due to counterion shielding is also evident in double-stranded DNA. MD studies carried out by Bai et al. [24] simulated two 12-bp DNA duplexes joined by a flexible polyethylene glycol (PEG) tether at various ionic conditions to observe this. [See Figure 1.12.]

![Figure 1.12: Data from Monte Carlo simulations of two dsDNA strands linked by a PEG tether show increased flexibility at high salt concentrations. The left image is a schematic and sequence of the tethered duplex system. The images on the right show data, from left to right and top to bottom, of 0.02, 0.06, 0.17, 0.3, and 2.0M concentrations of monovalent ions. The colored dots represent the location of the distal end of the other DNA duplex, and the color represents the energetic difference between the conformer and the minimum-energy conformer observed in the ensemble. Adapted from Herschlag et al. [24].](image-url)

Ions can assist in the function of junction-interacting proteins by inducing conformational changes in both proteins and DNA. Some proteins bind to junctions only when the junction is in certain conformations, and local salt concentrations and DNA sequences can drive the DNA to adopt those conformations. Thus DNA-ion-protein interactions are motivations at the forefront of studies involving the dynamics of ion binding to DNA. Having an in-depth understanding of ionic interactions with DNA junctions in the absence
Figure 1.10: Conformations have an expected FRET efficiency as either low or high (a). FRET efficiency plotted as a function of ion concentration for magnesium, calcium, cobalt hexamine, terbium, europium and neodimium ions show that increasing ion concentrations leads to higher FRET efficiency (b,c). The points are experimental measurements, and the black lines are curves fit to the experimental data to calculate ion-junction binding dissociation constants for each ionic species. Energy transfer efficiency increases as Mg$^{2+}$ concentrations increase from 0 to 400M. Adapted from Litke et al. 2016 [20].

(a) Low [Mg$^{2+}$]. This timeseries implies branch migration is more frequent at lower salt concentration. The authors note that individual junction steps happen so fast they cannot be fully resolved.

(b) High [Mg$^{2+}$]. The authors explain that the plateau of FRET values at the black arrow is a period during which the junction is energetically trapped and unable to migrate.

Figure 1.11: FRET efficiency from a mobile junction, where FRET efficiency is a proxy for the distance between the ends of the arms. The upper blue curves are FRET efficiency sampled every 6ms, and the lower red curve is the blue curve after processing through a noise removal algorithm. These data show that increased Mg$^{2+}$ concentrations serve to limit branch migration processes. Adapted from Karymov et al. [21].
of proteins is important when beginning to understand protein-junction interactions in the presence of ions. References [16, 21, 25–29] describe studies on the binding specificity of junction-binding proteins.

DNA junction ion binding locations and mechanisms have been proposed upon interpretation of experimental and computational results. Litke et al. [20] interpret results from FRET and isothermal titration calorimetry experiments of DNA junctions as evidence that the junction center most likely has two Mg$^{2+}$-binding sites. Wheatley et al. [9] also predicted the distribution of counter-ions and solvating water in a DNA junction using MD simulations. Hyeon et al. [25], through simulations and smFRET experiments, observe dynamical heterogeneity of DNA junctions as a function of solvated Mg$^{2+}$ ions. [See Figure 1.13 for their calculations and Chapter 3.1 for our analogous results.] They understand their data as showing that “the secondary structure of the Holliday junction, particularly at the internal multiloop (which mediates the conformational transition between the two isoforms), is pinned by Mg$^{2+}$ ions”, and that various open-state topologies with different internal multiloops act as transition states between stacked isoforms. Lipfert et al. [26] provide a summary of current research and theoretical understanding of DNA interactions with ionic solvents, but DNA-ion binding processes such as junction-ion interactions, and their physical principles, implications and interpretations are still not completely understood.

![Figure 1.13: Radial distribution function between Mg$^{2+}$ and phosphates in a DNA junction with 10 base pairs per arm, calculated separately for phosphates at the terminal, middle, and junction locations of the strands, showing preferential binding of Mg$^{2+}$ at the junction. Data is from 100ns MD simulation at 310K, adapted from Hyeon et al. [25].](image)

### 1.3 DNA Junction Melting

Theoretical understanding of DNA junction assembly and melting has evolved significantly with the advent of more powerful computational resources and experimental techniques. In 1977, Kallenbach and Berman posited that “short duplex oligonucleotides tend to denature in all-or-none fashion, and significant populations of intermediates arise only in longer chains” [14]. In 1996, Peyrard and Dauxois [27] used...
computing resources at the Advanced Computing Laboratory of Los Alamos National Laboratory to simulate a 1-D system of particles coupled by nearest-neighbor anharmonic interactions and subjected to an on-site Morse potential. They show that “one-dimensional phase transitions can exist for mechanical systems of particles with a positions characterized by a continuous variable”, a theory which they describe as applicable to the DNA melting phase transition. In 2018, Ouldridge et al. used a minimal one-site-per-nucleotide coarse-grained model for DNA to study DNA junction self-assembly and melting. They find that junction assembly is “most successful in the temperature window below the melting temperatures of the target structure and above the melting temperature of the misbonded aggregates.” [28]

In grounding work for our study, Wang et al. [8] used an implicit ion model of DNA to simulate junction melting at various salt concentrations from 10 to 500mM [Na⁺]. They compared melting temperatures \( T_M \) calculated from replica exchange molecular dynamics (REMD) simulations with \( T_M \) inferred from absorption experiments on junctions of the same sequence, and found that their model closely mimics the experimental dependence of \( T_M \) on Na⁺ concentration. Figure 1.14 shows the agreement between the experimental and simulated melting temperatures.

![Figure 1.14](image)

**Figure 1.14**: Data from absorption experiments on DNA junctions (b) matches closely to data from the implicit ion model of 3SPN.2 (a). For the theoretical calculations, the authors carried out 17-20 REMD simulations for 6s each, with exchanges between replicas attempted every 200ps. The melting temperature \( T_M \) increases monotonically as a function of salt concentration for both the absorption experiments and the simulations. \( T_M \) differs by < 3% between implicit ion 3SPN.2 simulations and absorption experiments for all salt concentrations. Adapted from Wang et al. [8].
1.4 DNA Nanotechnology

DNA junctions have found novel applications in the field of nanotechnology. A popular method of developing nanotechnology is self-assembly, by which nanoparticles with binding affinity are brought together to form nanostructures. Because of DNA’s ‘sticky-ended cohesion’ binding mechanism, it has proved useful in building theoretical and experimental frameworks for self-organizing DNA-based materials such as lattices, nanotubes, and programmable chemical reaction networks. In 1998, Winfree et al. [29] fabricated 2D DNA junction arrays using two-unit and four-unit polymers, characterizing their constructs using gel electrophoresis and observing the patterned crystals using atomic force microscopy (AFM). In 1999, Mao et al. [30] synthesized 1D and 2D DNA lattices using rhombus-like molecules consisting of four junctions, and subsequently visualized these arrays by AFM. Benson et al. [31] demonstrated a method to generate arbitrarily shaped scaffolded DNA nano-structures, and visualized these 2D sheets in a 10mM MgCl2 buffer using AFM [See Figure 1.15 for depictions of 2D DNA junction nanolattices from these studies].

![Figure 1.15](image-url)

**Figure 1.15:** Two-dimensional DNA nanolattices made up of four-way junction monomers, visible through AFM. These lattices can be used as supports for nanostructures such as water purification systems [32], controllable membrane channels [33], and molecular electronic or plasmonic circuits [34]. The first diagram, from [31], shows a schematic of DNA tile assembly. The following two pictures, from [30] and [29], respectively, are AFM images of constructed DNA nanolattices.
DNA junctions can also be used in nanotiles which can self-assemble into tubes with lengths on the order of micrometers. In 2013, Zhang et al. [37] explored approaches to trigger DNA nanotube growth made of DNA double-crossover tiles containing a rigid core of two parallel DNA duplex helices, verifying their predictions on self-assembly and exploring kinetic models of catalyzed tube formation through time-lapse TIRF (total internal reflection fluorescence) microscopy and AFM. In 2017, Jorgenson et al. [38] used AFM, transmission electron microscopy (TEM), and fluorescence microscopy to visualize the architecture of various assembled DNA nanotube structures. Figure 1.16 shows schematics used to develop DNA nanotubes created from tiles consisting of two DNA junctions with sticky ends.

**Figure 1.16:** DNA nanotubes made up of monomers which include four-way junctions to increase the nanotube’s stability. The first two pictures are from Zhang et al. 2013 [35] and the second picture is from Jorgenson et al. 2017 [36].

It has been verified experimentally that DNA junctions can be integrated into chemical reaction networks which perform various computational and nanomechanical functions. In 2018, Cherry et al. [39] used branch-migration domains as subunits in a winner-takes-all neural network that identifies the number represented by a hand-written digit with high precision. A summary of the development of DNA nanotechnology can be found in [40] and references within. Understanding the mechanisms by which DNA junctions melt and undergo conformational transitions, and the environmental conditions which determine their conformational specificity and melting temperature and dynamics is important for pushing this burgeoning field forward.
1.5 Motivation and Objectives

The theory of DNA junction dynamics and junction-protein interactions have been thoroughly investigated using experimental methods as well as all-atom and coarse-grained MD simulations of junctions of various size and mobility in various solvent environments. In this thesis, we investigate the dynamics of a mobile and an immobile DNA junction using a coarse-grained implicit ion model for DNA by simulating DNA junctions at a range of temperatures and in various ionic concentrations on timescales large enough to gather statistics on equilibrium fluctuations, melting dynamics, and junction migration. In Chapters 3.1 and 3.3 we analyze the equilibrium structure at Na\(^+\) and Mg\(^{2+}\) concentrations around physiological levels. We detect preferential ion binding at the junction core and calculate junction isoform population distributions and transition rates to compare with ion binding and isoform population distribution data present in the literature. In a multidisciplinary collaboration, researchers use experimental techniques to probe the melting process in parallel with our simulations. Chapter 4 elucidates the dynamic process of junction melting by comparing fluorescence melting data on the stability of specific bases with data from our simulations. Finally, we simulate junction migration to show the explicit ion 3SPN.2 model returns qualitatively consistent results on the salt-concentration dependence of migration probability. From our data, we highlight the advantages and potential shortcomings of the 3SPN.2 coarse-grained implicit ion model for DNA to simulate DNA junctions and offer useful insights for future computational and experimental studies of DNA junctions.
Chapter 2

Model and Simulations

"The notion that a numerical result should depend on the relation of object to observer is in the spirit of physics in this century and is even an exemplary illustration of it."


In 1953, Metropolis et al. computed the first simulation of a liquid using the Metropolis Monte Carlo technique at Los Alamos National Laboratories on then one of the most powerful computers available [37]. In 1957, Alder and Wainwright simulated a system of hard spheres by integrating classical equations of motion [38] and in 1964, Rahman solved the equations of motion for a set of Lennard-Jones particles [39]. Since this groundwork, biologists’ interest in MD has surged due to its usefulness in explaining links between biomolecular structure and function. Simulating over relevant timescales, researchers explore how macro and micro structure contribute to functional mechanisms.

MD simulations integrate the equations of motion of a configuration of particles according to forces calculated from locations, masses, charges, and interaction potential gradients. Simulating many-particle systems is expensive since simulation time scales quadratically with number of particles. Coarse-graining, by modeling groups of atoms as one particle, grants the opportunity to simulate large systems for timescales intractable using atomistic models. Coarse-grained models have drawbacks; they cannot resolve details smaller than the particle discretization and must be parameterized to match system characteristics, which are potentially arduous to determine experimentally. Biological systems are usually immersed in solvents, most commonly water and ions, which can further increase simulation time and space requirements.
Section 2.1 - Coarse-grained DNA Model

2.1 Coarse-grained DNA Model

To simulate DNA junctions we use 3SPN.2, a coarse-grained three-site per nucleotide model for DNA developed by the de Pablo group at University of Chicago [15, 40, 41]. The 3SPN.2 model represents each DNA base with three particles. For each base, one particle represents the phosphate group ($\text{PO}_4^{3-}$), another particle represents the backbone sugar ($\text{C}_5\text{H}_{10}\text{O}_4$), and another particle represents the nucleoside, each placed at the center of mass of the corresponding moiety [See Figure 2.1]. Using a top-down parameterization relying on experimentally measured duplex and hairpin melting temperatures, base-step and base-stacking free energies, and equilibrium bond lengths, bend angles, and dihedral angles, this model captures properties of B-DNA such as sequence and salt-concentration dependent duplex and hairpin melting temperatures, major and minor groove widths, and persistence length of single-stranded and double-stranded DNA. With anisotropic potentials between force sites, this model can accurately simulate DNA to study DNA hybridization, DNA-protein binding, and nano-engineered DNA-hybrid materials such as DNA origami and DNA liquid crystals [15].

Figure 2.1: All atom (left) to coarse grained (right) representation of a 10 base pair long strand of B-DNA using the 3SPN.2 model. The middle image shows the mapping of the coarse-grained model on top of the atomistic model, with the coarse-grained force sites placed at the center of mass of the corresponding moiety. Gold particles represent phosphate group sites, yellow particles represent sugar sites, and purple and blue particles represent nucleoside sites. Adapted from [42].

Potentials, parameterized from experimental data, are categorized as either bonded or non-bonded. Bonded potentials include harmonic and anharmonic linear bond potentials, harmonic angle potentials, and Gaussian well dihedral potentials. Melting temperatures $T_m$ for various B-DNA sequences were obtained via UV absorbance measurements, and inter-strand non-bonded interactions were adjusted until the free energies of hybridized and dehybridized states were equal at $T_m$. Intra-strand base stacking free energies were measured by nicking DNA duplexes and examining their relative electrophoretic mobility, and stacking potential strengths were adjusted until the simulated free energy of stacking agreed with the
Section 2.1 - Coarse-grained DNA Model

Experimental data. Non-bonded potentials consist of excluded volume contributions, intra-strand base stacking, inter-strand cross-stacking, base pairing interactions, and the electrostatic potential. Equilibrium bond lengths, bend angles, and dihedral angles are obtained from the B-DNA fiber crystal structure [41]. Relative entropy coarse graining, in which “reference all-atom simulations are targeted in an iterative approach to find the effective coarse-grained potential that preserves the information in the all-atom ensemble” [15], was used to model the pair correction and the bond and bend angle potentials. A repulsive Lennard-Jones potential of the form

\[ U_{\text{exc}} = \sum_{i<j} \begin{cases} 
\varepsilon_r \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \varepsilon_r & r < r_c \\
0 & r \geq r_c 
\end{cases} \]  

(2.1)

accounts for core repulsion as well as attraction between all sites that do not participate in bonded or base pairing interactions, i.e., between ionic and nonionic sites. The nonbonded effective potentials between ions are modeled with a Coulombic interaction and a correction term, \( U_{\text{corr}} \), as follows

\[ U_{\text{ion-ion}}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0\epsilon(T)r_{ij}} + U_{\text{corr}}(r_{ij}) \]  

(2.2)

"where \( q_i \) and \( q_j \) are the charges of the \( i^{\text{th}} \) and \( j^{\text{th}} \) ions, \( \epsilon_0 \) is the dielectric permittivity of vacuum, \( \epsilon(t) \) is the solution dielectric, and \( r_{ij} \) is the intersite separation” [41].

This model was previously used by Prytkova et al. [43] to simulate DNA melting in small-molecule-DNA-hybrid dimer structures (SMDHs). This model has also been coupled with the mW-ion model by Demille et al. [42] to model DNA with explicit solvation by water and ions. A previous version of the 3SPN.2 model, which included ions implicitly in the potentials as opposed to an explicit representation, was used by Wang et al. [8] to validate this previous version of 3SPN.2 as a model that accurately reproduces experimental data. Their paper confirms the implicit ion model’s ability to reproduce experimental results about DNA junctions, including the preference for a stacked isoform at high salt, the existence of a square-planar intermediate between stacked states, and salt concentration dependent melting temperatures. Figure 1.14 shows junction melting curves from the implicit ion 3SPN.2 model at various salt concentrations and from experimental fluorescence melts. Part of the work described in this thesis verifies the 3SPN.2 explicit ion model returns similar melting dynamics and isoform distributions for DNA junctions that the 3SPN.2 implicit ion model does.

All simulations were performed with serial computation using LAMMPS (Large-scale Atomic/Massively Parallel Simulator), an open-source, general purpose MD simulation engine developed and distributed by Sandia National Laboratories. We input the initial configuration, temperature, a thermostat random
seed to assure trajectories do not replicate data, 3SPN.2 pair coefficients and potentials, neighbor list parameters, initial velocity distributions, time step, simulation length, output file formats, and output rate. Output data include thermodynamic information such as temperature, kinetic, potential and total energy, pressure, and volume. For our analyses we investigate the trajectories, which capture the positions of all particles in the system at each snapshot.

2.2 Molecular Simulations

We simulated three different ensembles of DNA junctions, split into these categories:

1. An immobile junction ensemble at constant temperature, to determine junction isoform population distributions at various salt concentrations.

2. An immobile junction ensemble at physiological salt concentration above the melting temperature, to describe the dynamics of the melting process and compare with fluorescence melting data.

3. A mobile junction ensemble to investigate the explicit ion 3SPN.2 model’s capacity to predict the dependence of junction mobility on salt concentration.

For the equilibrium and melting ensembles, we simulated a 34 base pair junction J3 previously studied with the 3SPN.2 implicit ion model by Wang et al. [8] and fluorescence melting experiments by Savage, McDonald and Litke [4,44,45]. The terminological J3 refers to the choice of the base sequence at the core of the junction, which determines the tendency for the junction to stack. The immobile junction sequence has a total 136 base pairs and 404 force sites.

<table>
<thead>
<tr>
<th>Strand</th>
<th>Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>CCTCCGTCTTAGCAAGGGGCTGCTACCGGAAGGG</td>
</tr>
<tr>
<td>H</td>
<td>CCCTTCCGTTAGCAGCCTGAGCGGTGGTTGAAGG</td>
</tr>
<tr>
<td>R</td>
<td>CTTTCAACCACCGCTCAACTCAACTGCAGTCTGG</td>
</tr>
<tr>
<td>X</td>
<td>CCAGACTGCAGTTGAGTCCTTGCTAGGACGGAGG</td>
</tr>
</tbody>
</table>

Table 2.1: Sequence of strands that comprise the mobile junction.

For the equilibrium ensemble, we solvated the junction in a 200 Å box with periodic boundary conditions at 16 different combinations of Mg$^{2+}$ and Na$^+$ concentrations. The salt concentrations and the number of solvating Mg$^{2+}$, Na$^+$, and Cl$^-$ ions at each of these concentrations are given in Table 2.2.

At each of these 16 salt concentrations we simulated 10 different trajectories for 1 μs, starting from
the open configuration. Particle velocities were fixed at each timestep using a Langevin thermostat at 310K with a damping parameter of 400 time units with the Gronbech-Jensen/Farago formulation so that the total force on each atom has the form

\[ F = F_c + F_j + F_r = F_c + \left( \frac{m}{\text{damp}} \right) v + c \sqrt{\frac{k_B T_m}{dt * \text{damp}}} \]  

(2.3)

Langevin dynamics mimics the quasi-random movement of particles within a fluid due to collisions. Since our simulation box contains no water particles, Langevin thermostatting enables solvent-induced diffusion dynamics, i.e. a thermal noise accounting for the lack of solvent. All force site coordinates were unwrapped, so if a particle passes through a periodic boundary, the particle position was logged as if it had not been unwrapped into the periodic box. Particle positions were saved every 20,000 timesteps, corresponding to 0.2ns of simulated time between trajectory snapshots.

For the melting ensemble, we first simulated the same junction as in the equilibrium ensemble at a range of temperatures below and above the known melting point to determine the best temperatures at which to analyze melting dynamics. For each temperature from 320K to 380K in steps of 5K we equilibrated 20 replicas of the open conformation of J34 in a 200Å box with periodic boundary conditions at [Na\(^+\)] = 177.2mM (854 Na\(^+\) and 722 Cl\(^-\) sites) and [Mg\(^{2+}\)] = 0mM for 20ns. Then we heat the system by increasing the thermostat temperature and simulate an additional 1μs. At each of these temperatures we calculate the mean fraction of broken bases and plot it versus temperature [See Figure 2.2]. From this curve we determine temperatures above 355K have a mean fraction of broken bases above 0.5. Then we launched 100 simulations at each of 4 different temperatures (360, 365, 370, and 375K) above the melting point, using the same simulation protocol as previously; the simulations begin with a 20ns equilibration at 310K, followed by 0.5μs at the higher temperature. Data from melting simulations was compared with data from fluorescence melts of DNA junctions formed by strands with the same sequences we simulated.

### Table 2.2: Salt concentrations and number of solvating particles at each concentration for the equilibrium ensemble.

<table>
<thead>
<tr>
<th>[Na(^+)] (mM)</th>
<th>37.3</th>
<th>127.2</th>
<th>227.2</th>
<th>327.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Mg(^{2+})] (mM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>180</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>0.8</td>
<td>4</td>
<td>180</td>
<td>56</td>
<td>4</td>
</tr>
<tr>
<td>9.9</td>
<td>48</td>
<td>180</td>
<td>144</td>
<td>48</td>
</tr>
<tr>
<td>49.8</td>
<td>240</td>
<td>180</td>
<td>528</td>
<td>240</td>
</tr>
</tbody>
</table>

Number of Mg\(^{2+}\), Na\(^+\) and Cl\(^-\) particles at each salt concentration.
Using fluorescent nucleotide analogs inserted in place of regular nucleotides at specific bases of interest along the junction, fluorescence melt data shows the relative stability of specific bases [See Chapter 4 and figures therein].

Figure 2.2: Melting simulation protocol and preliminary data to determine optimal temperatures to simulate melting dynamics. For the blue dotted curve, 20 replicas were launched at temperatures from 320 to 380K in steps of 5K. For the orange curve, we simulated 100 replicas at temperatures from 360 to 380K in steps of 5K. Our analysis of melting dynamics is carried out using these higher temperatures.

For the mobile junction ensemble, we use 44 base pair DNA sequences [See Table 2.3] developed by Wujie Wang to study extended branch migration through polyAT regions at the junction center, as described by Karymov et al. [21]. At each of the salt concentrations at which we simulated the immobile junction equilibrium ensemble [See Table 2.2], we ran 5 independent simulations which start with 0.05ns of equilibration followed by 1.5 microseconds of production. Trajectory data was dumped every 20,000 timesteps, or 0.2 nanoseconds.

<table>
<thead>
<tr>
<th>Strand</th>
<th>Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>TAAGCTTGCAAGCATATATATATATCTCGTAATTTCCGGTTA</td>
</tr>
<tr>
<td>H</td>
<td>TAACCGGAAATTACGAGATATATATAGATGCAATGCAAGCTTC</td>
</tr>
<tr>
<td>R</td>
<td>GAAGCTTGCAATGCAATATATATATATCTCACTGAGGCTTAGG</td>
</tr>
<tr>
<td>X</td>
<td>CCTAGCGGCTACGTAATATATATATATATATATATGCTTGAAGCCTTA</td>
</tr>
</tbody>
</table>

Table 2.3: Sequences of strands that comprise the mobile junction. The poly-AT region at the center of each strand allows this junction to exchange base pairs along strands more easily than the immobile junction, which contains both purine (A and G) and pyrimidine (C and T) base pairs at the junction center.
Section 2.3 - Hidden Markov models and our implementation

2.3 Hidden Markov models and our implementation

Hidden Markov models describe systems with internal states governed by the “limited horizon assumption” that the following state depends only on the previous state. An application for implementing hidden Markov models is a coarse model for weather, which can have one of three possible states: rainy, cloudy, or sunny. Tomorrow’s weather depends on today’s; if today is cloudy we have a 50% chance of rain tomorrow, a 25% chance of sun and a 25% chance of another cloudy day. These transition probabilities and the probabilities for tomorrow’s weather given today is rain or sunny compose a transition matrix $T$ which determines weather evolution. A cloudy day today, the unit vector $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$, multiplied by $T$, returns a vector $\begin{bmatrix} 0.5 & 0.25 & 0.25 \end{bmatrix}$ representing tomorrow’s weather possibilities. Continued matrix multiplication operations yield the steady state describing the normalized probability of each weather condition after infinite time. The steady state can also be calculated directly from the transition matrix with the equation

$$(T^T - (I + 1))^{-1} \cdot v \quad (2.4)$$

where $I$ is the identity matrix, $1$ is the matrix of all ones, and $v$ is a unit vector the same size as $T$.

These models are useful for classification of molecular structures because they can “automatically identify key conformational states in a way that is unbiased, human-readable, convenient, and rigorous” [46]. A literature review of hidden Markov model applications to biophysics is available from Shukla et al. [46]. In 2006, McKinney et al. [47] used Markov models to analyze FRET signals from fluorescent labels placed at the ends of the junction’s arms, showing DNA junction isomerization. These models were trained to recognize a two-state system with Gaussian emissions, and identified state probabilities, transition rates, and emission spectra for each isomerization state. Okamoto et al. [11] also used Markov models applied to single-molecule FRET trajectories of mobile junctions to study junction migration dynamics. In 2002, Thayer et al. [48] implemented Markov models “based on probabilistic roll/tilt dinucleotide models of sequence-dependent DNA structure” to classify protein binding sites. They then proceeded to test the trained models ability to predict the potential for binding sites in unknowns. As Shukla et al. note, “there are still challenges with identifying the best decomposition of conformational space...and connecting [Markov models] to experimental data” [46]. In our discussion of the data output by our Markov models, we consider the limitations and errors imposed by using this method to analyze our results.
For our implementation of hidden Markov state models, we used the Mathematica function EstimatedProcess, which accepts two inputs: the timeseries of measurements (either angle, distance, or both angle and distance data [See Chapters 3.2.2 and 3.2.1 for how we take these measurements]), from every simulation at specific salt concentrations as a table of TemporalData, and a HiddenMarkovProcess. Prior to running the EstimatedProcess, we input the number of states and the functional form of the emission probability densities into the HiddenMarkovProcess. These emission probability densities can take any functional form, the most common being exponential and Gaussian. Our emission probability densities are encoded as a ProductDistribution composed of either four (for when we analyze angle data or distance data) or eight (for when we analyze angle and distance data together) NormalDistributions. Thus each “state” [see Figure 3.3 and Chapter 3.2.4 for the explanation of how we label and classify these states]) has either four or eight Gaussian emission probability densities.

The task of the training script is learning the emission probability density parameters by maximizing the log-likelihood of the inputted sequence of observations. After training, the code returns the transition matrix between states and the mean and standard deviation of each Gaussian. From the transition matrices and Equation 2.4 we derive the stationary state of the system which constitutes the isoform population distributions at each salt concentration. A previous implementation of Markov models to DNA junction MD simulations as described by Hyeon et al. [25] use the number of states as a variable parameter; here, we classify three states (Iso1, Open, Iso2) as in Wang et al. [8], for the simplicity of analysis using this technique and to compare this analysis to other classification methods outlined in Chapters 3.2.2 and 3.2.1.
Equilibrium Structure and Dynamics

"Under normal conditions the research scientist is not an innovator but a solver of puzzles, and the puzzles upon which he concentrates are just those which he believes can be both stated and solved within the existing scientific tradition."

Thomas S. Kuhn, *The Structure of Scientific Revolutions*, 1962

3.1 Radial distribution function

The radial distribution function, or density-density correlation function, shows the probability a particle from one chemical species is a distance from a particle of another species, and is calculated by:

\[
g(r) = \frac{V}{2\pi r^2 N^2} \left[ \sum_i \sum_j \delta(r - r_{ij}) \right]
\]

where \( V \) is the simulation box volume, \( N \) is the number of atom pairs, the sums are over all the atoms in each species, and \( r_{ij} \) is the distance between the \( i^{th} \) atom in the first species and the \( j^{th} \) atom in the second species. We split the phosphates into two “species” to investigate preferential ion binding to the junction center. Figure 3.1 and 3.2 show \( g(r) \) for Na\(^+\) and phosphates and Mg\(^{2+}\) and phosphates at physiologically relevant salt concentrations ([Na\(^+\)] \geq 100).
Section 3.1 - Radial distribution function

Figure 3.1: As $r$ approaches the box size (200 Å), $g(r)$ approaches the ionic concentration. For $\sqrt{3} \times 200 \, \text{Å} > r > 200 \, \text{Å}$, $g(r)$ approaches zero because no ions exist outside the corner-to-corner distance. The radii of the peaks of $g(r)$ remain constant across salt concentrations and pair selections, while the peak heights vary. $g(r)$ decreases as Mg$^{2+}$ concentration increases as a consequence of Mg$^{2+}$ ions displacing Na$^+$ ions from phosphate sites. The first peak in $g(r)$ is higher for phosphates at the center of the junction, consistent with $g(r)$ calculations from Hyeon et al. [25] [See Figure 1.13].

Figure 3.2: The first peak in the Mg$^{2+}$ $g(r)$ begins at a larger radius than the first peak of the Na$^+$ $g(r)$ most likely because the Lennard-Jones radius of Na$^+$ in 3SPN.2 is 2.494 Å, as opposed to 4.0 Å for Mg$^{2+}$. For our calculations we convert from a probability density into a molar density of ions to make a one-to-one comparison between the radial distribution of ions and the total box ion concentration. By observing the radius at the largest peak in $g(r)$, we obtain the most likely distance an ion resides from any given phosphate in the junction, and by integrating the distribution function over the first peak, we obtain the mean number of nearest neighbor ions.
Section 3.1 - Radial distribution function

To predict nearest-neighbor solvation numbers, we use the following integration:

\[ N = \int_0^{r'} 4\pi \rho r^2 g(r) dr \]  \hspace{1cm} (3.2)

where \( g(r) \) is in units of number density, \( \rho \) is the number density of ions in the box, and \( r' \) is the radial distance to the end of the first peak. The mean number of ions in the first solvation shell of phosphates at the center of the junction and phosphates not at the center of the junction are in Tables 3.2 and 3.1.

### Table 3.1: Number of first nearest Na\(^+\) neighbors

<table>
<thead>
<tr>
<th>[Mg(^{2+})] (mM)</th>
<th>0</th>
<th>1</th>
<th>10</th>
<th>50</th>
<th>0</th>
<th>1</th>
<th>10</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.19</td>
<td>1.04</td>
<td>0.34</td>
<td>0.12</td>
<td>0.83</td>
<td>0.76</td>
<td>0.29</td>
<td>0.11</td>
</tr>
<tr>
<td>100</td>
<td>4.74</td>
<td>4.41</td>
<td>1.56</td>
<td>1.38</td>
<td>3.37</td>
<td>3.19</td>
<td>1.23</td>
<td>1.17</td>
</tr>
<tr>
<td>200</td>
<td>7.82</td>
<td>7.56</td>
<td>6.20</td>
<td>4.15</td>
<td>5.30</td>
<td>5.21</td>
<td>4.67</td>
<td>3.44</td>
</tr>
<tr>
<td>300</td>
<td>13.61</td>
<td>13.4</td>
<td>11.27</td>
<td>7.89</td>
<td>9.79</td>
<td>9.64</td>
<td>8.56</td>
<td>6.51</td>
</tr>
</tbody>
</table>

Table 3.1: Number of first nearest Na\(^+\) neighbors. \( g(r) \) was integrated to 4.3 Å for each salt concentration. There are more ions (about twice as many) residing near phosphates at the center of the junction than there are ions bound to other phosphates located closer to the junction center. As the magnesium concentration is increased, sodium ions are displaced from the center of the junction, and as the sodium concentration is increased, magnesium ions are displaced from the center of the junction.

### Table 3.2: First solvation shell occupancy for Mg\(^{2+}\) ions

<table>
<thead>
<tr>
<th>[Mg(^{2+})] (mM)</th>
<th>1</th>
<th>10</th>
<th>50</th>
<th>1</th>
<th>10</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.004</td>
<td>0.324</td>
<td>2.078</td>
<td>0.002</td>
<td>0.206</td>
<td>1.423</td>
</tr>
<tr>
<td>100</td>
<td>0.002</td>
<td>0.063</td>
<td>1.801</td>
<td>0.001</td>
<td>0.041</td>
<td>1.195</td>
</tr>
<tr>
<td>200</td>
<td>0.001</td>
<td>0.121</td>
<td>1.499</td>
<td>0.000</td>
<td>0.065</td>
<td>1.000</td>
</tr>
<tr>
<td>300</td>
<td>0.001</td>
<td>0.099</td>
<td>1.223</td>
<td>0.000</td>
<td>0.053</td>
<td>0.823</td>
</tr>
</tbody>
</table>

Table 3.2: First solvation shell occupancy for Mg\(^{2+}\) ions. \( g(r) \) was integrated to 5.3 Å for each salt concentration. As the sodium concentration is increased, magnesium ions are displaced from the first solvation shell at the center of the junction, and as magnesium concentrations increase, the first solvation shell fills up with magnesium ions. At [Na\(^+\)] 10mM and [Mg\(^{2+}\)] 50mM, the occupancy of the first solvation shell is around 2, which agrees with data from Litke et al. [20] that suggest two divalent cations at the center of the junction are necessary for specific folding of the junction.
3.2 Methods to determine junction isoform

To classify microscopic coordinates into macroscopic data concerning junction conformations with implications for DNA junction biology, biophysics, and nanotechnology, with our goal robust analyses with the ability to measure the isoform during junction melting, we tried four criteria to determine junction conformations in thermodynamic equilibrium.

1. Using distances between base pairs on opposing strands at the center of the junction.
2. Using angles between three successive phosphates in the backbone at the center of the junction.
3. Using the RMSD of configurations from an ideal isoform configuration.
4. Using data from the first two methods as input for Markov models to automatically classify conformations.

![Diagram showing different junction conformations]

**Figure 3.3:** In the Iso1 conformation (a), distances $d_{TT} \cup d_{CC}$ are small (<12 Å), $d_{AG} \cup d_{AG}$ are large (>12 Å), B (yellow) and R (green) strands are linear, and H (red) and X (blue) strands are stacked. In the Iso2 conformation (c), $d_{AG} \cup d_{AG}$ are small, $d_{TT} \cup d_{CC}$ are large, B and R strands are stacked, and H and X strands are linear. In open conformations, all distances are large and angles between adjacent arms are $\geq 90^\circ$. Adapted from Wang et al. [8]

### 3.2.1 Central base distances

Simulating DNA junctions with the implicit ion 3SPN.2 model, Wang et al. [8] used distances between bases at the junction center as a metric to determine isoform [see above for an outline of this process and Figure 3.4 for a histogram of central base distances and isoform population distributions calculated from this method]. Figure 3.4 shows the implicit ion 3SPN.2 model predicts the junction preferentially adopts the stacked conformations at high salt, with a 36/58 ratio between Iso1 and Iso2. Their calculations agree with results presented by McKinney et al. [10] who find a 33/77 ratio between stacked isoforms. Transition probabilities calculated from this model show that the open state acts as an intermediate between stacked...
isoforms. The authors note that this ratio underpredicts the bias toward Iso2 by about 15% when compared to experiments by Ha et al. [10]. They also note that these distances could be accessible experimentally using fluorescent nucleotide base analogs at the junction center.

Since these data were calculated from the implicit ion version of 3SPN.2 on an identical junction sequence, we perform the analysis on explicit ion model simulation trajectories to see if results are consistent across model versions. Figure 3.5 shows a normalized histogram of the central base distances from our simulations, plotted for each salt concentration. Qualitatively, we observe that at low salt concentration, the distributions reflect mostly open junctions, and at higher salt concentrations, the peaks reflect stacked junctions becoming more populated.
Section 3.2 - Methods to determine junction isoform

Figure 3.5: Normalized histogram of central base distances for all salt concentrations. We plot data from each Mg$^{2+}$ concentration, with different Na$^+$ concentrations shifted vertically by 0.1, and label the peaks of each graph as Iso1, Iso2, or Open according to rules prescribed in Figure 3.3. The dotted vertical line is the estimated cut-off distance of 12 Å.

Figure 3.6: Isoform population distributions from base distances criterion. The shade of red and the inset number correspond to the isoform population calculated using this method at each respective salt concentration.
3.2.2 Consecutive phosphate angles

Observations of higher than expected Iso1 probabilities at high [Na$^+$] led us to double check our results using alternative methods to measure isoform. Since junctions are described by which strands are linear versus stacked, we calculate angles between the three successive phosphates at the center of each strand as alternative criteria to determine the isoform [See Figure 3.7]. We label them by their strands (ie. H, B, X, R) and plot their normalized distributions in Figures 3.8. Bimodal angle distributions led us to use a cutoff angle method as previously used with the base distance criteria.

![Figure 3.7: Schematic of the angles between successive phosphates measured for this analysis. Phosphate, sugar, and base sites, and the linear bonds between them are visualized. Measuring this angle helps us to visualize junction configurations in a way that we cannot with only the distances between bases at the junction center.](image)

At lower salt concentrations, where stacked isoforms do not dominate, these histograms have a less clear cut-off angle between the two peaks. To validate this method’s ability to classify isoforms despite the peaks not being as clear-cut as the peaks in the distance histograms, and to check if the choice of cut-off angle alters isoform population distributions, we measure the isoform population distributions using 5 different angle cutoffs from 110° to 130° in steps of 5°. We observe no qualitative difference in isoform distributions for these different angle cut-offs.

These population distributions are similar to those calculated from the base distances criterion, leading us to believe the results returned by the base distances criteria. Even though the angles from simulation trajectory data do not have a direct analogy to measurements that can be easily made experimentally, we learn about the junction structure from these data. At high salt concentrations, the distributions reflect a stacked junction, and at low salt concentrations, the distributions reflect a more planar conformation. In our analysis of this data using hidden Markov models, we posit that a tetrahedral structure is a more likely alternative than the square planar conformation to be an intermediate between stacked conformations.
Section 3.2 - Methods to determine junction isoform

Figure 3.8: Angles from opposing strands are binned together. In Iso2, B and R arms are stacked and H and X arms are quasi-linear. In Iso1, B and R angles are quasi-linear (>90°) and H and X angles are small (<90°). Open and tetrahedral configurations exist as intermediates between the stacked states.

Figure 3.9: Isoform population distributions from central phosphate angles criterion with a cut-off angle of 130°. These distributions are qualitatively similar to those calculated from the base distances criterion.
3.2.3 RMSD between ideal isoform structures and current snapshot

As a third method to determine the junction configuration, we compare the structure of the junction at each trajectory snapshot to the idealized structure of each isoform. Root mean squared distance (RMSD) calculations quantify how much a molecule deviates from a given structure, and are numerically evaluated using

\[ RMSD(v, w) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} ||v_i - w_i||^2} \]  \hspace{1cm} (3.3)

where \( v_i \) is the location of the \( i \)-th atom relative to the molecule’s origin at time \( v \), \( w_i \) is the \( i \)-th atom’s location at time \( w \), and the sum is over all \( n \) simulated particles in the structure.

Since molecules rotate and translate during simulation, we transform the junction coordinates onto the idealized isoform structure to minimize the RMSD with respect to the ideal structures. Using the Python package ‘RMSD’ \[49\], which implements the Kabsch and quaternion algorithms \[50, 51\] to find the best transformation to relate two sets of vectors, we calculate the RMSD between the current snapshot and each ideal structure, classifying the isoform at that snapshot as that with the smallest RMSD value. Isoform distributions calculated from this classification method are below. This method struggles to produce results that match the literature; open populations are small at low \([Na^+]\) and \([Mg^{2+}]\), and stacked Iso2 populations dominate. Some results are consistent with other methods, however; there is an increase in the Iso1 state at \([Na^+] = 327\) mM, which is also visible in population distributions from the angle and base distances methods. We considered center-averaging the RMSD timeseries to remove noise. However, given the success of the previous two methods relative to this one, we did not continue this exploration.

![Figure 3.10: Isoform distribution by using RMSD from ideal isoform.](image-url)
3.2.4 Markov model state classification

In this section, we present the results obtained from training the Markov state models on our angles and distance timeseries. Tables 3.3, 3.4, and 3.5 show the means of the four Gaussian emission probability densities for the Markov models trained on the distance data from bases at the center of the junction, the phosphate angles at the center of the junction, and the combination of the base distances and the phosphate angles. To create these tables, we order states by the first value in the list of Gaussian emission means (AG for the base distances table, H for the angles and distances tables). For the base distances Gaussian emission probabilities, standard deviations were on the order of 1-10 Å, and for the angle Gaussian emission probabilities, standard deviations were on the order of 10-35°. Since each trained model returns a transition matrix, leading to 16 independent transition matrices, we do not plot them here [refer to Chapter 3.3 for our analyses of equilibrium dynamics from transition matrices]. Figures 3.11, 3.12, and 3.13 show isoform population distributions for the base distances, angles, and base distances and angles Markov models, calculated from the transition matrices by Equation 2.4. As a double check that Equation 2.4 works, we also empirically derive the stationary state by feeding the temporal data back into the trained model to classify the isoforms at each timestep, and find that the analytical solution and the empirical evaluation of isoform population distributions agree remarkably.

There are advantages and drawbacks of using these automated classification methods. First, the states labeled as Open are classified with angles larger than 90°, which points to a tetrahedral form of the junction being a probable transition state between stacked isoforms. This is an advantage of this method; we recover details of the equilibrium structure that we could not have with simple cutoff methods. A drawback is that some states were “miscategorized” at low salt. At low salt concentrations, when the stacked isoforms do not dominate, we still used the model to classify three states. At [Na⁺] = 27mM, 127mM and [Mg²⁺] = 0mM, 1mM, where we do not expect the Iso1 population to be present, the isoform population distribution predicts a large population [See Figures 3.11, 3.12, and 3.13]. Referencing the tables at the concentrations where we see the population is larger than expected, i.e. [Na⁺] = 27mM [Mg²⁺] = 0mM in Table 3.4, the mean angles (138, 111, 119, 106) and mean distances (26, 26, 25, 26) are indicative of a tetrahedral or square planar conformation rather than a stacked one. Thus we determined that these states were miscategorized; the Markov model determined these as separate states from the open configuration because we told the model to classify three states, and we mislabelled the states as Iso1 because we order and then label the states based on the first of these mean Gaussian emission values.
### Markov model mean base distances for trained classes

<table>
<thead>
<tr>
<th>[Na$^+$] mM, [Mg$^{2+}$] mM</th>
<th>“Iso2” AG AG TT CC</th>
<th>“Open” AG AG TT CC</th>
<th>“Iso1” AG AG TT CC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distance (Å)</td>
<td>Distance (Å)</td>
<td>Distance (Å)</td>
</tr>
<tr>
<td>27, 0</td>
<td>12, 13, 23, 23</td>
<td>19, 20, 20, 20</td>
<td>26, 26, 25, 26</td>
</tr>
<tr>
<td>127, 0</td>
<td>8, 9, 23, 23</td>
<td>18, 19, 22, 23</td>
<td>20, 21, 16, 16</td>
</tr>
<tr>
<td>227, 0</td>
<td>8, 8, 23, 23</td>
<td>17, 18, 21, 21</td>
<td>22, 22, 10, 10</td>
</tr>
<tr>
<td>327, 0</td>
<td>8, 8, 23, 23</td>
<td>17, 18, 20, 21</td>
<td>22, 22, 9, 9</td>
</tr>
<tr>
<td>27, 1</td>
<td>11, 12, 23, 23</td>
<td>19, 20, 19, 20</td>
<td>25, 25, 24, 25</td>
</tr>
<tr>
<td>127, 1</td>
<td>8, 9, 23, 23</td>
<td>18, 19, 21, 22</td>
<td>21, 21, 14, 14</td>
</tr>
<tr>
<td>227, 1</td>
<td>8, 8, 23, 23</td>
<td>17, 18, 20, 21</td>
<td>22, 22, 9, 10</td>
</tr>
<tr>
<td>327, 1</td>
<td>8, 8, 23, 23</td>
<td>16, 17, 21, 22</td>
<td>21, 22, 11, 12</td>
</tr>
<tr>
<td>27, 10</td>
<td>8, 8, 23, 23</td>
<td>18, 19, 19, 20</td>
<td>22, 22, 8, 8</td>
</tr>
<tr>
<td>127, 10</td>
<td>8, 8, 23, 23</td>
<td>17, 18, 18, 19</td>
<td>22, 21, 7, 8</td>
</tr>
<tr>
<td>227, 10</td>
<td>8, 8, 23, 23</td>
<td>16, 18, 20, 21</td>
<td>22, 22, 9, 9</td>
</tr>
<tr>
<td>327, 10</td>
<td>8, 8, 23, 24</td>
<td>15, 16, 21, 21</td>
<td>22, 22, 9, 10</td>
</tr>
<tr>
<td>27, 50</td>
<td>8, 8, 23, 23</td>
<td>17, 18, 19, 19</td>
<td>22, 22, 8, 8</td>
</tr>
<tr>
<td>127, 50</td>
<td>8, 8, 23, 23</td>
<td>16, 17, 20, 20</td>
<td>22, 22, 8, 8</td>
</tr>
<tr>
<td>227, 50</td>
<td>8, 8, 23, 24</td>
<td>17, 18, 18, 19</td>
<td>22, 22, 8, 8</td>
</tr>
<tr>
<td>327, 50</td>
<td>8, 8, 23, 23</td>
<td>17, 18, 18, 19</td>
<td>22, 22, 8, 8</td>
</tr>
</tbody>
</table>

**Table 3.3:** Means of Gaussian emissions from models trained on distances. States are ordered from left to right by the value of the first mean (the first AT distance).

![Isoform population distributions](image)

**Figure 3.11:** Isoform population distributions from base distances Markov models. Take care in reading these; in the table above, states are labeled from left to right as Iso2, Open, Iso1, whereas this plot is in the reverse order.
Section 3.2 - Methods to determine junction isoform

### Table 3.4: Mean of Gaussian emission probability distribution from Markov models trained on phosphate angles data. States are ordered from left to right by the value of the first mean (the H angle).

<table>
<thead>
<tr>
<th>[Na⁺]mM, [Mg²⁺]mM</th>
<th>“Iso2”</th>
<th>“Open”</th>
<th>“Iso1”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angle (°)</td>
<td>Angle (°)</td>
<td>Angle (°)</td>
</tr>
<tr>
<td>27, 0</td>
<td>76, 126, 77, 123</td>
<td>83, 115, 135, 112</td>
<td>138, 111, 119, 106</td>
</tr>
<tr>
<td>127, 0</td>
<td>68, 138, 68, 135</td>
<td>83, 122, 122, 121</td>
<td>139, 109, 125, 105</td>
</tr>
<tr>
<td>227, 0</td>
<td>68, 138, 69, 136</td>
<td>108, 125, 115, 127</td>
<td>137, 93, 137, 75</td>
</tr>
<tr>
<td>327, 0</td>
<td>67, 138, 68, 137</td>
<td>113, 125, 119, 117</td>
<td>139, 71, 139, 69</td>
</tr>
<tr>
<td>27, 1</td>
<td>83, 117, 131, 113</td>
<td>93, 132, 73, 126</td>
<td>138, 110, 127, 104</td>
</tr>
<tr>
<td>227, 1</td>
<td>68, 138, 69, 136</td>
<td>112, 123, 121, 120</td>
<td>138, 72, 139, 70</td>
</tr>
<tr>
<td>327, 1</td>
<td>67, 139, 68, 137</td>
<td>105, 128, 116, 128</td>
<td>136, 90, 138, 73</td>
</tr>
<tr>
<td>127, 10</td>
<td>67, 138, 68, 136</td>
<td>118, 122, 124, 115</td>
<td>139, 69, 139, 68</td>
</tr>
<tr>
<td>127, 10</td>
<td>69, 138, 70, 136</td>
<td>120, 122, 125, 117</td>
<td>138, 70, 138, 69</td>
</tr>
<tr>
<td>227, 10</td>
<td>67, 138, 68, 136</td>
<td>113, 125, 122, 118</td>
<td>139, 72, 139, 70</td>
</tr>
<tr>
<td>327, 10</td>
<td>67, 139, 68, 137</td>
<td>107, 132, 118, 123</td>
<td>139, 68, 139, 67</td>
</tr>
<tr>
<td>127, 50</td>
<td>67, 138, 67, 137</td>
<td>120, 124, 124, 116</td>
<td>139, 68, 139, 67</td>
</tr>
<tr>
<td>227, 50</td>
<td>67, 138, 67, 137</td>
<td>120, 122, 124, 118</td>
<td>139, 67, 139, 66</td>
</tr>
<tr>
<td>327, 50</td>
<td>67, 139, 67, 137</td>
<td>120, 123, 128, 113</td>
<td>139, 68, 139, 66</td>
</tr>
</tbody>
</table>

*Figure 3.12: Phosphate angles Markov model isoform population distribution. When we take into account the misclassification of states, the model trained on phosphate angles data does capture expected equilibrium structures; open state population decreases with increasing salt concentration.*
Section 3.2 - Methods to determine junction isoform 37

Markov model mean angles and distances for trained classes

<table>
<thead>
<tr>
<th>Na,Mg</th>
<th>“Iso2”</th>
<th>“Open”</th>
<th>“Iso1”</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.0</td>
<td>68,137,69,134,10,10,23,23</td>
<td>80,117,117,114,18,19,21,21</td>
<td>133,111,121,106,23,23,22,22</td>
</tr>
<tr>
<td>127.0</td>
<td>67,138,68,135,8,9,23,23</td>
<td>96,127,111,127,16,17,22,22</td>
<td>133,105,132,99,22,22,19,19</td>
</tr>
<tr>
<td>227.0</td>
<td>68,138,69,136,8,8,23,23</td>
<td>110,125,116,122,17,18,21,21</td>
<td>139,77,139,70,22,22,10,10</td>
</tr>
<tr>
<td>327.0</td>
<td>67,139,68,137,8,8,23,23</td>
<td>112,124,118,117,17,18,20,21</td>
<td>139,69,139,68,22,22,9,9</td>
</tr>
<tr>
<td>127.1</td>
<td>69,137,69,134,9,10,23,23</td>
<td>81,119,119,115,18,19,21,21</td>
<td>135,113,121,106,22,22,21,21</td>
</tr>
<tr>
<td>127.1</td>
<td>66,138,67,136,8,9,23,23</td>
<td>110,123,119,119,18,19,21,22</td>
<td>136,90,136,85,21,21,14,15</td>
</tr>
<tr>
<td>227.1</td>
<td>68,138,69,136,8,8,23,23</td>
<td>112,122,120,118,17,18,20,20</td>
<td>138,70,139,68,22,22,9,9</td>
</tr>
<tr>
<td>327.1</td>
<td>67,139,68,137,8,8,23,23</td>
<td>103,130,113,126,15,16,21,22</td>
<td>136,87,138,79,21,21,12,12</td>
</tr>
<tr>
<td>127.10</td>
<td>67,138,68,136,8,8,23,23</td>
<td>117,120,124,113,18,18,19,20</td>
<td>139,68,139,67,22,22,8,9</td>
</tr>
<tr>
<td>127.10</td>
<td>69,138,70,136,8,8,23,23</td>
<td>120,121,124,116,17,17,18,19</td>
<td>138,69,139,68,22,21,8,8</td>
</tr>
<tr>
<td>227.10</td>
<td>67,138,68,136,8,8,23,23</td>
<td>113,124,121,117,17,18,20,20</td>
<td>139,71,139,69,22,22,8,9</td>
</tr>
<tr>
<td>327.10</td>
<td>67,139,67,136,8,8,23,24</td>
<td>105,132,116,123,16,16,21,21</td>
<td>139,68,139,67,22,22,8,9</td>
</tr>
<tr>
<td>127.50</td>
<td>67,138,67,137,8,8,23,23</td>
<td>119,122,123,114,17,18,19,19</td>
<td>139,68,139,66,22,22,8,8</td>
</tr>
<tr>
<td>127.50</td>
<td>67,138,67,137,8,8,23,24</td>
<td>112,126,122,119,16,17,20,20</td>
<td>138,68,139,66,22,22,8,8</td>
</tr>
<tr>
<td>227.50</td>
<td>67,138,67,137,8,8,23,24</td>
<td>120,119,124,115,17,18,19,19</td>
<td>139,67,139,66,22,22,8,8</td>
</tr>
<tr>
<td>327.50</td>
<td>67,139,67,137,8,8,23,23</td>
<td>120,121,127,111,17,18,18,19</td>
<td>139,67,139,66,22,22,8,8</td>
</tr>
</tbody>
</table>

Table 3.5: Mean of Gaussian emission probability distribution from Markov models trained on distances and angles. States are ordered from left to right by the value of the first mean (the H angle).

Figure 3.13: Base distances and angles Markov model isoform population distribution. This model returns the expected result that stacked states are more probable at higher salt concentrations.
3.3 Transitions between conformations

In order to confirm that, like in experiments, the open configuration acts as an intermediate state between the two stacking conformers in our model, we calculate transition rates between conformations for states identified using the phosphate angles isoform criterion and show the transition matrices for high and low salt concentrations. [See Figure 3.14.] From these we determine that the addition of ions serves to stabilize the stacked isoforms and the junction must transition through the open configuration to exchange stacking strands. Our results for transition probabilities between stacked and open states are consistent with results from the implicit ion 3SPN.2 study by Wang et al. [8], who report an ≈0.1 transition rate between Iso2 and Open and a ≲ 0.1 transition rate for the Iso1-Open transition at 300mM [Na⁺]. Moreover, the transitions rates directly between Iso1 and Iso2 are essentially zero, demonstrating that the open state provides the only pathway to move between stacked isoforms. This result is also consistent for transition matrices derived from isoform timeseries where isoforms were classified using the base distances criterion and the Markov models. Those figures are not shown here for brevity.

Figure 3.14: Transition matrices from phosphate angles criteria for low and high salt concentrations. Transition rates between the two stacked isoforms (the upper left and lower right values) are vanishingly small, whereas the off diagonals representing transition rates between stacked and open isoforms, are nonzero. As Mg²⁺ concentrations increase, stacked states become more stable.
Chapter 4

Melting Dynamics

"The foundation stones of the material universe remain unbroken and unworn. They continue this day as they were created - perfect in number and measure and weight."

Clerk Maxwell, Treatise on Electricity and Magnetism, 1873

Figure 4.1: Five frames from the first 200ns of a single melting simulation, depicting the stages of the melting process. Melting initiates with a bubble of broken bases forming at the center of the junction (1), which then grows outward (2,3), resulting in two arms dissociating from the junction (4). Melting proceeds by the final two intact strands dissociating from each other (5). The phosphates are visualized as one continuous strand. Strand colors are red (H), blue (X), yellow (B) and green (R). This chapter presents data which provide quantitative and qualitative analyses of the junction melting process. Frames were rendered in Visual Molecular Dynamics (VMD) [52].

In this chapter we communicate our findings on the melting dynamics of the J3 junction as observed in our simulations and experimental measurements carried out by Ishita Mukerji et al. First, we discuss the method by which we determine the dissociation of individual base pairs [See Figure 4.2]. We use the
relative “time-to-melt” $t_m$ as a proxy for base stability in our simulations; if a base pair takes a longer
time to dissociate than other base pairs, that base pair is relatively more stable. As an example of the
dynamic process of melting and an introduction into our analysis of the relative stability of whole strands,
we plot the times at which individual arms of the junction have dissociated on top of the number of intact
base pairs versus time for one simulation [See Figure 4.3]. To investigate the relative stability of strands in
the junction, we plot the mean time-to-melt for all base pairs along two opposing strands (H and X). Next,
we compare our measurements on the stability of these strands with measurements of melting temperature
obtained from fluorescent nucleoside analogs placed at specific locations in the junction and show that
our results are consistent with those from the experiment. As a last check to determine our analysis is
consistent and to provide another structural viewpoint of junction melting, we measure the isoforms of
our melting ensemble and show that junctions adopt a stacked conformation prior to melting with isoform
population distributions comparable to those in our equilibrium ensemble at similar salt concentrations.
These data report that the 3SPN.2 explicit ion model can replicate experimental data on DNA junction
melting and provide insights into DNA junction melting dynamics.

4.1 Method to determine base dissociation

To determine melting times of individual base pairs, with the goal of outlining regions of instability within
the junction, we measure the distance between the canonical base pairs on each strand. Using an analysis
method developed by Wujie Wang, we denote a canonical base pair as melted if the distance between
bases is larger than 50 Å for more than 50 simulation snapshots (6 nanoseconds). The 50 Å distance and
6 nanosecond time are large enough values to assure that base pairs are not classified as melted during
thermal fluctuations. Figure 4.2 shows the distance between three canonical base pairs for the first 500ns
of one simulation in our melting ensemble at 360K; see the bottom panel for an example of thermal
distance fluctuations which are not falsely classified as full base pair dissociation.

Figure 4.3 shows the number of intact base pairs and dissociation time of each of the arms for one
specific simulation at 360K. In this run, the melting process occurs relatively slowly and on a per-strand
basis, whereas in other simulations at higher temperatures, melting occurs rapidly and the dynamics are
not as easily visible. Melting begins in this simulation with the dissociation of almost half the base pairs
in the junction between 0 and $\approx 50$ns, followed by dissociation of the XR arm. From 50ns to $\approx 300$ns, the
number of base pairs remains approximately constant. At 300ns, the BH strand dissociates. In the next
50ns, the number of intact base pairs drops from $\approx 20$ to $\approx 15$, after which the BX strand dissociates. The
Figure 4.2: Left: timeseries of distances from 3 base pairs, with the determined melting time in orange. Right: the locations of the base pairs in J3 which are plotted on the left. At the start of the simulation, the distances are small, corresponding to intact base pairs with fluctuations due to thermal noise. After some time the distance gets large and stays large, corresponding to a dissociated base pair. The analyzed base locations correspond to locations of fluorescent nucleoside analogs used for fluorescence melting experiments.

Figure 4.3: Number of base pairs and $t_m$ by arm for 1 simulation. The solid vertical lines correspond to the time at which the last base pair in each arm has dissociated, and the dotted vertical lines depict the time at which 50% of the base pairs in that arm have dissociated. Initially, the number of base pairs is 64, representing a fully intact junction. After around 1000ns, the number of base pairs has dropped to zero, with all strands dissociated. We see that the melting of individual arms occurs after a sharp decrease in the number of bonded base pairs (the noisy line going from green to red), which matches our physical intuition.
Section 4.2 - Preferential melting

next 400ns period ends with dissociation of the RH strand and the number of intact base pairs drops to zero. This graph and others like it offer unique insights into the dynamics of individual junctions melting which are not as transparent as insights offered by other experimental methods.

4.2 Preferential melting

To investigate the dynamics of melting and the stability of the strands and arms of the junction, plot the mean base pair dissociation time for junction simulations at 360K, which we use as a metric for junction base pair stability [See Figure 4.4]. Our expectations are that because the X strand has two fewer hydrogen bonds (88) than the H strand (90) when all canonical Watson-Crick base pairs are intact, and also from theoretical calculations of strand melting temperatures [See Figure 4.6], the X strand will be less stable than the H strand, and calculations from our models show this to be the case. We also calculated the average distance between base pairs before that base pair has melted at the four different temperatures in the higher range of temperatures we simulated, and see that the ends and center of each arm have higher base distance fluctuations prior to melting than the middle of each arm. These results on the instability of the junction center and the ends of the arms is consistent with experimental data from Wujie Wang, Rachel Savage, and Julie McDonald [44, 45], whose data are not shown here but show that melting either occurs from the outside of the strands inwards, or from the center of the junction outwards. Our images of mean base pair melting times show that the center of the junctions dissociate, on average, before the ends of the arms, and also that the X strand is more stable (i.e. has a lower melting temperature) than the H strand. Because not all simulations at 360K fully melted, and our base pair melting time calculation returns the length of the simulation if the base pair is not dissociated by the end of the simulation, the standard deviation of the average base pair melting time is on the order of the scale of the y-axis and is not shown.

Experiments that were motivated by our results on preferential melting of individual strands used fluorescent nucleotide analogues in place of bases on individual strands to obtain base pair resolution mapping of junction stability. These experiments proceeded as follows: equal concentrations of three DNA strands corresponding to the three of the four strands in the junction are placed in a buffer solution with the final DNA strand, which has a fluorescent probe in place of one nucleotide. Annealing of this solution creates fully intact DNA junctions, which are then heated while fluorescence is measured as a function of temperature. The fluorescence of the probe is modulated by its distance to the base pair on the antisense strand. Initially, the junction is intact, the probe forms a base pair with the other strand,
Figure 4.4: Mean base pair dissociation time for all melting simulations at 360K, with base pairs separated by their corresponding strands, or pseudoduplexes. We see the X strand dissociates on average around 200 ns before the H strand. This points to the X strand being a less stable strand than the H strand.

and minimal fluorescence is observed. As the junction melts, the probe dissociates from its base pair and begins to fluoresce.

By tracking fluorescence as a function of temperature we can obtain a melting temperature, and by comparing melting temperatures across probes placed on different strands, we elucidate the relative stability of those strands. Figure 4.5 shows data from fluorescence melt experiments on DNA junctions of the same sequence as in our simulations. The different curves correspond to fluorescence melts taken for each of four fluorescent probes placed along two strands in the junction. Red curves represent the fluorescence melt curves from probes placed on the X pseudo-duplex, and black curves represent those on the H pseudo-duplex. The temperature required for the normalized fluorescence to reach 50% corresponds to the melting temperature of that base pair. We see that the melting temperatures of the probes along the H strand are higher than the melting temperature of the probes along the X strand, which is consistent with the base pair dissociation time from our simulations [see Figure 4.4]. These data represent the first demonstration of junction melting on a preferential arm-by-arm basis and demonstrate that an alternate model of junction melting which requires the collective melting of the entire junction is not valid.

As another comparison on melting dynamics to compare with previous results, we look at theoretical melting temperatures of the DNA sequences that comprise the strands of interest in this DNA junction.
These theoretical melting temperatures are reported by two different methods; the first is from an online tool published by Integrated DNA Technologies (IDT) which calculates the melting temperatures of duplex strands based on sequence, sodium and magnesium concentration, using experimental DNA enthalpy and entropy values defined by Allawi and Santalucia \[53\]. The second method also uses enthalpy and entropy values from SantaLucia’s experiments input into a Mathematica script. These two calculations differ in their prediction of absolute melting temperatures due to differences in the implementation of salt concentration dependence, but the relative melting temperatures between strands are the most important to consider and are consistent across methods. These theoretical models predict that, if the H strand and the X strands existed in the absence of a DNA junction, that the X strand would be less stable than the H strand. This agrees with data from our simulations and also from experimental fluorescence measurements of DNA junctions by Savage and McDonald \[44, 45\].

![Figure 4.5](image)

**Figure 4.5:** Normalized fluorescence intensity versus temperature for four different probe locations. These experimental data show that the X pseudo duplex is more unstable than the H pseudoduplex, which agrees with the results from our measurements of the mean base pair melting time for bases along these strands.

### 4.3 Isoform transitions during melting process

By probing the structure of DNA junctions during the melting process, we shed light on how melting occurs and offer another way in which to validate our predictions of prefential melting from the mean base pair dissociation times. For each melting simulation we calculate the isoform as predicted by our base-distances cut-off method. Figure 4.7 shows an example isoform timeseries as calculated by the
Table 4.1. Theoretical melting temperatures of junction pseudo-duplexes

<table>
<thead>
<tr>
<th></th>
<th>Integrated DNA Technologies</th>
<th>Starr/SantaLucia Mathematica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tm (°C)</td>
<td>Relative Stability</td>
<td>Tm (°C)</td>
</tr>
<tr>
<td>X34</td>
<td>78.53</td>
<td>2</td>
</tr>
<tr>
<td>H34</td>
<td>83.88</td>
<td>1</td>
</tr>
<tr>
<td>X17</td>
<td>59.29</td>
<td>4</td>
</tr>
<tr>
<td>H17</td>
<td>69.35</td>
<td>1</td>
</tr>
<tr>
<td>B17</td>
<td>62.59</td>
<td>3</td>
</tr>
<tr>
<td>R17</td>
<td>62.97</td>
<td>2</td>
</tr>
</tbody>
</table>

As expected, the relative stability of the duplexes does not differ between the two modes of calculations while the absolute melting temperature does. Also important to note is that the Integrated DNA Technologies melting temperature, which takes into account the DNA and NaCl concentration, is only 1 °C higher than both experimental absorbance and fluorescence melts of the X34 duplex. The average absorption melting temperature for D3X is 76.6 ± 0.1 °C and the average fluorescence melting temperatures for 6MI-labeled D3X is 77.0 ± 0.8 °C, while Integrated DNA Technologies predicted a melting temperature of 78.5 °C. Such accuracy between experimental and theoretical melting temperatures is a promising result that shows that the experimental techniques of quantifying the melting temperature of DNA were precise.

Although the melting temperatures of the duplexes were consistent with the model, these melting temperatures were still an average of 12 °C higher than the

Figure 4.6: Theoretical melting temperatures of the DNA sequences that comprise the junction. The two columns correspond to different ways in which these theoretical melting temperatures were calculated. The blue rows correspond to theoretical melting temperatures of the 34 base pair DNA sequences that comprise the X and H strands. The last four rows are theoretical melting temperatures for the 17 base pair DNA sequences that make up the junction arms. Adapted from Savage, 2017 [44].

distances method) for the first 100ns of a single simulation. In Figure 4.8 we have plotted the isoform population distributions as a function of time before and after melting. We see that the junction population distribution during melting mirrors the isoform population during equilibrium. This data is compatible with our analysis of strand stability and offers an insight into the dynamics of junction melting. As we consider both the preferential melting of individual strands, and the most probable stacked isoforms during melting, we paint a picture of the dynamics of strand dissociation and offer an explanation for junction melting not previously explored in literature.
Figure 4.7: Isoform timeseries calculated from base-distances criterion for the first 100ns of run 8 at 360K. During equilibration, when the temperature is 310K, the junction does not transition between isoforms quickly. After equilibration, the temperature increases and faster dynamics leads to frequent isoform transitions. After 40ns, the distances between bases at the junction center lead to an open classification. This corresponds to dissociation of opposing strands. Junctions at 360K do not usually melt this fast, nor does this timeseries paint the full picture of melting dynamics from the context of junction isomerization. This plot is simply shown as an example of an isoform timeseries for the melting system.

Figure 4.8: Isoform population distribution for all simulations at 370K. We made this plot by creating an isoform timeseries (as in Figure 4.7) for each melting simulation. Then, for each timeseries, we set as “zero” the time at which all strands have completely dissociated during that simulation. Finally, we “stack” the isoform timeseries on top of each other centered at their new zeroes and tally the number of simulations that are in each state 250ns before and after melting occurs.
Branch Migration Dynamics

“Just because we don’t understand doesn’t mean that the explanation doesn’t exist.”


Branch migration occurs when the junction is in the open state. Cations stack the junction arms by shielding the negatively charged phosphates in the backbone of DNA. Thus branch migration is reduced in the presence of high salt concentrations and occurs on much slower timescales than the exchange of stacking arms. The correlation between junction isomerization and branch migration should be observable in our simulations, given the data presented in this chapter and in Chapter 3.2, but we do not take junction isoform measurements in our branch migration simulations, so we cannot carry out this analysis. Nonetheless, we show that the explicit ion model of 3SPN.2 can be used for studies of DNA junction migration in the presence of Na\(^+\) and Mg\(^{2+}\) ions. We calculate timeseries’ of the location of crossover between the strands, and from those calculate a “transition probability” for each salt concentration. This reproduces the expected result that the addition of cations serves to decrease junction migration. Since the version of 3SPN.2 we use includes explicit ions and can recreate junction migration dynamics, it might be useful in future studies of the affects of ion binding and unbinding at the heart of the junction on junction isomerization and branch migration. In principle, our simulations have the capability to generate statistics from such measurements. A comparison of data on branch migration dynamics as collected from the 3SPN.2 explicit ion model to data collected by Xinyu Zhu on the implicit ion model remains to be done.
5.1 Methods to determine junction location

Code developed by Xinyu Zhu to measure the junction location in simulations of junctions from the implicit ion model, and adapted for our specific system, determines the junction location for our migrating junction simulations. For each simulation snapshot, the code walks from the end of each arm towards the center of the junction, measuring the distance between canonical base pairs. The junction location for that arm is recorded as the base pair at which the distance becomes larger than a certain cut-off distance. The “true” junction location for that snapshot is calculated as the mean of the junction location from each of the four arms. The developers of this algorithm tested different cut-off distances to determine the optimal cut-off distance at which to determine the junction location; because of time constraints, we did not perform this same analysis.

5.2 Junction migration probability

In Figures 5.1 and 5.2, we plot the junction location as a function of simulated time for all simulations of the migrating junction at low and high magnesium concentrations. We do not filter this data, so some noise in the signal may falsely represent branch migration. We see that at low magnesium concentration, the junction location fluctuates on shorter timescales due to the junction being in the open configuration more, where junction migration is possible [see Figure 1.3]. At higher magnesium concentration, the junction adopts a stacked conformation and is less likely to migrate. We see similar trends in the FRET timeseries of migrating junctions from Karymov et al. in Figure 1.11.

In Figure 5.3 we show the total number of our simulation snapshots in which the junction location is different from the previous simulation snapshot, divided by the total number of snapshots. This is a crude approximation for the probability of the junction undergoing a branch migration step, especially since our junction location data was not filtered to remove any noise [See 1.11 for how filtering branch migration signals can help visualize data and improve analysis], but it offers a qualitative picture of how different ionic concentrations affect the dynamics of DNA junction branch migration as simulated in the 3SPN.2 explicit ion model. We average these migration probabilities over the 5 simulations of the migrating junctions at each concentration to get the probability that a junction will migrate on the timescale between snapshots. As expected from experimental FRET data on migrating junctions from Karymov et al. [21], increasing salt concentrations decreases junction migration probability.
Figure 5.1: Timeseries of junction location in simulations with no Mg$^{2+}$. At this concentration, junctions are in the open state often, where branch migration possible. Here the branch point of the junction can shift by more than 5 base pairs during the course of the 1.5µs simulation.

Figure 5.2: Timeseries of junction location at high magnesium concentration. At high salt concentrations, the junction is more likely to be in one of the stacked isoforms, where branch migration is not possible. Here, the branch point shifts less frequently and makes smaller jumps (around 2 base pairs) than during simulations in junctions with no magnesium. Since these data were not filtered to reduce noise, some artifacts exist in the timeseries and are most likely not “real” branch migration steps.
Figure 5.3: Migration probability for mobile junction at various salt concentrations. As salt concentration is increased, the probability of junction migration decreases. This result is more drastic for low Na\(^+\) concentrations ([Na\(^+\)] = 127 mM) as [Mg\(^{2+}\)] increases than it is for higher Na\(^+\) concentrations. This shows that Mg\(^{2+}\) ions are more effective at pinning the junction location when Na\(^+\) concentrations are low. More detailed descriptions of how Mg\(^{2+}\) ions interact with the junction center and are available in a publication by Hyeon et al. [25], but an in-depth analysis of the dynamics of junction isomerization, branch migration, and ion binding remains for future studies of this system.
Conclusion

“...there are two moments that are important. There’s the moment when you know you can find out the answer and that’s the period you are sleepless before you know what it is. When you’ve got it and know what it is, then you can rest easy.”

Dorothy Hodgkin, For our daughters: how outstanding women worldwide have balanced home and career, 1996

We used the 3SPN.2 implicit ion model to examine DNA junctions in equilibrium, during melting, and undergoing the branch migration process. We confirmed our ability to reproduce, to a first approximation, the structural and dynamic aspects of DNA junctions through simulations using this model. Here we discuss the strengths and weaknesses of our methods and evaluate the degrees of success with which the implicit ion model of 3SPN.2 captures the structural and dynamic properties of DNA junctions in the context of our analyses.

6.1 Evaluation of ionic distributions around the junction

We compare with theoretical calculations by Litke et al. [4, 20], who predict two magnesium binding sites at the center of the junction, and find that our results reflect their experimental predictions for Mg$^{2+}$ nearest neighbors at the junction center. We also compare with theoretical calculations of radial distribution functions of ions around the junction from Hyeon et al. [25] and find that our calculations match, showing
preferential ion binding at the junction center. These calculations provide motivation to use the explicit ion 3SPN.2 model to further investigate ion binding at the junction center and its effects on dynamical junction processes such as isomerization, melting, and branch migration.

### 6.2 Evaluation of junction isoform determination methods

Central to our discussion is our model’s bias toward stacked conformations at lower salt concentrations. Our analyses show that at low salt concentrations ([Na$^+$] = 27mM and 127mM and [Mg$^{2+}$] = 0mM and 1mM), the open state dominates and stacked states have probabilities close to zero. As salt concentration increases, the fraction of junctions in the open state decreases while the fraction of junctions in the Iso2 state increases, but this transition of stacking partners is sensitive to which ion concentration changes. For example, we see at [Na$^+$] = 27mM along the increasing [Mg$^{2+}$] axis, the Iso2 probability increases whereas the Iso1 probability does not increase. In comparison, looking at [Mg$^{2+}$] 50mM along the increasing [Na$^+$] axis, the Iso1 probability increases and the Iso2 probability decreases. These results are contrary to the expected result that salt concentration does not affect the junction bias for which arms stack, which could be a statistical error attributed to the number and length of simulations or a systematic error attributed to the model’s lack of ability to capture dynamic properties of DNA at high ionic strength. The developers of the 3SPN.2 model note that “it is unclear how appropriate the model is for calculations of dynamic properties at moderate to high ionic strengths.” [41]

The dissonance between our observations and expectations may be a consequence of several shortcomings in our study. Simulations may not be extensive enough to calculate a representative ensemble average. (i.e a single trajectory may not sample the entire junction configurational space). Short equilibration times and simulations beginning from the open state may have lead to an overclassification of the open configuration. Our model also shows a larger than previously measured ratio of Iso1 and Iso2 populations, which could be due to similar reasons as for the oversampling of the Open state. Since simulation time is finite, those simulations that transition from Open to Iso1 might stay in Iso1 for the duration of the simulation, especially at high salt concentrations. In the implicit ion model study by Wang et al. [8], 10 times as many simulations were run for twice as long at each salt concentration to gather equilibrium statistics on isoform population distributions. Then, distributions were double-checked by running extra simulations starting from different initial configurations to see that had any effect on the isoform population distributions. In future work using this model, efforts might be better concentrated in simulating longer runs for fewer salt concentrations to generate more in-depth statistics on a few relevant
salt concentrations.

Second, the 3SPN.2 model might simply be more prone to predicting open configurations of DNA junctions. The contributions of ions might dampen the dynamics in the system make the junction less likely to adopt stacked conformations or transition between stacked conformations on our simulated timescales at high salt; we might see an exchange of stacking partners at higher salt concentrations because of a crowding effect of ions around the junction center and arms. The 3SPN.2 explicit ion model might fail to return the expected result at such high ion concentrations since it is only meant to handle low ion concentrations. Hence we also recommend only simulating lower salt concentrations for future research using this model.

We are not entirely convinced of the validity and utility of our Markov model classification method. Here we discuss this method’s results and significance. We see that the trained Markov models return drastically different classifications for different salt concentrations; the means and standard deviations of the Gaussian emissions differ between salt concentrations. This causes confusion when analyzing the Markov model’s population distributions. As a potential improvement upon this analysis, we hypothesize that providing more data by inputting every single trajectory snapshot from all salt concentrations into the Markov model training script might remove the problem in which the Markov model miscategorizes an open state that we labeled Iso1. Although this method would not return transition matrices and isoform population distributions for each concentration (we would have to measure these populations for each salt concentration empirically, after training), we postulate that more data over a range of salt concentrations will lead to better fitting classifications of isoform geometries. Since these models take hours to train, as compared to minutes when simply tallying states using a cut-off distance or angle criterion, we are unsure if they are as useful as our other analyses using cutoffs. However, the Markov models do return important details, such as showing that a tetrahedral form of the junction is more likely than a square planar form. Despite the Markov model’s inherent usefulness in describing conformational states, questions of the replicability of our data, how long the analyses take compared to simpler methods, and time constraints lead us to recommend critically evaluating the benefit of using these types of models for analysis in future studies.

### 6.3 Evaluation of melting data

We predict using the base pairs at the center of the junction would not be a good method to determine isoform during melting especially since we see that the junction preferentially melts from the center;
regardless, we were able to give a dynamical picture of melting by measuring the isoform populations distributions for junctions before and after all strands were completely dissociated. For our analysis of melting simulations, data could have been biased by the fact that not all junctions melted during the simulation time, even though simulations were carried out above the melting temperature. Because the base dissociation time code returns the length of the simulation if the base pair is not melted by the end, our data on the average base distance prior to melting and the mean base dissociation time for each strand [See Figure 4.4] may be biased to lower values. In the future we might consider launching more simulations for longer timescales to gather more statistics, although it is unclear if that would remove or lessen this bias caused by unmelted junctions.

### 6.4 Future directions

This project was pushed in many directions since its inception, and we are still left with questions to answer and data to analyze. We propose the following recommendations for further research and analyses that could be carried out using our current simulation trajectories and thermodynamic data, as well as recommendations for carrying our further simulations using the 3SPN.2 explicit ion model:

1. Compare data on branch migration collected by Xinyu Zhu from the implicit ion model with data from our simulations, specifically, histograms of preferred locations in the branch migration process, probable branch migration base step sizes, and the probability of the different junction isomers when the junction branch point is in different locations.

2. Simulate longer trajectories at lower salt concentrations to gather higher quality statistics on the isoform distributions.

3. Simulate longer trajectories of melting to remove some bias regarding junctions not melting during the extent of the simulation.

4. Use isoform characterization algorithms to determine the isoform of the mobile junction before, after and during junction migration steps, to measure how the Open state is in fact an intermediate for both junction migration and crossover isomerization.

5. Analyze thermodynamic data to investigate free energies of junction conformer transitions, branch migration transitions, ion binding, and the dynamics of these processes.
In conclusion, the explicit ion 3SPN.2 model provides a robust framework on which to study the dynamics of DNA junction isomerization, melting and branch migration. Our results offer a portrayal of ion binding consistent with the literature, multiple ways in which to classify junction conformations and infer isoform population distributions at various salt concentrations, a dynamic picture of junction melting which indicates that junction melting occurs on a per-strand basis, and branch migration statistics consistent with the familiar trends. In addition, we provide suggestions for future computational research of DNA junctions.
Bibliography


