UNIVERSITY OF CALIFORNIA,
IRVINE

Computational Studies of Hoogsteen Base Pairs in Nucleic Acids and Developments in Enhanced Sampling Simulation Techniques

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

James McSally

Dissertation Committee:
Professor Ioan Andricioaei, Chair
Professor Douglas Tobias
Professor Craig Martens

2019

DEDICATION

To my family, my friends, and my love, Dian.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I most certainly must begin by thanking my wonderful adviser, Ioan Andricioaei. Your guidance has furthered my passion for science. You have set an aspiring example of what it means to excel and truly enjoy one's career. You made a lasting impression on me on my very first visit to UCI and have continued to impress me for the duration of my time here. You have always treated me with respect as both a person and as a burgeoning scientist. You gave me room to explore on my own and used a guiding hand when I would occasionally lose my way. I can say with confidence that I am prepared for the road ahead, because of the tools you have given me, and for that I cannot say thank you enough.

Thank you to the current and past members of the Andricioaei group. Specifically, I would like to thank Gianmarc Grazioli, Anupam Chatterjee, Moises Romero, and Dhiman Ray. Through various conversations and sometimes outright collaborations, you have all helped me reach this goal.

Thank you the Al-Hashimi group at Duke University, for a fruitful collaboration over my years here.

To the members of my advancement committee: Doug Tobias, Craig Martens, Vladimir Mandelshtam, Filipp Furche, and Ray Luo. Thank you for helping me see the merits of looking at a problem from as many sides as possible. I appreciate your time and consideration, and thank you for your trust in me to succeed here at UCI since my candidacy exam.

I want to thank UCI and the UCI Chemistry department, for giving me this amazing opportunity to grow. I have had a fantastic experience in the excellent collegiate atmosphere so strongly cultivated here.

I would like to acknowledge the various computing resource organizations that have made my work possible. Specifically, STAMPEDE at XSEDE, Triton Shared Computing Cluster (TSCC) is UC San Diegos primary research HPC system, and UCI's very own Greenplanet.

I would also like to acknowledge the NIH for their financial support through grant number: 5R01GM089846-09. As well as the NSF for their support through grant number: CMMI 1404818

I must also thank Thomas Douglas at St. John Fisher College, Jannette Carey at Princeton University, and David Reha at the Center for Nanobiology and Structural Biology, Czech Republic, for taking a chance on me with my various undergraduate research projects.

Additionally I would like to thank my High school and Junior High chemistry teachers: Mrs. Malecki and Mr. Donahue. Thank you for entertaining a million questions from me and encouraging me to ask even more.

Thank you to all of my friends. Both the ones I've made here at UCI and those in NY, many

of you will never know the true magnitude of how you've helped me.

I want to thank all of my family, both of my parents: Jim and Barb, my grandparents: Dick and Arlene, and my sisters: Mollie and Victoria. Thank you for all of your love and support. It wasn't easy to travel across the country to pursue this incredible opportunity, but all of you helped make it possible. I would never have been able to do any of this without each and everyone one of you, and for that I will forever be grateful.

Finally my bride to be, Dian Romonosky. You have meant the world to me and have been a continuing source of inspiration through this process. Thank you so much for your understanding and patience, that made this all possible. You were there for all of the highs and lows in this journey. If I have learned one thing, it is that with you by my side, there is nothing that we can't accomplish. We made it!

CURRICULUM VITAE

James McSally

EDUCATION

Doctor of Philosophy in Chemistry **2019**
University of Califorina, Irvine *Irvine, California*

Bachelor of Science in Chemistry **2014**
St. John Fisher College *Rochester, New York*

RESEARCH EXPERIENCE

Graduate Research Assistant **2014–2019**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2014–2019**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

m1A and m1G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs **2016**
Nature Structural and Molecular Biology

Modulation of Hoogsteen dynamics on DNA recognition **2018**
Nature Communications

ABSTRACT OF THE DISSERTATION

Computational Studies of Hoogsteen Base Pairs in Nucleic Acids and Developments in Enhanced Sampling Simulation Techniques

By

James McSally

Doctor of Philosophy in Chemistry

University of California, Irvine, 2019

Professor Ioan Andricioaei, Chair

The further study of the fundamental physical properties of nucleic acids, can have far reaching impacts. These highly dynamic biological macromolecules can be difficult to study as their behavior can be dictated by processes that act on timescales that range over several orders of magnitude. Certain experiments of nucleic acids in solution can scratch the surface of the sub millisecond regime, but experiments that probe processes much faster than that can become difficult, and often do not contain significant atomistic detail, with the notable exception of site-specific liquid-state NMR. From the nanosecond to microsecond, biological macromolecules can be readily studied by simulation. Through the use of molecular dynamics I've studied an exciting feature of nucleic acids, the Hoogsteen base pair. In recent years support for its biological relevance has increased. Understanding the mechanics of how this base pair forms and it's energetic comparison to the Watson-Crick base pair, can lead to developing a greater understanding of it's role in biology. Often comparing to experiments I have seen a stark contrast in the abilities of DNA and RNA to maintain this base pair. In an attempt to explain this difference, I have seen how cooperative shifts in the sugar puckers of DNA, that are necessary for the Hoogsteen base pair to form, are unobtainable in more rigid RNA sugars. I have also been able to observe how the dynamic equilibrium that exists between Watson-Crick and Hoogsteen base pairs within DNA can be influenced by

the binding of intercalating drugs, such as echinomycin. Finally, I have combined previous computational techniques to develop a novel way of obtaining time correlation functions from accelerated milestoning techniques, with potential application to biological systems such as the Watson-Crick Hoogsteen base pair transition.

Chapter 1

Introduction

It is very possible that all life on earth was based on nucleic acids at one point [1, 2, 3, 4]. This idea is known as the RNA world hypothesis. Cells are more sophisticated today and consist of many types of biological macromolecules. The primary actors within the cell are proteins and nucleic acids. Within every cell is a type of nucleic acid known as DNA. It acts as the storage vessel of genetic information. This information contains the blueprints for constructing all the proteins the cell shall use. Partnered with this process is another type of nucleic acid, RNA. While performing many roles in the cell, RNA most notably is used as a messenger between DNA and proteins bringing everything together[5]. The differences in their function and fundamental characteristics are still being explored today. With tasks so vital as keeping and facilitating the use of genetic information, nucleic acids play a pivotal role in the study of life, and likely have since life has existed. Studying nucleic acids at their fundamental level, can have wide reaching impacts and is an important goal of many scientists today[6].

This integral role nucleic acids have in all walks of life gives the study of nucleic acids broadly reaching impacts in the field of medicine. Cancer is wildly understood to be heavily related to

genetic material. The use of cancer therapeutic techniques centered around the manipulation of DNA directly is a promising area built on the foundation of a strong understanding of how DNA works in the human body [7]. Studying nucleic acids, can lead to not only improving the understanding of, but combating pathogens. Viruses, such as HIV, work using RNA to hijack a cells natural nucleic acid machinery, which can lead to developing autoimmune diseases such as AIDS [8]. And even health problems such as aging, are associated with breakdowns in the maintenance of the information stored in DNA [9]. In all of these areas of medicine and more, building a strong fundamental understanding of the inner workings of nucleic acids is necessary in the guidance of developing novel medicinal therapies.

While the study of nucleic acids is an important goal, this can be a difficult task. The original work by Watson, Crick, and Franklin, was a landmark step in determining the structure of DNA [10], but this really only showed a static image. Structural and sequential information gained about RNA and DNA, give great insight into the roles those particular sequences may play in a cell. However, in their participation in the cycles of life, these biological macromolecules are highly dynamic. The time scales of these dynamics can vary greatly from motions at the femtosecond scale to full seconds and longer[11]. Information on these dynamics can rarely be associated directly to sequence and structural data alone. Sophisticated experimental techniques have come a long way in recent years for helping to elucidate the dynamics of nucleic acids[12, 13]. Often in tandem with these experiments, models used in computer simulations have been growing for some time [14]. These computer simulations can allow for insights currently unattainable by experimental means alone. Here I present my work in using these computer simulation techniques to help give insight at the atomic level on Hoogsteen base pairs in nucleic acids. In addition I have worked on improving techniques that can be applied to studying biological macromolecules, such as nucleic acids.

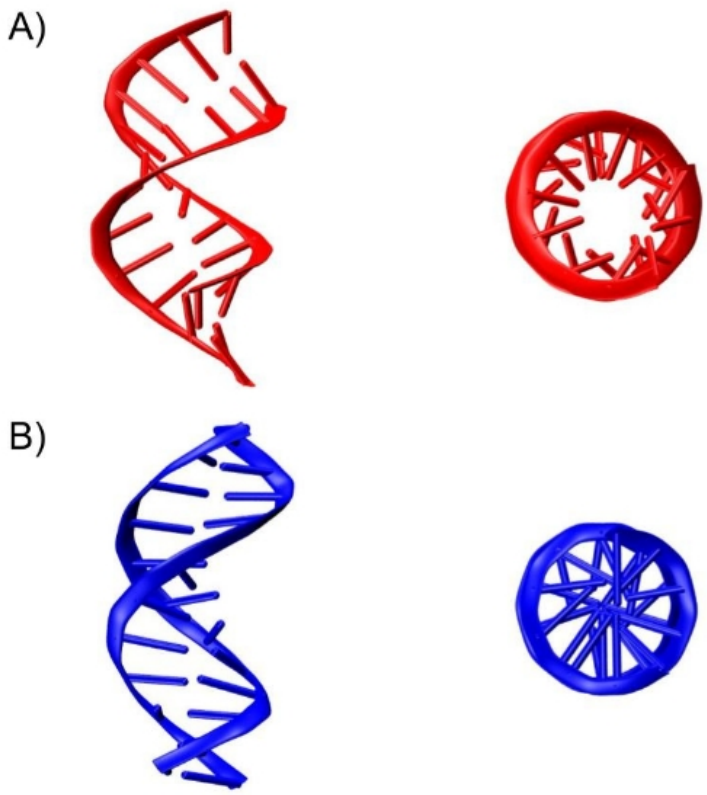

Figure 1.1: A) Side and top view of A-form RNA B) Side and top view of B-form DNA

1.1 Nucleic acids studies

1.1.1 RNA vs DNA

Nucleic acids consist of two majors parts: a backbone consisting of repeating units of a phosphate and furanose sugar, and nitrogenous bases bonded to each sugar. The fundamental difference between RNA (ribonucleic acid) and DNA (deoxyribonucleic acid) is the sugar. In addition, the bases each will contain are slightly different. In DNA, typically one of four bases are seen: adenine, cytosine, guanine, or thymine. In RNA, the first three will be the same, but in place of thymine, will be uracil. These molecular differences result in fairly significant conformational differences when a double helix is formed in either RNA or DNA[15].

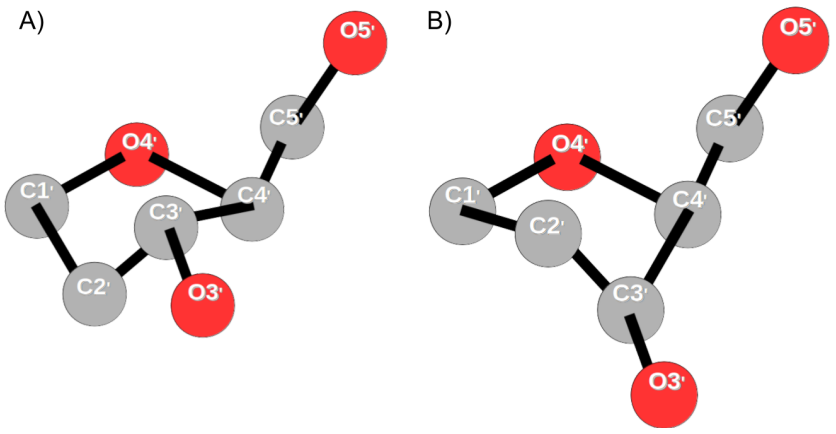

Figure 1.2: A) C3' endo Sugar pucker, predominant conformation of A-form helical nucleic acids B) C2' end Sugar pucker, predominant conformation of B-form helical nucleic acids

Though primarily seen as a single strand, RNA can sometimes can fold in on itself forming a helical hairpin loop. The shape of the helix that is formed in this case is known as an A-form helix, seen in Fig 1.1 part A. The predominant helical structure of DNA is a B-form helix Fig 1.1 part B. These distinctions contribute to the respective abilities to interact with proteins and small molecules. Both contain a major or minor, with the A form having a deeper major groove and wider minor groove. The distance between successive base pairs is narrower in A-from as well. Looking from above one can see the looser radius of the A-form[15]. All of these properties allow for significantly different access to the bases within the helix.

In addition to these global helical differences in conformation, there are internal differences as well, primarily seen and related to the sugar. The distance between the C1' on each sugar (the carbon bonded to the nucleobase) is slightly larger in the case of A form. This is related heavily to the tendency of the A-form RNA sugar to adopt a pucker conformation in the C3' endo state (Fig. 1.2 A). The primary sugar pucker conformation for B-form DNA results in the C2' endo state (Fig. 1.2 B). This favoritism is a direct consequence of the ribose having an additional oxygen on the C2' carbon. The C3' endo position gives a larger distance between the O4' oxygen and this ribose specific oxygen, alleviating energetically disfavorable van der waals interactions[16]. The C2' endo state allows for the greater inter

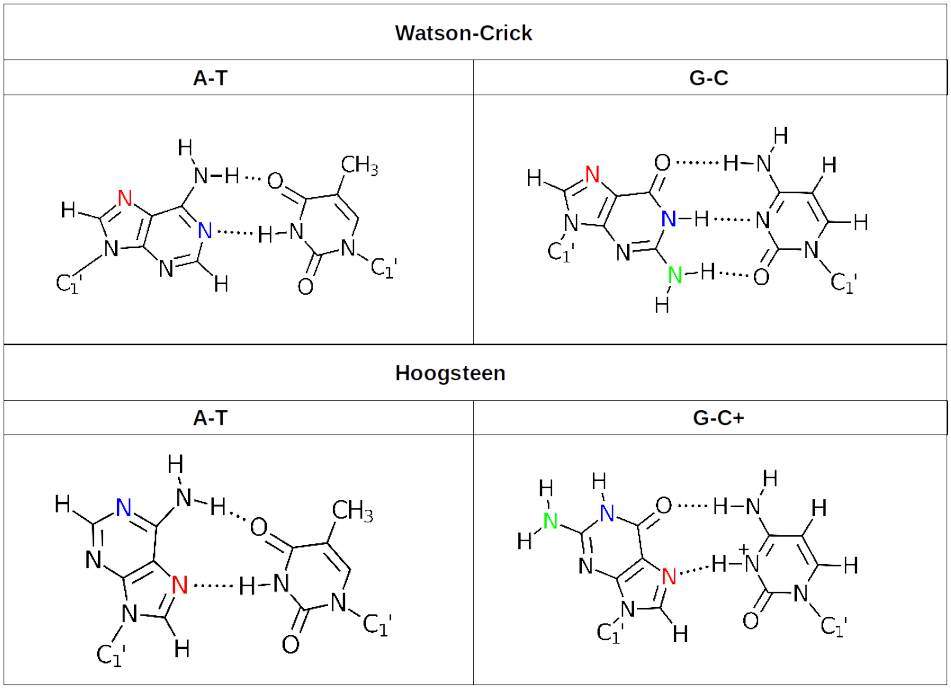

Figure 1.3: Top) Watson-Crick base pairs for A-T (left) and G-C (right) Bottom) Hoogsteen base pairs for A-T (left) and G-C+ (right)

base pair distance, as well as the closer C1' distance between nucleotides of the same base pair. Though these are the primary states for each respective helix, there are 10 states that they tend to cycle through dynamically while in equilibrium. The distribution among these sugar pucker states, tends to be greater in B-form helices.

1.1.2 Watson Crick vs Hoogsteen

Beyond the global conformations of A and B forms for the nucleic acids, and the various sugar puckers of the individual nucleotides, each base pair can adopt various conformations. The primary base pair conformations considered here will be the Watson-Crick (WC) and Hoogsteen (HG) base pairs.

The Watson-Crick base pair is named so by the famous scientists that originally solved the double helix structure of DNA, and predicted that the nucleobases of each strand would

interact with one another in a particular manner[10]. Adenine will pair with thymine (or uracil in RNA) while guanine pairs with cytosine. Watson and Crick also suggested a very specific hydrogen bond network between each (Fig 1.3 top). For A-T pairs, two hydrogen bonds are formed, only one of which is unique to the WC base pair. This is between nitrogen 1 (blue) of adenine and a nitrogen on thymine. In the case of G-C pairs, there are three hydrogen bonds in total, with two uniquely suggested by Watson and Crick. A similar hydrogen bond to the WC characteristic H-bond, from nitrogen 1 (blue) of guanine and a nitrogen on cytosine. In addition there is a hydrogen bond between nitrogen 2 (green) of guanine and an oxygen on cytosine. These particular hydrogen bonds help to stabilize the base pairs to form in this particular conformation and ultimately the double helical structures.

An alternative to this conformation of base pairs, was first proposed by Hoogsteen (HG) just a few years after Watson and Crick proposed theirs (Fig 1.3 bottom)[17]. The HG conformation is mostly distinguished by a 180^o rotation of the glycosidic bond that links the purine nucleobase to it's sugar. Consequence of this rotation result in the breaking of the WC characteristic and forming of hydrogen bonds unique to the HG arrangement. The A-T HG base pair now has the nitrogen 7 (red) of the adenine now paicipates in hydrogen bonding with a nitrogen on thymine. While in G-C HG base pairs, there is now only a total of 2 hydrogen bonds, with the unique one to HG being the nitrogen 7 (red) of the guanine with a now protonated nitrogen on cytosine. This rotation results in some rather large consequences for the nucleic acid's ability to interact with other macromolecules of biological importance.

Almost immediately it becomes apparent of the change in potential hydrogen binding sites of the bases to other potential hydrogen binding molecules. The hydrogen bonding availability in the minor groove is completely removed in both A-T and G-C pairs, while the number of potential sites increase in the major groove. Further the slight difference in affinity for

hydrogen bond formation between newly exposed hydrogen bond donors and acceptors may be relevant.

Though proposed more than 50 years ago, the biological relevance of Hoogsteen base pairs is still not completely understood. Evidence for the importance of HG base pairs has been mounting in the for some time in areas of DNA replication, damaged DNA, and DNA-Protein interactions [18]. According to structural surveys of the Protein Data Bank, HG base pairing is likely often overlooked in structure determination and many HG base pairs are perhaps currently mislabeled as WC [19]. However until the past decade, they've primarily been seen only in rare binding events.

Hoogsteen base pairs were seen readily in unique cases of protein bound DNA[20, 21], as well as in DNA-small molecule complexes[22]. However, evidence for Hoogsteen base pairs in naked DNA in solution was first seen by Nikolova et al in 2011 [23]. It was observed that WC and HG base pairs are in dynamic equilibrium with one another in some base pairs, and most prevalent in sequences with moderately long stretches of A-T pairs combined with 5'3' CA steps.

1.1.3 Molecular Dynamics

Molecular dynamics (MD) simulation techniques center around a classical approximation of molecular structure. To evaluate the dynamics of a particular system one needs not only some starting configuration, but the ability to evaluate the classical equations of motion. In MD this is done by determining the acceleration of some configuration of an atom $\vec{R}_a$ by

relating it to the force on that atom $\vec{F}_a$ through Newton's second law of motion:

$$\vec{F}_a = m_a \cdot \frac{d^2\vec{R}_a}{d^2t} \tag{1.1}$$

By using one of various numerical methods to solve this differential equation one can map out the positions of each individual atom as a function of time and generate a trajectory. Though this sounds simple enough, the accuracy of this model is highly dependent on the ability of generating the forces. The forces on each atom can be determined from the negative gradient of a potential energy surface (eq 1.2) that models the interactions of each atom with all the other atoms in the system. This is often referred to as a force field.

$$\vec{F} = -\nabla V \tag{1.2}$$

To do generate a molecular mechanics force field, typically a collection of harmonic approximations are used to simplify, what would otherwise need quantum mechanical treatments of covalent bonds, angles, and dihedral angles. In addition to these harmonic models of bonds, a collection of pairwise interactions between individual atoms are considered, to model electrostatics and van der waal forces. The combination of these models together creates what is known as the potential energy surface of molecular mechanics:

$$V = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 \sum_{dihedrals} + K_\phi(1 + \cos(n * \phi - \delta))^2 \tag{1.3}$$

$$+ \sum_{impropers} K_\omega(\omega - \omega_0)^2 + \sum_{LJ} \left(\frac{r_{ij}^{min}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{min}}{r_{ij}}\right)^{6} + \sum_{elec} \frac{q_{ij}}{\epsilon r_{ij}}$$

These are the primary terms of the CHARMM potential energy function, or force field[24]. The various constants (K_b,b_0,K_θ,θ_0, etc.) are parameters determined by fitting to various combinations of experimental data and quantum mechanics calculations. In this simplest form, this is sometimes referred to as the "ball and spring" model, from the significant use of the harmonic approximations. This approximation is just that and is often the limiting case on the accuracy of the results obtained from simulation. However, this is well understood within the community and parameterization is improved upon constantly to achieve the most accurate results possible [25].

In general this force field allows us to simulate biological molecules in solution efficiently with reasonable accuracy. Most simulations demonstrated in this thesis will be variations of molecular dynamics, the results of which offer insight onto the physical behavior of the Hoogsteen base pair.

1.2 Enhanced Path Sampling

1.2.1 Langevin Dynamics

While molecular dynamics is a useful technique, it does have its limitations. Often biological timescales can be far beyond those capable of molecular dynamics, especially in systems with a large number of atoms. Proteins in particular can have systems on the order or beyond 1,000,000 atoms. For an MD simulation this would require evaluation of forces on each atom for each timestep in the simulation. But often there are particular portions of a system that is the focal point of the study. In these cases, further approximations can be incorporated,

one such approximation results in Langevin dynamics.

Langevin dynamics creates a single particle whose dimensionality is determined by number of degrees of freedom deemed relevant to the system [26]. The other degrees of freedom are model by a stochastic term, often referred to as white noise, or a random kick. This addition of a random kick is not sufficient however, for modeling molecular systems. An additional term, that accounts for the friction that this high dimensional particle would necessarily experience, to further account for the difference from the molecular mechanics model. This combination results in the Langevin equation, written here in 1-d for simplicity:

$$F_r = -\gamma m\dot{r}(t) - \nabla V(r(t)) + \xi(t) \tag{1.4}$$

Where the first term, is the frictional component, with the coefficient of friction γ to dampen the momentum of the particle $m\dot{r}(t)$. The second term is the negative gradient of the potential energy surface $V(r(t))$ the particle evolves on. Finally the stochastic, random noise, term $\xi(t)$ introduces random motion to, what would otherwise be a completely deterministic system[26]. This then becomes a stochastic differential equation, requiring some unique means of integration, to ultimately obtain the equations of motion of the Langevin particle.

Though having various colloquial uses, the "white noise" term in the Langevin equation has a very specific meaning. This noise $\xi(t)$ obeys the fluctuation dissipation relation which gives it the following properties:

$$< \xi(t) >= 0 \qquad ; \qquad < \xi(t)\xi(t') >= 2k_B T m\gamma\delta(t - t') \tag{1.5}$$

The traditional example of the utility of this approximation is to use the coefficient of friction and the stochastic term to model the effects of solvent on a molecule[26]. The degrees of freedom of the molecule then become the dimensions in which the Langevin particle evolves, which allows the molecule to move according to the equations of motion described by the potential energy surface. All though this is of course sacrifices some accuracy with the increased number of approximation, the primary advantage here is the reduction in calculation time, as the solvent is typically the largest component of a molecular dynamics simulation.

As mentioned before, this recasts the problem from a deterministic one to a stochastic one. This allows for some other interesting advantages. Namely the fact that the dynamics are now influenced by Gaussian white noise, allow for the unique manipulations of trajectories generated in this way. Precisely, the probability a trajectory occurring under one set of conditions, can be determined when a trajectory is actually generated under a completely different set of conditions [27]. This leads to ways of accelerating dynamics, but returning natural kinetics which will be described in detail in the tail end of this thesis.

1.2.2 The Liouville and Fokker-Planck Operators

In particular when dealing with stochastic dynamics, and often in deterministic cases, it can be useful to assesses not individual trajectories, or coordinates evolving with time, but probability distributions with time. This section follows heavily the derivations of both the Liouville and Fokker-Planck operators from those in reference [28].

Rather than beginning with a starting configuration, one begins with the probability distri-

bution over all possible configurations:

$$f(x,t)_{t=0} = f(x,0) \tag{1.6}$$

As this by definition contains all the possible configurations, as time progresses, naturally at any given time, t, the probability distribution should be normalized:

$$1 = \int dx f(x,t) \tag{1.7}$$

This implies a conservation law that relates the changes in configurations with respect to time and the changes in the probability density with respect to time. For a completely deterministic system one has:

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\frac{\partial x}{\partial t} f(x,t) \tag{1.8}$$

By defining the Liouville operator, one is able to rewrite the conservation law in eq (1.8) and thus achieve the Liouville equation:

$$\hat{L} \equiv -\frac{\partial}{\partial x}\frac{\partial x}{\partial t} \qquad ; \qquad \frac{\partial f(x,t)}{\partial t} = -\hat{L} f(x,t) \tag{1.9}$$

This differential equation can then be solved in terms of it's initial conditions, giving the time dependent probability distribution for a deterministic system[28]:

$$f(x,t) = e^{-t\hat{L}} f(x,0) \tag{1.10}$$

With the inclusion of a stochastic term into equation (1.8) a similar treatment allows one to determine the time dependent probability distribution for systems described by Langevin dynamics.

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(\frac{\partial x}{\partial t} f(x,t) + \xi(t) f(x,t)\right) = -\hat{L} f(x,t) - \frac{\partial}{\partial x}\left(\xi(t) f(x,t)\right) \tag{1.11}$$

This differential equation will have a similar solution as before, the first term in fact is identical, but now the time dependent probability distribution $f(x,t)$ must also be dependent on previous random kicks from the stochastic term, as well as the initial distribution $f(x,0)$

$$f(x,t) = e^{-t\hat{L}} f(x,0) - \int_0^t dt' e^{-(t-t')\hat{L}} \frac{\partial}{\partial x} \xi(t') f(x,t') \tag{1.12}$$

Subsequently this solution can be substituted back in to the conservation in equation (1.11) to further discern the impacts of the stochastic integral term.

$$\frac{\partial f(x,t)}{\partial t} = -\hat{L} f(x,t) - \frac{\partial}{\partial x}\left(\xi(t) f(x,0)\right) + \frac{\partial}{\partial x}\left(\xi(t) \int_0^t dt' e^{-(t-t')\hat{L}} \frac{\partial}{\partial x} \xi(t') f(x,t')\right) \tag{1.13}$$

The iterative nature of this problem begins, to become apparent. In order to combat this, the average over the noise must be taken. Keeping in mind that $\xi(t)$ has zero mean and obeys the fluctuation dissipation relation in eq (1.5). This means the second term must go to 0, under averaging with respect to the noise and the remaining stochastic terms can be combined. This results in a Fokker-Planck equation:

$$\frac{\partial f(x,t)}{\partial t} = -\hat{L}f(x,t) + \frac{\partial}{\partial x} 2k_B T m\gamma \frac{\partial}{\partial x} f(x,t) \tag{1.14}$$

Which allows for a definition of a Fokker-Planck operator:

$$\mathcal{D} \equiv \frac{\partial}{\partial x}\frac{\partial x}{\partial t} + 2k_B T m\gamma \frac{\partial^2}{\partial^2 x} \tag{1.15}$$

This operator allows for the evaluation of the time dependent probability distribution of dynamical observables, that can be described by stochastic diffusive processes.

1.2.3 First passage times

When studying the transitions of a molecular system, it is often useful to frame the problem in a similar manner to the Kramer's problem[29]. This involves a particle leaving a potential energy well and crossing a potential barrier. Naturally for discussion on dynamics, one might ask how long such a process might take. This is often referred to as the notion of a first passage time. This treatment again will follow heavily the discussion by Zwanzig on the matter of first passage times [28].

If a particle following Langevin dynamics is trapped in a volume V, the first passage time τ, would be the amount it takes for the particle to be absorbed at a boundary ∂V. The location of this particle could be described by the configuration x, who's starting configuration would be x_0. The corresponding time dependent probability distribution, $f(x,t)$ would necessarily have some specific properties, when being evaluated Fokker-Planck operator $\mathcal{D}$:

$$\frac{\partial f(x,t)}{\partial t} = \mathcal{D} f(x,t) \tag{1.16}$$

The solution of course being:

$$f(x,t) = e^{-tD} f(x,0) \tag{1.17}$$

This initial distribution,$f(x,0)$, would be perfectly described by the delta function: $\delta(x - x_0)$. If no time has been able to pass then the current configuration must be the starting configuration. Second, the limit of the time dependent distribution is known as well, for given an infinite amount of time, the particle must be absorbed by the boundary.

Next, it is unimportant, how the particle leaves V, but simply that it does. So the focus must be on the time dependent probability not over the configurations at time t but the total probability over the entire volume:

$$S(t,x_0) = \int_V dx f(x,t) \tag{1.18}$$

This of course is still dependent on the initial configuration. This distribution contains all possible probabilities for the particle to leave the volume at some time. The time derivative of this therefore tells us how likely a given time τ is to be the time for how long it took for the particle starting in x_0 to be absorbed by the boundary. In other words, it gives the distribution of first passage times for a given starting configuration :

$$K(t, x_0) = \frac{dS(t, x_0)}{dt} \tag{1.19}$$

This distribution of first passage times will be one of the central quantities obtained in the enhanced sampling methods shown in this thesis.

1.2.4 Markov Chains and Random walks

A Markov Chain process, is one in which there are specific states that are able to be obtained, and that the current state is what dictates evolution in to the next [26]. In other words if rather than a continuous configuration x, a system is described by a collection of n discrete states one has:

$$\mathbf{X}_n = \{X_1, X_2, ..X_n\} \tag{1.20}$$

In this scheme states in between X_1 and X_2 are not allowed, but all defined states are accessible. There can be transitions between any two states, that have a non zero transition probability. For the second condition of a markov chain to remain true, these transition probabilities must be solely dependent on the currently occupied state. The collection of all

pairwise transition probabilities K_{ij} results in the transition matrix:

$$\boldsymbol{K} = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1n} \\ K_{21} & K_{22} & \dots & K_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n1} & K_{n2} & \dots & K_{nn} \end{bmatrix} \tag{1.21}$$

With the construction of this matrix for a given markov chain one is able to then conduct random walks, using tools such as rejection sampling or other Monte Carlo based methods[26].The matrix contains the probability of any step in a walk from the X_i configuration to the X_j configuration, which would be K_{ij}. With this knowledge an appropriate step can be chosen, and each individual step is only dependent on its' current configuration X_i.

With enough information to construct the transition matrix, and appropriately selected states to construct the markov chain, random walks can be extremely fast and efficient tools.

1.2.5 Time Correlation functions

Time correlation functions are invaluable quantities often calculated in biophysics[30, 31, 32]. They can lead to significant insights on dynamical information, often seen as rate constants or order parameters for various processes [28]. Through simulation of biological molecules, trajectory data can be used to determine these time correlation functions [33].

The general form of a time correlation function of a dynamical observable $A(x,t)$ would be:

$$C(t) = \langle A(x_0,0)A(x,t)\rangle = \int dx A(x_0,0)A(x,t)f(x,\infty) \tag{1.22}$$

This gives the relationship between the variable at time t, and what it was when it was at the beginning, weighted by the probability density at equilibrium of it being in configuration x, $f(x,\infty)$. At $t=0$ this is simply the variance $C(0) =< A^2 >$. In the long time limit $C(t)$ should converge to the mean squared, $< A >^2$. In other words, when the two measurements of A are so far apart in time that they are no longer correlated, on average, you would expect to see two measurements of the mean.

If from simulation, one can obtain the solution to the Fokker-Planck equation, $f(x,t)$, there are alternative means to calculating the time correlation function [34]. Specifically, first one calculates the expected value of arriving observing $A(x,t)$, given that is started at x_0 at time $t=0$, with this extra condition added, the time dependent probability distribution will be denoted as g for distinction from f which has no such requirement:

$$\langle A(x,t;x_0,0)\rangle = \int dx A(x)g(x,t;x_0,0) \tag{1.23}$$

Replacing this expression in to eq. (1.22) for $A(x,t)$, accounts for the various initial conditions, limiting g, ultimately resolving the restriction and returning the true time correlation

function.

$$C(t) = \int dx A(x_0, 0) \left(\int dx A(x) g(x, t; x_0, 0) \right) f(x, \infty) \tag{1.24}$$

Exploiting this feature allows for the unique means of calculation of time correlation functions, shown in this thesis.

Chapter 2

Biased and Equilibrium Molecular Dynamics studies of the Hoogsteen Base Pair in A6 DNA and RNA

2.1 Introduction

This section serves to illustrate my contributions to a journal article published in Nature Structural and Molecular Biology in 2016 by Zhou et al. entitled "m1A and m1G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs[35]."

The identification of HG base pairs in naked DNA was in part through the use of site specific methylation[23]. The methylation of adenine bases in DNA, at the N1 site inhibits the formation of the N1-N3 hydrogen bond unique to A-T WC base pairs, which in turn encourages the formation of HG base pairs[36]. This particular base modification was indicated to be extremely prevalent in the RNA transcriptome of eukaryotes [37]. It raises, therefore, interesting questions about the possibility of RNA to form HG base pairs in its primarily

A-form helical structures. While DNA has been shown to transiently adopt HG base pairs in unbound double helix structures, little has been seen with regards to the formation of HG base pairs in RNA duplexes. Though it typically adopts an A form helix in contrast with DNA's B-form, nothing on it's face should suggest RNA has a more difficult time to adopt the HG conformation over DNA. Exploring the possibility of RNA adoption the HG conformation was the focus of this work.

Highly sensitive NMR relaxation dispersion experiments were used to assess the viability of Hoogsteen (HG) base pairs in A-form RNA[38, 39, 40]. In contrast to B form DNA, which has been shown through the same methods to have a limited exchange between WC and HG states [23], the equivalent signal indicative of the transfer between the two states is not seen for A-RNA. Though this signal is weak, even in the context of B-DNA, it was concluded that this transition signal was outside the scope of the measurements, by varying temperature and pH conditions associated with heavily modulating the WC to HG relationship in DNA.

Further NMR experiments were attempted with the use of methylated forms of adenine and guanine. When purines are methylated at the N1 site it is seen to induce the HG state in DNA by preventing the hydrogen bond scheme formed in the WC state [41]. When these methylated purines were introduced to A-RNA, not only was melting of the methylated base pair site observed, but disruption of the neighboring WC base pairs were also observed.

With the experimental data suggesting heavily that A-RNA lacks the ability to maintain the HG state, molecular dynamics simulations were used to try and elucidate why this inability was seen. To make this assessment both biased molecular dynamics that induce transitions between WC and HG states, and separate simulations sampling the WC and HG states.

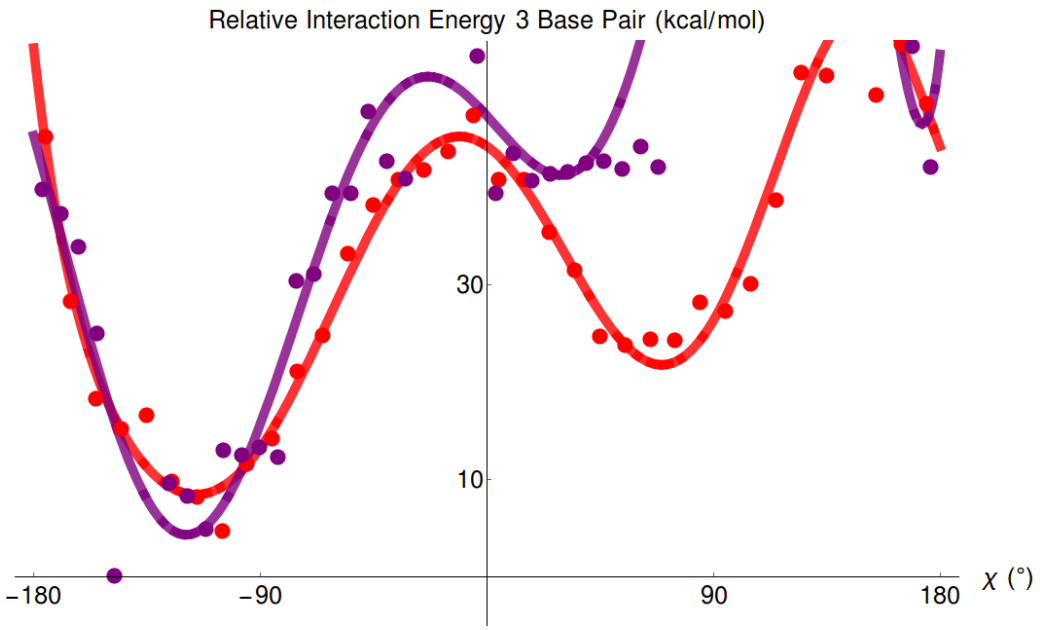

Figure 2.1: The mean interaction energy of the A16-U9 (T9 for DNA) base pair with surrounding base pairs (above, below) as a function of the glycosidic torsion angle (χ) for A6-DNA(red), A6-RNA(purple). Data points have been fitted with a polynomial function to guide the eye for each data set.

2.2 Results

2.2.1 Biased Molecular Dynamics

Biased Molecular Dynamics(bMD) were used to evaluate transitions of A16 in A6 DNA and A6 RNA between the WC and HG states [42]. In this context the reaction coordinate was defined by the distances between the hydrogen bonding atoms for the intended transition ie (WC to HG or HG to WC). As each state has a unique hydrogen bond, minimizing this distance encourages a flip.

The interaction energy of the base pair with the flipping base(U/T9-A16) as well the base pairs above and below the flipping base(U/T8-A17 and G10-C15) was calculated through out the trajectories of successful flips. After pairing the interaction energy with its corresponding value of the glycosidic torsion angle, the (χ,IE) pairs were grouped into 50 bins. Plotting

the (χ,IE) pairs for A6-RNA against A6-DNA allows for the comparison for each flipping environment (Figure 2.1)[35].

It can be seen in Figure 2.1 that A6-DNA (red) exhibits a clear two well system along the glycosidic angle. This is to be expected given previous experimental and computational studies demonstrating the stability of the HG state in the syn region (positive) in DNA[23]. In significant agreement with the NMR studies A6-RNA (purple) shows a significantly higher energy in the syn region than seen in A6-DNA[35]. Furthermore, when the transitions are induced in A6-RNA there's appreciable steric interactions with neighboring base pairs, in order to accommodate the flipping base.

2.2.2 Equilibrium Molecular Dynamics

Unbiased simulations of the A16-U/T9 HG base pair in various contexts were conducted as well. Simulations were run for with A16 in either WC or HG in various contexts including: A6-DNA, A6-RNA-hairpin, an inverted $3' \rightarrow 5'$ sequence of A6-RNA-hairpin, rA16-A6-DNA(ribose sugar replaces deoxyribose at A16), m^1A-A6-DNA, m^1A-A6-RNA. Viewing how the HG base pair compares to the WC in these various contexts leads to several conclusions about the most important conditions for the HG state.

Unsurprisingly the WC and HG simulations of A6-DNA remained stable for the duration of the simulations. In the case of A6-RNA starting in HG, in 2 of 10 trajectories, the A16 and U9 bases spontaneously swung out into the major groove. This matches well with the NMR optical melting experiments of methylated RNA. In several simulations of the inverted sequence A6-RNA, within the short window of simulation time, the HG base pair spontaneously flipped to WC. The results of the rA16-A6-DNA simulations were also consistent with the NMR data, in that little to no effect was seen.

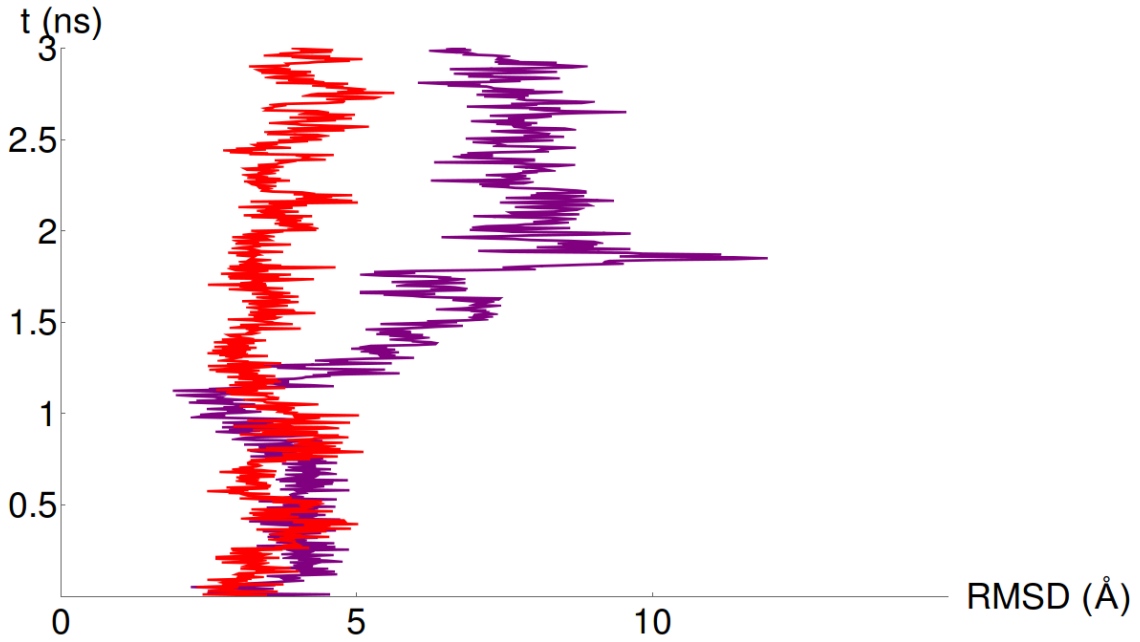

Figure 2.2: The simulation time vs global RMSD for m^1A-A6-DNA(red) and m^1A-A6-RNA(purple) has been plotted.

Finally comparison of the m^1A HG structures in A6-DNA and A6-RNA resulted in melting, or disruption of the RNA strands, while the DNA strand remains intact for the duration of the simulation. Figure 2.2 demonstrates this with monitoring the global RMSD over the time length of a single trajectory of each nucleic acid. Starting the RNA in the HG state and essentially blocking its ability to form the WC base pair results in denaturing of the secondary structure of the A-form RNA, as it does not seem to be able to maintain the HG state[35].

In an attempt to explain part of the reason for these differences in the various HG state contexts, the hydrogen bonds specific to HG base pairs were monitored(Table 1). Hydrogen bonds were considered formed when the appropriate hydrogen bond donor-acceptor distance and minimum donor-hydrogen-acceptor angle were satisfied; 3.6 Å and 120° respectively, following specifications from Goldsmith et al[43]. Due to the transient nature of the hydrogen bonds through out the duration of each trajectory, the percent of trajectory time in which a hydrogen bond is present, is shown(Table 2.1).

	N7—H-N3 %	O4—H-N6 %
B-DNA	87.3	98.4
A-RNA	2.4	99.5
A-RNA*	82.7	37.5
rA16-B-DNA	97.5	98.5
m^1A-B-DNA	61.6	99.7
m^1A-A-RNA	6.6	37.5

Table 2.1: The presence of the HG specific H-bonds are shown for various simulations. *Denotes a trajectory in which A16 and U9 flip out into the major groove.

The hydrogen bond presence for single trajectories representing each system gives great insight into why the various results are observed for each context of HG studied. It is clear that the characteristic N7-H-N3 formed between A16 and U/T9 of HG base pairs is barely formed in the case of A-RNA and m^1A-RNA[35]. A second A-RNA trajectory is shown to demonstrate the difference in hydrogen bond schemes when the A16 and U9 flip out into the major groove. It is only after the O4-H-N6 hydrogen bonds are broken that a hydrogen bond can be formed for N7-H-N3.

2.3 Discussion

Between the NMR experiments and MD simulations, there is considerable evidence to suggest that the HG state is difficult to maintain in A-form RNA. This difficulty arises primarily from the inability of RNA to form the HG characteristic hydrogen bond. This is not limited to simply the sugar of the flipping base, as demonstrated with the deoxyribose sugar at A16 in RNA. This suggests that the stabilization of HG in DNA may be more closely related to the B-form helical structure, as opposed to the specific chemical difference in sugars. The melting of neighboring base pairs when starting RNA simulations in HG paired with the breathing of neighboring base pairs when transitions between WC and HG, supports the idea that the A-form RNA is too rigid to accommodate the HG base pair.

The recognition of the destabilizing effects of the m^1A in RNA heavily suggests it's association with its biological relevance. With this additional categorical difference between RNA and DNA, further understanding on why they perform such different roles within the cell can be developed. Namely, of particular interest in DNA's ability to accept the damaging effects of the methylation, which can ultimately lead to mutations [44]. Therefore, the chemical modification can be used as a switch to regulate messenger RNA's ability to perform transcription[37]. Naturally it can be believed that the ability, or lack there of, to maintain the HG base pair, could play a role in these processes.

2.4 Methods

Structure generation for MD simulation: hp-A6-RNA, hp-A6-RNA 35, and A6-DNA helices were built using make-na [45] with all bases in WC conformation. In the case of hp-A6-RNA, a duplex structure was generated using make-na and the UUCG loop attached and annealed using the CHARMM simulation package [24]. Rotating along the glycosidic bond angle χ by 180^o created structures with HG conformation at A16. Unbiased MD equilibrium simulations were performed as follows. All structures were simulated using constant temperature MD with CHARMM36 forcefield [25] and a generalized Born molecular volume (GBMV) implicit solvent [46]; parameters for m^1A were taken from Xu et al[47]. Integration used a velocity-Verlet algorithm with a timestep of 1 fs. The cutoff for non-bonded list generation was 21 Å, the cutoff for non-bonded interactions was 18 Å, and the onset of switching for non-bonded interactions occurred at 16 Å. The SHAKE algorithm was used to constrain the covalent bonds to hydrogen atoms involved. Each structure was heated to 300.0 K with harmonic constraints on all non-hydrogen atoms, heating occurred in 1 ps increments of 1.0 K for a total of 300 ps steps, followed by 200 ps equilibration at 300.0 K. Harmonic constraints were then gradually removed during a sequence of 4 reductions for 50 ps each.

Unbiased production-run simulations were then run for 3 ns without constraints for each system. Ten independent simulations with hp-A6-RNA and rA16-A6-DNA with A16 in HG conformation were produced from independent conformations obtained during the heating and equilibration method described above. A6-DNA in HG was repeated twice.

Global RMSD was calculated from the single 3 ns trajectories of m 1 A starting in HG for both hp-A6-RNA and A6-DNA,in which $r_i(t)$ is the instantaneous coordinate of atom i and r_i^R is the position of the reference structure.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(r_i(t) - r^R)^2}{N}} \tag{2.1}$$

H-bond presence was evaluated using CHARMM's COOR HBOND module for each trajectory with cutoff distance and angle of 3.6 Å, and 120^o following Goldsmith et al[43].

Biased MD simulations: The protocols for minimization, heating, and solvation were identical to those used for the unbiased simulations. The biased molecular dynamics method [42] implemented in the CHARMM package was used to force conformational transitions between WC and HG states using a biasing potential $W(\rho(t))$ applied according to Equation 2.2,

$$\rho(t) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq 1}^{N} (r_{ij}(t) - r_{ij}^R)^2 \tag{2.2}$$

and

$$W(\rho(t)) = \begin{cases} \frac{\alpha}{2}(\rho(t) - \rho_a(t))^2 & \text{if } \rho(t) < \rho_a(t) \\ 0 & \text{if } \rho(t) \geq \rho_a(t) \end{cases} \text{ where } \rho_a(t) = \max_{0 \leq \tau \leq t} \rho(t)$$

$\rho(t)$ is a collective distance between the instantaneous (r_{ij}) and the reference structure (r_{ij}^R), and α the strength of the half-harmonic bias. In all cases, biases were placed between pairs of atoms that share a hydrogen-bond in the target structure, ensuring that the adenine base would not only perform the roughly 180^o flip, but also form the definitive hydrogen-bonding structure of the desired WC or HG configuration. After the biased trajectories were generated, they were post-processed in CHARMM, outputting the χ-angle dependence of the relative interaction energy value in the absence of the bias. The relative interaction energy was calculated for the base pair that includes the flipping base as well as the base pairs above and below the flipping base. Angle-energy pairs were binned into 50 bins and the mean of the energy was evaluated within each bin. Plots of relative interaction energy as a function of the χ-angle were thus generated.

Chapter 3

Biased and Equilibrium Molecular Dynamics studies of the Hoogsteen Base Pair in DNA in the presence of Echinomycin

3.1 Introduction

This section serves to illustrate my contributions to a journal article published in Nature Communications in 2018 by Xu et al. entitled "Modulation of Hoogsteen dynamics on DNA recognition.[48]"

The Hoogsteen base pair has been seen in numerous of contexts[18]. It has been fairly well established that a dynamic equilibrium exists between the WC and HG states in DNA[23]. Prior to this it was only seen in rather peculiar binding situations. One such area was the binding of drugs such as echinomycin [22][49]. This drug is a bisintercalator meaning it in-

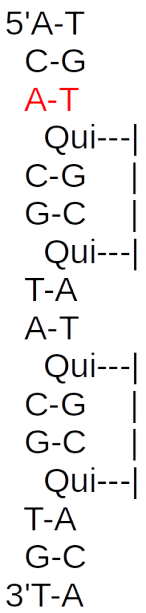

Figure 3.1: Palindromic sequence of DNA used in NMR and MD analyses with location of bisintercalated quinaxolone groups of echinomycin

serts its quinaxolone groups directly in between base pairs in the strands of DNA it binds to. This clearly would disrupt the stacking effects of neighboring base pairs and potentially fixes some in the Hoogsteen state. Hear it was demonstrated through NMR relaxation dispersion techniques and MD simulation, that in fact the drug allows for a dynamic equilibrium between WC and HG states, as in naked DNA, but a reduction in the energetic gap between the two states.

The DNA strand in Fig. 3.1 was evaluated as it is based on previous sequences to demonstrate for HG affinity in the presence of echinomycin [50]. Relaxation dispersion techniques were used to identify, the predominant presence of HG base pairs at the T6A7/A7T6 locations. While A3-T10 remained primarily WC, the relative population of HG saw an increase by 9 times. The identity of these conformational shifts were verified with trapping techniques, that use modified sequences at varying pH's known to induce HG modification.

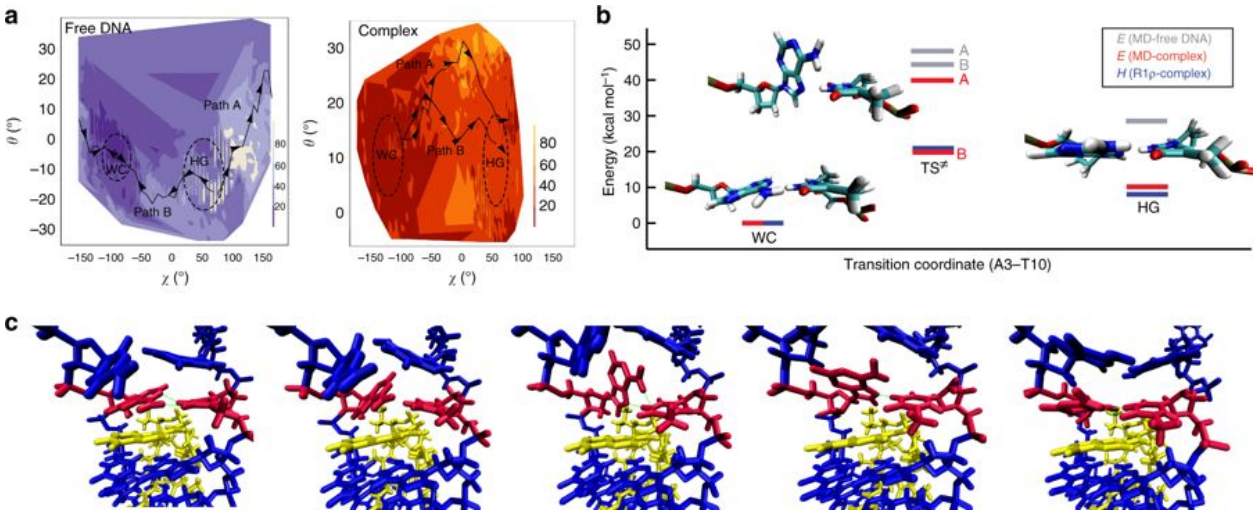

Figure 3.2: A) Contour plots showing the relative interaction energy (E kcal/mol) as a function of the base opening (θ) and flipping (χ) angles from multiple bias trajectories of Free DNA (blue) and DNA-echinomycin complex (orange). Dashed regions show the primarily sampled χ and θ angles of the unbiased simulations. Two paths can be seen for both systems. B) Comparison of the interaction energies of the WC,HG, and transition states of both paths observed in the DNA-drug complex and Free DNA to the enthalpies calculated from the relaxation dispersion experiments. C) Snapshots of an A3-T10 in the DNA-echinomycin complex transitioning between WC and HG states via path B.

3.2 Molecular dynamics studies

Biased and equilibrium MD simulations were used to examine whether or not Hoogsteen to Watson-Crick and Watson-Crick to Hoogsteen transitions are feasible in the presence of a nearby bound echinomycin molecule. Conversely, the signals observed by the NMR experiments, could readily be seen in cases where the HG and WC base pairs had become trapped at the time of drug binding. Due to the protonation of cytosine that occurs in the transition to the HG state, the focus of the simulations were on the A-T transitions, specifically at the A3 position, highlighted in red in Figure 3.1.

In total thirty biased simulations were run, each starting with a different initial velocity for both the Free DNA and DNA-echinomycin complex. Control simulations on A3T10 in the free DNA duplex resulted in several (4 of 30 simulations) successful transitions between WC and HG resulting in a Hoogsteen base pair. By comparison, all 30 simulations re-

sulted in successful transitions between WC and HG without echinomycin dissociation in the DNAechinomycin complex. Relative to the free DNA, the success of the transitions indicate a lower energetic barrier height for the WC to HG transition in the complex DNA as well as a decrease in energy of the HG base pair (Fig 3.2 part B). Both the barrier height and energetic differences computed by MD are in good agreement with the NMR RD measured counterparts (Fig 3.2 part B).

Successful transition pathways were predominantly clockwise about the χ angle in the case of the drug complex, but in the case of free DNA, rotation of the glycosidc bond was seen in both directions. The transition state of the predominant path (path B in Fig 3.2) part A) features a purine base that is near-orthogonal to its paired pyrimidine resulting in disruption of the neighboring WC base pair. Some stacking interactions with the quinoxaline rings are disrupted as well. The intercalating portion of the drug remains stacked on the flanking GC base pairs which can be seen in Fig 3.2 part C. The base flipping appears to be unhindered due to the flexibility of the neighboring base pairs that exhibit collective conformational changes during the transition.

The A7 location was also identified as likely to have a dynamic relationship between the WC and HG state. However, it is suspected that the neighboring A-T pair cooperatively transitions. In other words the T6A7/A7T6 simultaneously swap between WC and HG states. Simulations in which a single base pair was flipped successfully from Hoogsteen to Watson-Crick without disrupting the bound echinomycin were done. This supports the feasibility of having HG to WC transition in the presence of the bound echinomycin.

3.3 Discussion

There are two primary conclusions that can be reached from this work. First that the relative stabilities of WC and HG pairs must be in some capacity influenced by base pair stacking effects. Second, that intercalation did not inhibit the internal rotation of base pairs and still allows for an exchange between WC and HG states. With intercalators such as echinomycin that have been considered for sometime now as a possible cancer treatment [51, 52]. It is possible the mechanism by which it is able to contribute medicinally is related to it's effect on the regulation of HG/WC base pair exchange.

3.4 Methods

Coordinates for the E12DNAechinomycin complex were obtained by downloading the 1XVN structure 38 from the Protein Data Bank (PDB). The coordinates of the DNA portion of the complex were loaded into the CHARMM molecular modeling package and coordinates for the terminal two bps were generated using internal coordinate tables within CHARMM [24]. Both A3 bases were rotated 180 at the glycosidic bond to begin in the Watson-Crick conformation. Structures for control simulations of free DNA were generated through the use of make-na [45]. Each A7 was rotated 180^o along the glycosidic bond to begin in the Hoogsteen conformation. The coordinates of a single echinomycin molecule were loaded into Schrodingers Maestro program [53], to generate bond parameters. Bond parameter and coordinate information for echinomycin were entered into CHARMM CgenFFs automated atom typing program for generation of CHARMM force field parameters for the echinomycin [54, 55, 56, 57]. The DNA-echinomycin complex and control free DNA were each placed into cubic water boxes with side lengths of 87 Åwith 20,440 and 20,558 TIP3P water molecules, respectively [58]. To insure the neutrality of each system 31 Na^+ cations and 9 Cl^- anions

were added as well. Each system was equilibrated using constant temperature and pressure dynamics. Temperature was maintained at 300 K and pressure at 1 atm using the Nose-Hoover Thermostat [59]. Particle-mesh Ewald summation [60],[61] was used with cutoffs of 14 Åto calculate electrostatic potentials. Equilibration for each system ran for 300 ps using a leap verlet algorithm. From the final structure produced from the equilibration, for each system, 30 simulations were run under the exact same conditions of equilibration while varying the initial starting velocities sampling the immediate space near either the Watson-Crick or Hoogsteen states. The biased MD method [42] implemented in the CHARMM package was used to assess conformational transitions between Watson-Crick and Hoogsteen bps for A3 and A7 both in the presence and absence of echinomycin, using a biasing potential $W(\rho(t))$ applied according to equation 3.1,

$$\rho(t) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq 1}^{N} (r_{ij}(t) - r_{ij}^{R})^2 \tag{3.1}$$

and

$$W(\rho(t)) = \begin{cases} \frac{\alpha}{2}(\rho(t) - \rho_a(t))^2 & if \rho(t) < \rho_a(t) \\ 0 & if \rho(t) \geq \rho_a(t) \end{cases} \text{where} \rho_a(t) = \max_{0 \leq \tau \leq t} \rho(t)$$

$\rho(t)$ is a collective distance between the instantaneous (r_{ij}) and the reference structure (r_{ij}^{R}), and α the force constant of the half-harmonic bias in kcal/mol Å^{-4}. In all cases, biases were placed between pairs of atoms that share a hydrogen-bond in the target structure, ensuring that the adenine base would flip 180^o in the χ-direction, and form the appropriate hydrogen-bonding structure of the target conformation. Trajectories were post-processed in CHARMM, outputting the χ and θ angle dependence of the relative interaction energy value (this excludes the bias potential for the biased simulations). The relative interaction

energy was calculated for the bp that includes the flipping base as well as the bps above and below the flipping base. In this calculation, each atom of the base was evaluated individually for both bonded and non-bonded terms for the CHARMM force field, which includes the interaction with the ligand and the solvent effect. The (χ, θ, energy) points were binned into a 50 50 grid of bins for both angles and the mean of the energy was evaluated within each bin. Contour plots of relative interaction energy as a function of both θ and χ were generated.

Chapter 4

Computational studies of the relative stability of Hoogsteen base pairs in DNA vs. RNA

4.1 Introduction

The Hoogsteen (HG) base pair has gained considerable interest in the biophysical community in recent years [18]. With the exciting discovery of their appearances in naked DNA, the prospect of its biological relevance in various context, has received appreciable support [23][19]. In particular through the methylation of adenine and guanine at the N1 site, forces the adoption of the HG base pairs[41]. This subsequently allows for repair enzymes to recognize the site and check for inconsistencies in the genetic code[62, 63].

It is possible that the HG base pair's biological relevance does not end at processes concerned with DNA. This same methylation of purines in RNA is seen widely as post-transcriptional switches [37]. It was shown recently that this modification in A-form RNA results in base pair

melting, primarily due to an increase in instability of the HG base pair in RNA compared to DNA [35]. Further NMR experiments and MD simulation have characterized the energetics of why A-form RNA cannot maintain the HG base pair[64].

Here for the first time the relative conformational entropy of the HG base pair will be assessed in both B-form DNA and A-form RNA. It will be shown that in comparison to their respective WC base pairs, the HG base pair is more entropically disfavored in B-form DNA and than A-form RNA. However, the well established flexibility of 5'-3' CA steps[65] of B-form DNA, a cooperative effect that does not exist in A-form RNA, results in a greater diversity of sugar puckering states. This paired with a consistently being able to form the HG characteristic hydrogen bond, results in a much lower value in the PMF difference in DNA than RNA. This is supported by state-of-the-art enhanced sampling methods for calculation of Potential of Mean Force (PMF) surfaces, qualitatively consistent with previous surfaces [66].

4.2 Results

The conformational entropy was calculated for A6-DNA B-form with A16 in either WC or HG and A6-RNA-hairpin A-form with A16 in either WC or HG. Due to the flexibility of the terminal base pairs and the loop in RNA, these were omitted from the entropy calculation. In the case of A6-DNA it was observed that the all WC strand was more favorable than the HG containing strand, with a $T\Delta S = -3.2$ kcal/mol. In contrast,there was an increase in the conformational entropy for the HG state in A6-RNA-hairpin with a $T\Delta S = 6.44$ kcal/mol. The direction of both of these entropy differences were seen in various subsegments including the backbone of the side of the strand with the flipping base, the C15A16 nucleotides, the sugars of these two nucleotides, the T/U9-A16 base pair, and the single A16 nucleotide. All of which can be seen in the table below.

Segment considered	DNA $T\Delta S$ (kcal/mol)	RNA $T\Delta S$ (kcal/mol)
Majority of strand	-3.2	6.4
Backbone of flipping side	-6.0	1.9
C15,A16	-6.9	2.8
C15,A16 sugar only	-6.5	1.0
T/U9-A16	-0.69	0.015*
A-16	-3.0	2.5

Table 4.1: Conformational entropy difference between simulations containing HG and WC various segments of A6-DNA B form and A6-RNA-hairpin A form. Segments considered included the base pairs between G2-C23 and G11-C14 for both strands, removing the most flexible portions of each strand, referred to as the full strand. In addition, the sugar-phosphate backbone of the A-16 flipping side between C14 and C23, the nucleotides of C15 and A16, solely the sugars of C15 and A16, the nucleotides of T/U9 and A16, and just the A16 nucleotide were considered. *In the case of the T/U9 and A16 of RNA, the calculations of each states entropy did not converge to within the magnitude of the difference between them.

This suggests an increase in rigidity from the WC to HG state within this C-A step of DNA, which is commonly associated with an increased presence of HG. With a considerable change in entropy within the sugars of C-A step, the sugar pucker distribution for each was considered for both DNA and RNA.

The calculation of the pseudorotation angle (P) as described by Altona and Sundaralingam, allows for the determination of the sugar pucker conformation of each nucleotide. The distribution of sugar pucker conformations adopted by the A16 and C15 nucleotide were the only appreciably different distributions in the HG containing duplex when compared with to the distributions of the purely WC duplex for both DNA and RNA.

Namely the primary conformation of the sugar pucker of A16 nucleotide in WC is shifted from an average C2' endo conformation to a O4' endo conformation when in the HG state. Paired to this the immediate neighbor C15 nucleotide's sugar pucker adjusts from an average C1'exo conformation to a C2' endo conformation. Beyond these average conformational shifts, it was observed that the entire distribution of pucker conformations is adjusted with a decrease in variance for both. Though a minor shift to C4' exo occurs in both A16 and C15 sugar puckers for RNA the dominant sugar pucker remains to be C3' endo, in addition

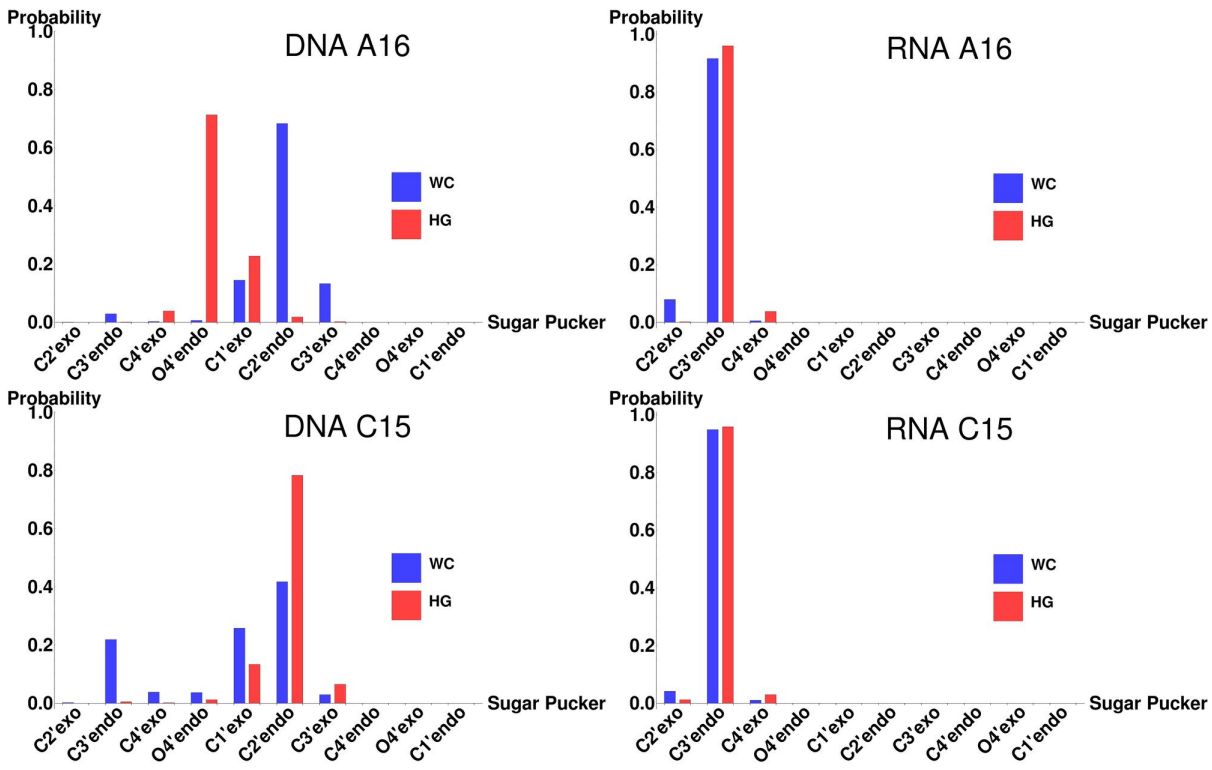

Figure 4.1: Probability of sugar pucker conformations for A16 (Top) and C15 (bottom) in DNA(left) and RNA(right) WC(blue) or HG(red)

the variability in pucker states remains mostly unchanged from it's narrow distribution in the WC case.

We also used the newly developed meta-eABF method to calculate the two dimensional potential mean force (PMF) for the WC to HG transition in A6-DNA and A6-RNA segment. We could get a converged PMF within 200 ns compared to 6-40 s simulations for other computational studies. The presence of Hoogsteen base pair is in the DNA is confirmed by the presence of clear deep minima at $\chi \approx 50^o$ and $\theta \approx 0^o$. The Watson-Crick base pair is depicted by the deep stretched minima between -180^o and -50^o of the glycosidic angle χ for θ close to 0^o. Meanwhile, for the Hoogsteen base pairs. This Hoogsteen-like structure is energetically more unstable relative to the Hoogsteen structure of the DNA likely due to a presence of only a single hydrogen bond. There is also a high barrier in RNA for direct transition between HG and the other minima and the lower energy path is through higher

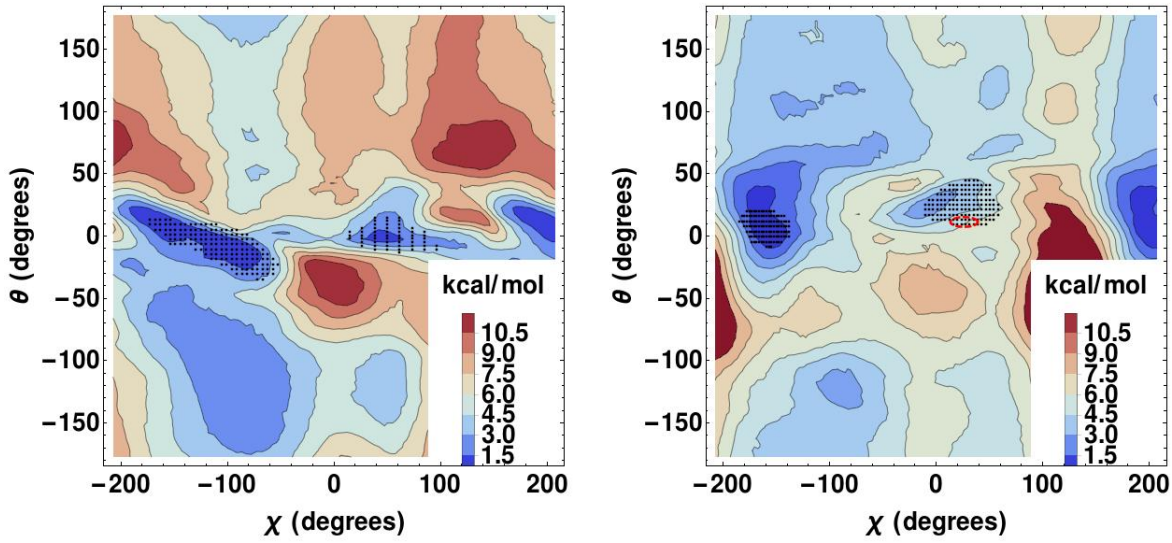

Figure 4.2: PMFs of the glycosidic torsion angle (χ) and the flip out angle(θ) for A6-DNA(left) and A6-RNA-hairpin (right). Overlayed with black dots showing sampled angles of the equilibrium simulations. A dashed red circle shows the portion of the RNA HG simulations that allowed for the HG characteristic hydrogen bond.

values of θ indicating base pair melting.

For comparison to the equilibrium studies the sampled χ and θ angles of the equilibrium simulations are overlayed onto the PMFs. In these simulations less than 2% of all the configurations sampled had the characteristic N7-N3 hydrogen bond formed between A16 and U9. The region in which this hydrogen bond was formed can be seen in the red dashed circle. This region is clearly distinct from the majority of configurations sampled and the intermediate HG well expected by the PMF calculations.

4.3 Discussion

From this information two primary inferences can be made. A primary contributing factor for the 5'3' CA step allowing for an increased propensity of HG base pairs in DNA, is likely the large flexibility in the cytosine sugar. The clearly cooperative effect between the sugars

seems to allow the loss of entropy, as the cytosine sugar of DNA is already in a broad distribution of states. Secondly, though there is a large increase in entropy for RNA, it is clearly not localized to the 5'3' CA step as it is in DNA. Couple this with the fact that RNA lacks the ability to form the characteristic HG hydrogen bond, suggest explanations to for RNA not actually maintaining a HG base pair, but leads closer to the melting of base pairs, which would be a highly entropic state, but also increase in energy compared to the hydrogen bond stabilized base pairs.

4.4 Methods

4.4.1 System preparation and equilibration

All input files were generated using CHARMM-GUI web server [67, 68] and VMD [69]. All simulations were performed using NAMD [70] package in GPU. For A6-DNA and A6-RNA the nucleic acid structures were made using the make-NA server [45]. All the structures were solvated using TIP3P water in rectangular solvation box with 17Å of water padding in each direction. CHARMM36 force field for nucleic acids [25]. Energy minimization was performed for 10000 steps using conjugate gradient algorithm.

The system was gradually heated to 298 K temperature at a rate of 1K/ps in NVE ensemble with harmonic constraints of 3 kcal/molÅ^2 on the nucleic acid heavy atoms. Then the constraints were removed in steps 0.5 kcal/molÅ^2 per 200 ps. The terminal base pairs are restrained with a small harmonic force constant of 0.1 kcal/molÅ^2 and the system was equilibrated for 3 ns in NVE ensemble and 10 ns of NPT ensemble with force constant 0.05 kcal/molÅ^2.

4.4.2 Enhanced sampling simulation

The newly developed meta-eABF method [71] implemented in NAMD 2.12 package through colvars module [72] has been used to obtain the PMF surface. The glycosidic angle χ and the pseudo-dihedral angle θ of the A-T base pair were chosen as order parameters following earlier studies [23, 48, 73, 74]. The collective variable space was spanned from -180° to 180° with 5° bin width for both the dihedral angles. For all the systems the meta-eABF simulation was started from the final structure from the NPT equilibration and continued till all the bins are explored and the RMS difference of the PMF's are converged. For most systems we got converged PMF within 0.2 μs of simulation. The one-dimensional PMF along glycosidic angle χ was obtained using the following expression

$$A(\chi) = -k_B T \ln \left(\int \exp \left(- \frac{A(\chi, \theta)}{k_B T} \right) d\theta \right) \tag{4.1}$$

4.4.3 Equilibrium simulations

Starting from the final frame of the NPT equillibration step, each of the 4 systems (A6-RNA WC/HG, A6-DNA WC/HG) were run in NAMD NPT ensemble with periodic boundary conditions. Simulations were held at 298 K and 1 atm using the Nose-Hoover Thermostat [59]. Particle-mesh Ewald summation was used to account for electrostatics with periodic images [60, 61]. Each simulation ran for a total of 25 ns with 1.0 fs time steps, saving coordinates every picosecond. Each simulation was repeated a total of 5 times, with different starting velocities

4.4.4 Conformational Entropy Calculations

Trajectories were post processed in CHARMM [24]. After removing the first nanosecond from each trajectory, the translational and rotational degrees of freedom were reduced and the molecule was oriented using CHARMM's ORIE command. Each of the 5 trajectories for each system were combined, resulting in 120 ns of sampled conformations for each system.

Conformational entropy calculations were performed with varying atoms, the collections of which can be seen in Table 4.1. These calculations were performed using quasihamoinc analysis frequency analysis [75, 76]. These frequencies (ω_i) are calculated, from the eigenvalues of the mass weighted covariance matrix, determined from the 120 ns trajectories. The frequencies are used to calculate the absolute entropy of each conformation with:

$$S = k_B \sum_{i}^{3n-6} \frac{\hbar\omega_i / k_B T}{e^{\hbar\omega_i / k_B T} - 1} - \ln(1 - e^{-\hbar\omega_i / k_B T}) \tag{4.2}$$

For each conformation $T \cdot S$ had converged within 0.05 kcal/mol for the last 10% of sampled configurations.

Chapter 5

Long-time correlation functions from biased Langevin dynamics and Markov chain walks

5.1 Introduction

Calculations of time-correlation functions can be invaluable in linking simulations and experiments of biological macromolecules [77, 78, 79]. They can be used for attaining pertinent kinetic information such as rate constants and order parameters. However the various timescales relevant to biological processes can be inaccessible to many simulation techniques [80]. This barrier can make it difficult to calculate time-correlation functions that capture longtime events. For kinetic information to be obtained, samples of configurations need to be collected in sequence with knowledge about the time between the configurations occurring [81]. Collection of all of this information can be time consuming and attempts at acceleration of this process can lead to loss of information. In an effort to overcome these temporal

barriers methods such as Steered Molecular dynamics (SMD) are sometimes employed [82]. This method uses a perturbing force in the MD simulation to induce conformational changes that otherwise would not occur within the time span of the simulation. These can give new information on low energy pathways and allow sampling of configurational states that were otherwise inaccessible. The trade off when using methods such as this is that information about the natural time between the various configurations along the pathway is lost. In other words, though a new high energy state may have been reached, and knowledge about the appropriate order of intermediate configurations between two states might be learned, information about how long process would take along the way is lost. When applying a similar concept of a perturbing force to Langevin dynamics there are means of recollecting this temporal information through a rewieghitng scheme detailed in the theory section of this paper [27]. This takes advantage of a the stochastic term within Langevin dynamics to determine how likely a particular trajectory is to have occurred on the unperturbed potential. This means dynamical information about the original potential energy surface can be determined faster than their normal evolution on that surface. This frame work, however, is still limited to the time spans accessible to a single trajectory of Langevin dynamics.

Milestoning is a technique that places hyperplanes in phase space to divide it into subsections that can be simulated separately [83]. The primary computational benefit to running simulations in this way, is that allows each subsection to be run in parallel [84]. This allows for exploration of a large amount of configurations in much shorter time then exploring the sames space from long continuous trajectories. In its inception this was an excellent tool for determining thermodynamic properties. Over the past several years there have been enhancements to the milestoning algorithm to improve on the ability to obtain these thermodynamic values [85], as well as novel ways of using the information obtained from these simulations to obtain kinetic data [84]. More recently a milestoning method incorporating the reweighting scheme with Langevin dynamics, known as Wind-Assisted Reweighted Milestoning Method (WARM) was proposed by Grazioli and Andricioaei [86]. Further, the same

authors proposed a means for the calculation of time-correlation functions from Milestoning simulations through the use of Markov chain random walks [87]. In this paper, the two methods shall be combined for the first time.

5.2 Theory

5.2.1 Wind-Assisted Reweighting Milestoning

As mentioned earlier, Milestoning has allowed for the calculation of equilibrium quantities from the beginning of it's use. The way this is accomplished is by first determining the probability distribution of being at some particular milestone A at time t, $P_A(t)$ shown in eq (5.1) [83]. Which must be the probability of arriving at milestone A, in t amount of time, from any neighboring milestone described by: $Q_A(t)$ combined with the probability of not leaving milestone A within that same amount of time frame. The probability of leaving, or transitioning from milestone A to some neighboring milestone B after it has been at A for τ length of time, can be expressed by probabilities of various transition times τ between milestone A and B: $K_{AB}(\tau)$ [83]. For a set of M milestones, all cases where milestones A and B are not nearest neighbors then $K_{AB}(\tau) = 0$ as any trajectory is to be terminated upon reaching a milestone.

$$P_A(t) = \int_0^t Q_A(t') \left[1 - \sum_{B=0}^{M} \int_0^{t-t'} K_{AB}(\tau) d\tau \right] dt',$$

$$Q_A(t) = 2\delta(t) P_A(0) + \sum_{B=0}^{M} \int_0^t Q_B(t'') K_{BA}(t - t'') dt'' \tag{5.1}$$

In the second line of eq 5.1, the probability of arriving at a milestone A, is determined from whether a trajectory begins at milestone A at time $t = 0$, corresponding to the first term, or reached a neighboring milestone B, sometime earlier and has now arrived at milestone A [83]. Thus all $P_A(t)$ and $Q_A(t)$ are determined from the collection of all $K_{AB}(\tau)$. So if one can generate each of these distributions, they shall contain all appropriate kinetic information needed for the description of transitions over the reaction coordinates divided by the milestones.

It is here that the WARM method becomes advantageous, as the reweighting scheme allows for expedited calculations of these distributions [86]. The $K_{AB}(\tau)$ distribution can be described by a conditional probability distribution of configurations. If M number of milestones are defined by a set of configurations $\{x_s\}$, the probability a transition to a neighboring milestone B occurs in τ amount of time provided that it was at milestone A at time 0 is:

$$K_{AB}(\tau) = P(x_B, \tau | x_A, 0) \tag{5.2}$$

Such a description of transition distributions pairs nicely with Langevin dynamics. The use of Langevin dynamics allows for a relatively simple means of determining these distributions from the generation of Langevin trajectories. The standard Langevin equation for trajectory generation in configuration space is shown in eq (5.3), where γ is a coefficient of friction, the potential energy of the system is defined by $V(x)$, and $\xi(t)$ is a random force, sometimes refer to as a "kicking" force [26]. This random force is typically described by Gaussian white noise with a mean of 0 and obeys the fluctuation dissipation theorem so that it's variance w is defined as: $w \equiv \langle \xi(t)\xi(t') \rangle = 2k_B T m\gamma$ [26].

$$m\ddot{x} = -\gamma m\dot{x} - \nabla V(x) + \xi(t) \tag{5.3}$$

Conditional probabilities such as eq (5.2) can be related to this stochastic random force term $\xi(t)$ in this way [28]:

$$P(x_B, \tau | x_A, 0) = \int D\xi W[\xi(t)] \delta(x(\tau) - x_B) \tag{5.4}$$

Using the delta function here, we select only for movements to nearest neighbor milestone configurations. Milestoning trajectories terminate once a milestone is reached, therefore only nearest neighbor transitions must be considered. Here, $W[\xi(t)]$ is the probability distribution of all random paths generated from $\xi(t)$. Since the noise is defined to be Gaussian distributed and it's mean and variance defined as 0 and w respectively, then naturally this distribution has the form:

$$W[\xi(t)] = \exp(-\frac{1}{2w} \int_0^t \xi(t')^2 dt' \ ) \tag{5.5}$$

Rearranging eq (5.3) to isolate $\xi(t)$ and squaring both sides, the integral within eq (5.5) becomes, what is known as the Onsager-Machlup action functional [88, 89]:

$$S[x(t)] = \int_0^t (m\ddot{x}(t) + \gamma m\dot{x}(t) + \nabla V(x(t)))^2 dt \tag{5.6}$$

Using the Ito formalism of stochastic calculus, a Wiener integral is able to be constructed and a variable change made in eq (5.5) from $\xi(t)$ to $x(t)$, with a functional Jacobian that

can be shown to be unity, the conditional probability can now be written in this form [90]:

$$P(x_B, \tau | x_A, 0) = \int_{(x_A,0)}^{(x_B,\tau)} DxW[x(t)]$$

$$W[x(t)] \equiv \exp(-\frac{S[x(t)]}{2w}) \tag{5.7}$$

This form allows for the determination of the likelihood a collection of random kicks will be generated from the white noise and produce a specific trajectory $x(t)$. This information can be used to compare the likelihood a trajectory being generated on a specific potential might be able to be generated on some other different potential [27]. This is what allows information generated from Langevin dynamics to be reweighted. This feature is exploited in the WARM method [86]. With inclusion of a perturbing force on some potential, a Langevin simulation can be coaxed in a specific direction. For the WARM method this is referred to as a "wind" force, $F_{\text{wind}}(x,t)$ [86]. With an additional wind force the Onsager-Machlup functional becomes:

$$S_f[x(t)] = \int_0^t (m\ddot{x(t)} + \gamma m\dot{x(t)} + \nabla V(x(t)) - F_{\text{wind}}(x,t))^2 dt \tag{5.8}$$

Comparing this action to one without the $F_{\text{wind}}(x,t)$ allows one to determine the probability

or weight a trajectory generated with the wind, would have without that wind.

$$\frac{W[x(t)]}{W_f[x(t)]} = \exp(-\frac{S[x(t)] - S_f[x(t)]}{2w}) \tag{5.9}$$

In practice, the discrete form of each action is calculated with an overdamped approximation, shown in eq (5.10).

$$S_f[x(t)] \approx \sum_i \left(m\gamma \frac{\Delta x_i}{\Delta t} + \nabla V_i - F_{\text{wind}}(x,t) \right)^2 \Delta t$$

$$S[x(t)] \approx \sum_i \left(m\gamma \frac{\Delta x_i}{\Delta t} + \nabla V_i \right)^2 \Delta t \tag{5.10}$$

In order to calculate each $K_s^{\pm}(\tau)$, a histogram of times from many Langevin trajectories are generated for each possible transition from every milestone. The $F_{\text{wind}}(x,t)$ is chosen to "blow" in the direction of the intended milestone for whichever transition distribution, a histogram is currently being generated. Each simulation will then reach the necessary milestone faster than without any assistance. Instead of counting each trajectory as a whole in the histogram, the relative weight, $\frac{W[x(t)]}{W_{F_{\text{ext}}}[x(t)]}$, of each trajectory is counted. This allows for the histograms to be generated faster with $F_{wind}(x,t)$, but the features of this histogram should be the same as if there were no wind [86]. The ability to reproduce equilibrium values, such as the equilibrium flux through each milestone, has been previously demonstrated. In this paper it will be shown that kinetic information is retained as well, through the incorporation of the random walk/ stochastic path integral method of calculating time correlation methods

from milestoning data, originally proposed by Grazioli and Andricioaei, detailed in the next section [87].

5.2.2 Time Correlation Milestoning

The calculation of a time-correlation function using milestoning data, is possible if the set of M number of milestones described by a set of configurations $\{x_s\}$, are an appropriate discrete approximation of the configuration space x [87]. Also required is the stationary probability distribution $P_s(\infty)$, which is attained with any milestoning simulation, where this approximation is valid.

In general a time-correlation function $C(t)$ of a time dependent property $A(x,t)$, where the equilibrium distribution of variable x is $f(x)$ would be:

$$C(t) = \langle A(x,0)A(x,t)\rangle = \int A(x_0,0)A(x,t)f(x)dx \tag{5.11}$$

In essence, for all times after time $t = 0$ one is comparing the time dependent property with its initial value, $A(x_0, 0)$. This time dependent property could be described by its expected value according to a time-dependent probability distribution which is also dependent on an initial configuration, $g(x,t;x_0,0)$. In the case of a system obeying Langevin dynamics, this distribution would be the solution to a Fokker-Planck equation [28]. The time dependent property becomes:

$$A(x,t) = \langle A(x,t;x_0,0)\rangle = \int A(x)g(x,t;x_0,0)dx \tag{5.12}$$

Which allows for substitution into eq (5.11):

$$C(t) = \int A(x) \left(\int A(x) g(x, t; x_0, 0) dx \right) f(x) dx \tag{5.13}$$

Focusing on the M sized set of $\{x_s\}$ defined by a set of milestones is a form discretizing the continuous space x. To use this information one must have the ability to calculate a discrete form of eq (5.13). Each initial condition can only be a member of this set and the observable A can only be calculated for members of this set. To approximate the integral of eq (5.12) a summation over all possible milestone configurations must be made, given some initial milestone x_0 where $g(x, t; x_0, 0)$ becomes the time dependent probability distribution for a given initial milestone configuration x_0, according to each milestone, $P_s(t|x_0)$. The discrete approximation for the time-dependent property is then:

$$\int A(x) g(x, x(0), t) dx \approx \sum_{s}^{M} A(x_s) P_s(t|x_0) \Delta x_s \tag{5.14}$$

The external integral must be also approximated as a summation over all milestone configurations. The equilibrium distribution $f(x)$ then becomes the probability at infinite time for each milestone $P_s(\infty)$. Though both summations are over the same set of $\{x_s\}$ configurations, the external summation must be taken over a distinct iterator from the internal summation, in this case i is chosen for this distinction, but note that $\{x_i\} = \{x_s\}$. Thus the

discrete form of the time correlation becomes [87]:

$$C(t) = \sum_{i}^{M} \left(A(x_i) P_i(\infty) \Delta x_i \sum_{s}^{M} A(x_s) P_s(t|x_i()) \Delta x_s \right) \tag{5.15}$$

Provided one can calculate both probability distributions $P_i(\infty)$ and $P_s(t|x_i)$ from milestoning simulations, it is clear from eq (5.15) that the time-correlation function shall be calculable. The equilibrium distribution $P_i(\infty)$ is commonly calculated from the equilibrium fluxes through each milestone in any milestoning simulation. With the determination of the transition distribution functions $K_s^{\pm}(\tau)$, it is possible to construct a Markov chain random walk, the trajectories of which can be used to calculate $P_s(t|x_i)$ [87].

5.2.3 Markov Chain Random Walk / Path Integral Methodology from Milestoning Data

In order to calculate a time dependent probability distribution with a given an initial configuration, $P_s(t|x_i)$, from milestoning data two objects are required. First, each $K_{AB}(\tau)$ must be calculated. From these distributions, the second object must also be calculated, which is a Markov transition matrix $\boldsymbol{K}$ outlined in eq (5.16), where each element is the probability of a transition occurring, based only on the current location. An illustration of such a Markov chain and it's corresponding matrix can be seen in Fig 5.1 and equation 5.16.

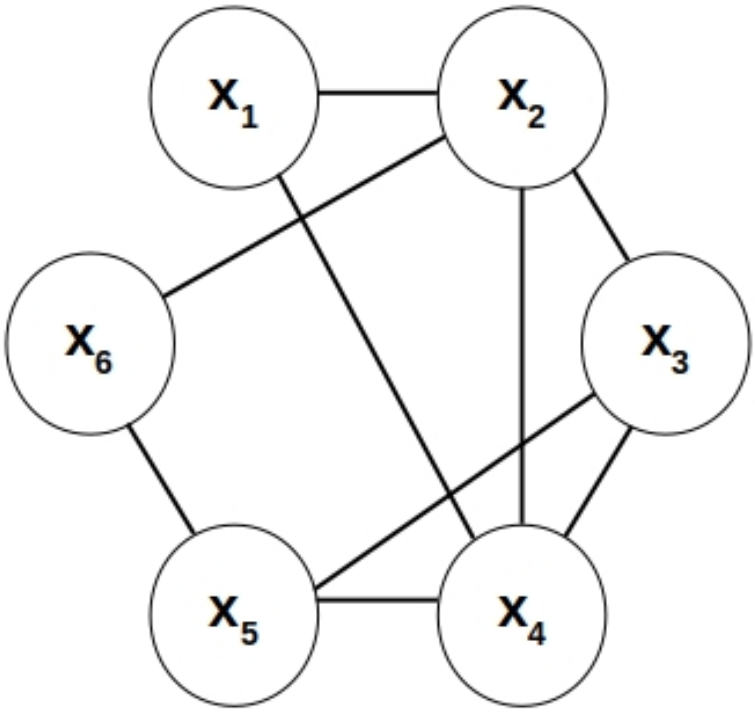

Figure 5.1: A six state Markov chain, with states labeled A-F. Nearest neighbors in configuration space are connected via black lines.

$$\boldsymbol{K} = \begin{bmatrix} 0 & K_{12} & 0 & K_{14} & 0 & 0 \\ K_{21} & 0 & K_{23} & K_{24} & 0 & K_{26} \\ 0 & K_{32} & 0 & K_{34} & K_{35} & 0 \\ K_{41} & K_{42} & K_{43} & 0 & K_{45} & 0 \\ 0 & 0 & K_{53} & K_{54} & 0 & K_{56} \\ 0 & K_{62} & 0 & 0 & K_{65} & 0 \end{bmatrix} \tag{5.16}$$

Such a matrix is constructed under two assumptions, which are built into the milestoning set up. First, only transitions between nearest neighbors in configuration space occur. In the example from Fig 5.1 while at milestone x_1 only transitions to either milestones x_2 and x_4 have non-zero probability in the matrix $\boldsymbol{K}$. Related to this, the second assumption is

that given an infinite amount of time at a particular milestone a transition to a different milestone shall necessarily occur, in other words the diagonal of the $\boldsymbol{K}$ matrix will have zero probability. To calculate the values of each non-zero element, $K_{AB}(\tau)$ are used. The total probability of a transition occurring is simply the integral over all possible times of the transition time distributions, eg $K_{AB} = \int_0^\infty K_{AB}(\tau)d\tau$. With the knowledge that a transition must occur and that it can only be a transition to a neighboring milestone, equation the sum of all M elements in a row in the $\boldsymbol{K}$ matrix must be 1, meaning for the first row of the example matrix in 5.16 where $A = 1$:

$$1 = \sum_{B=1}^{M} \int_0^\infty K_{AB}(\tau)d\tau = K_{12} + K_{14} \tag{5.17}$$

Once this $\boldsymbol{K}$ matrix is constructed, a Markov chain random walk can be easily generated corresponding to appropriate thermodynamics. However, this random walk does not have a constant Δt, as going between one pair of milestones can take drastically longer than another pair. In order to acquire physically relevant timescales of each step, once each direction for the step in a walk is determined, a time for that transition is randomly chosen from the $K_{AB}(\tau)$ that corresponds to said transition [87]. A demonstration of movement from milestone x_1 to x_4 in the example Markov chain in Fig 5.1 is shown in eq (5.18). Once the movement is accepted based on the probability of the transition by virtue of it being more probable than a random number R_1, the time it took for the movement to occur can be determined. Comparing the probability of τ' to a separately generated random number gives a reasonable sampling of times every time this transition occurs. This allows for the fast generation of a large collection of trajectories, each with the possibility of reaching timescales

much further than feasible by typical molecular dynamics methods.

$$if\ R_1[0,1] < K_{14}; \quad r(t+\tau) = x_4 \quad ; \quad if\ R_2[0,1] < K_{14}(\tau'); \quad \tau = \tau' \tag{5.18}$$

Each trajectory can be set to begin at a specific milestone configuration x_i. The collection of all trajectories generated under such a condition can be used to generate an appropriate $P_s(t|x_i)$ through means of path integrals [87]. In practice this becomes histograming the times it takes to arrive at each milestone after starting from a particular one, followed by normalizing over all milestones in each slice of time. With the discrete time dependent distributions in hand time correlations functions can easily be calculated using eq (5.15). Here it shall be demonstrated that this method works well in combination with the WARM method.

5.3 Numerical Demonstration

The inclusion of a force such as the constant used in the WARM method, can give directionality to an otherwise symmetric potential such as a two well system described by $V(x) = x^2(x-2)^2$, shown in fig 5.2, with varying external forces. Here it should be clear that passing through the same points in x in the same amount of time t will have significantly different likelihoods depending on which potential in fig 5.2 on which a trajectory is generated.

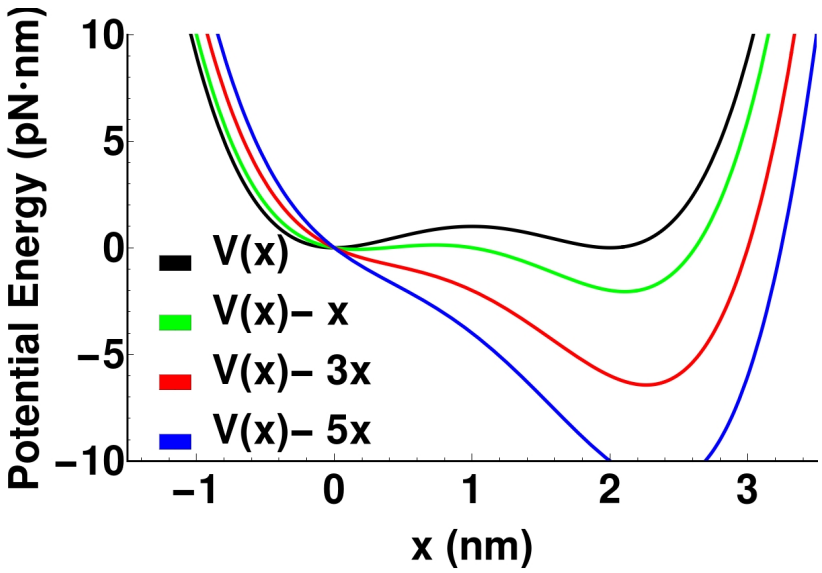

Figure 5.2: The impact of various perturbing forces on a two well potential $V(x) = x^2(x-2)^2 - F_{ext} \cdot x$ in $pN \cdot nm$. Four potentials are shown with applied external forces $F_{ext} = 0$, 1, 3, and 5 pN in black, green, red, and blue respectively.

Limits of Reweighting in the absence of Milestoning

The Reweighting scheme can be directly applied to the calculation of time-correlation functions when calculated via a summation over trajectories $x'(t)$. For the same gap in time for all trajectories generated under constant force f, the values of $x0$ and x_t are adjusted according to the weight of the trajectory at that point in time [27].

$$C(t) = \langle x_0 \cdot x_t \rangle = \frac{\int Dx' x_0 \cdot x_t \frac{W[x'(t)]}{W_{F_{\text{ext}}}[x'(t)]}}{\int Dx' \frac{W[x'(t)]}{W_{F_{\text{ext}}}[x'(t)]}} \tag{5.19}$$

There are two problems that arise from this formalism as it is necessary to generate many long time trajectories. The first problem is that each trajectory must be the full length of the time frame at which a time-correlation function will be calculated, this can be computationally expensive. Longtime effects in a time-correlation function is only obtained if information collected after a large number of small steps are generated in serial for many trajectories.

Second is that the magnitude of external force that will work for reweighting is heavily limited, since the trajectories are so long.

These issues can be demonstrated on a 1-D potential , Langevin dynamics simulations were run on a two-well, quartic potential energy function, $V(x) = x^2(x-2)^2$, with parameters $\beta = 0.5\ (pN \cdot nm)^{-1}$ and $D = 0.20\ nm^2/ps$. Time-correlation functions were calculated using a discrete approximation of eq 5.19 both with and without the use of reweighting. This can be seen in Fig 5.3. The top panel shows the extent the time-correlation function is changed by an external force, when no reweighting is applied. The stronger the constant external force, the more significant the change in the time-correlation function. When the functions are reweighted (bottom of Fig 5.3) it is clear that a stronger external force is more difficult to rewirght accurately to the time-correlation function with no external force. When a perturbation force of $5pN$ is reached, almost no accurate information is obtained for the latter part of the time-correlation function. This limit is related to the fact that the Onsager-Machlup action,S, accumulates over the entire trajectory. The longer a trajectory spends on one potential, the less likely that entire trajectory would be able to be generated on another potential.

The inclusion of milestoning circumvents these issues by using much smaller trajectories. With appropriately placed milestones, such as those shown in Fig 5.4, the number of steps needed to generate a trajectory can be reduced by orders of magnitude. By breaking up the configuration space, a wind force can be applied to encourage the successful transition to neighboring milestones, instead of applied to the entire potential as a whole. These much shorter trajectories, should be easier to reweight as the change in action ΔS accumulated over the course of simulation should be significantly smaller. They will also be much faster to generate as they are stopped once a neighboring milestone is reached. Finally as with all milestoning simulations, the trajectories can be generated in parallel saving on the overall time to sample the potential. The draw back when looking for time-correlation functions, is

that each trajectory individually no longer contains the necessary information to calculated the longtime time-correlation functions. Instead, the transition time distributions can be reweighted, and the Markov chain random walk/ path integral method described above can generate a collection of longtime trajectories, which in turn can be used to determine the time-correlation function.

5.4 Time correlation function from reweighted Langevin dynamics

In the original proposition of the WARM method, only constant force accelerations are used. The reweighting scheme allows for time-dependent external forces to be used as well [27]. Here, in addition to constant force winds, constant velocity pulling type winds will also be considered, for the determination of time-correlation functions with the use of 7 and 9 milestones evenly placed between -1 and 3 along the x axis. Langevin dynamics with parameters $\beta = 0.5\ (pN \cdot nm)^{-1}$ and $D = 0.20\ nm^2/ps$ were used to determine the transition time distributions, $K_{AB}(\tau)$, which were used in the Markov chain random walk/path integral scheme in order to calculate the time-correlation function.

5.4.1 Constant-force wind

The calculation of the time-correlation function for the 1-D two well potential,$V(x) = x^2(x-2)^2$, under various strength constant force winds can be seen in Fig 5.5. Evaluating the ΔS in overdamped conditions with constant force $F_{wind}(x,t) = F_{\text{const}}$, the expression for the

rewighting factor becomes:

$$\frac{W[x(t)]}{W_{F_{wind}}[x(t)]} \approx \sum_i \left(-\frac{\beta}{2}\frac{\Delta x_i}{\Delta t}F_{const} - \frac{\beta}{2m\gamma}\nabla V_i F_{const} + \frac{\beta}{4m\gamma}{F_{const}}^2 \right) \Delta t \tag{5.20}$$

The top panel shows the use of 7 milestones, and it is clear that wind forces of 5,10, and $15pN$ reproduce the true time-correlation function reasonably well. This is a significant improvement over the reweighting of an external force of $5pN$ shown in the bottom of Fig 5.3. Consistent with the findings of Grazioli and Andriciaoei, the bottom panel of Fig 5.5 shows that an increased number of milestones results in improved recapturing of the time-correlation. So much so, that when 9 milestones are used with wind forces of 5,10, and $15pN$ the true time-correlation function is captured almost exactly. While reaching winds of $20pN$ produces a much closer estimate of the true time-correlation function than the same force with 7 milestones.

5.4.2 Constant-velocity wind

With parallels to single molecule pulling experiments, often times a constant velocity pulling force is applied to simulations [82]. The reweighting factor, again in the overdamped limit, now with a constant velocity pulling force of the form: $F_{wind}(x,t) = -k(x - v_{\text{const}} \cdot t)$ becomes:

$$\begin{aligned}\frac{W[x(t)]}{W_{F_{wind}}[x(t)]} \approx \sum_i \Bigg(& -\frac{\beta}{2}\frac{\Delta x_i}{\Delta t}(-k(x - v_{\text{const}} \cdot t)) \\ & -\frac{\beta}{2m\gamma}\nabla V_i(-k(x - v_{\text{const}} \cdot t)) + \frac{\beta}{4m\gamma}(-k(x - v_{\text{const}} \cdot t))^2 \Bigg) \Delta t\end{aligned} \tag{5.21}$$

Simulations were done with 7 and 9 milestones with a constant velocity pulling wind force. Again it is clear that an increased number of milestones reproduces the true time-correlation function with greater accuracy. Pulling speeds v_{cons} of 5,10, and $15\frac{nm}{ps}$ show reasonable estimates of the true time-correlation function with 7 milestones and fairly accurate reproduction of the true time-correlation function with 9 milestones. While $20\frac{nm}{ps}$ becomes fairly inaccurate for 7 milestones, 9 milestones still gives a reasonably close time-correlation function. It should be noted that a constant velocity above $0.5\frac{nm}{ps}$ cannot produce a reasonable time-correlation function from re-weighting without milestones.

5.4.3 Kolmogorov-Smirnov Statistics for Transition Distributions Generated Using WARM

The accuracy of the reweighting scheme in the WARM method is best seen in it's ability to reproduce the transition time distributions, $K_{AB}(\tau)$. The importance of these distributions is mentioned above, as they dictate both the thermodynamics and kinetics of the Markov Chain trajectories, which are used for calculation of the time-correlation function. The Kolmogorov-Smirnov (KS) statistic, D_{KS}, is a measurement of the largest distance between the Cumulative Distribution Function of a pair of distributions [91]. For our purposes it appears as such:

$$D_{KS} = \sup |CDF_{F_{wind}}(\tau) - CDF_{F_{\text{wind}}=0}(\tau)| \tag{5.22}$$

The smaller this statistic is, the more likely a pair of distributions are in fact the same distribution. The KS statistic was calculated for each forward transition time distribution for both constant force and constant velocity pulling simulations with 7 and 9 milestones.

In the top panel of Fig 5.7 one can see that with an increase of constant force wind, the KS statistic increases. This increase is seen to a greater extent in the case of 7 milestones over 9 milestones. Also to note is that, for a given constant force, the largest distances were seen when leaving milestones at or near the bottom of each well in the potential. To a lesser extreme these same qualities can be seen in the bottom panel of Fig 5.7 for the constant velocity pulling experiments.

5.5 Discussion

This work has gone beyond previous uses of the WARM method in combining it with a Markov chain random walk/path integral formalism to calculate accurate time-correlation functions and explored some of the effects of doing this. The limits of the strengths of various types of winds used were assessed. For further improvement in the reweighting scheme, it is possible to consider alternate actions.

In the case of constant force winds, it was demonstrated that the incorporation of milestoning with the reweighting scheme allows for the use of much stronger forces than when no milestones are used. An increase in the number of milestones allows for even stronger forces. The position dependence of the KS statistic suggests that the use of a varying force, in which the same force applied to different milestone transitions, may improve the ability to reproduce the transition time distributions and consequently the time-correlation functions. A trade-off would need to be considered, as a decrease in the strength of the force for milestones near potential energy minima, may produce more accurate transition time distributions, this could lead to longer simulation times overall. In other words, the parts of a simulation needing the most speed up, are the most sensitive to the loss of accuracy from a speed up. It is possible that the inclusion of yet another method, the weighted ensemble method [92], could lead to improved evaluation of these areas. By assigning weight heavier

relative weights to the slowest moving trajectories and then incrementally terminating them, the amount of time spent computing these slower, yet important trajectories is reduced [93]. Alternatively a reweigthing scheme using a different action could also improve the ability to use stronger accelerating forces.

For the constant velocity simulations all that was seen in the case of constant force simulations, with slightly different extents. The constant velocity pulling speeds that can be used in the presence of milestones are an order of magnitude larger than the traditional reweighting scheme. This is likely due to the time dependence of this type of wind force. For short trajectories the minimum of the harmonic pulling force does not have enough time to move far and create such a large effective force. The bulk of the transition time distribution is on the shorter end, and longer transition times are highly unlikley. The success of the constant velocity wind led to an attempt of using an additional aspect of reweighting, a stretching of the time between points in the trajectory, originally proposed by Nummela and Andricioaei [27].

$$\frac{W[x(\alpha \cdot t)]}{W_v[x(t)]} = \alpha^{-\frac{n}{2}} \exp \sum_{j=1}^{n} (\frac{\beta m \gamma}{4\Delta t}(1 - \frac{1}{\alpha})\Delta x_j^2 + \frac{\beta \Delta t}{4m\gamma}(1-\alpha)F_j^2) \tag{5.23}$$

Above is an equation that allows for the reweighting of a trajectory generated under constant velocity pulling to a trajectory that has a constant velocity pulling speed that is α times slower. The intention behind this would be to use very high pulling speeds and recapture dynamics under much slower speeds and ultimately close to have no perturbing force at all. Conceivably orders of magnitudes timescales larger could be sampled if done appropriately. This was limited to a factor of about 2 in the in the absence of milestoning shown in the original proposal of the reweighting scheme [27]. With the incorporation of milestonines in this method, factors of even as small as 1.1 were unable to reproduce accurate transition time

distribution. This is believed to be do to large jump (high Δx) fast trajectories. These types of trajectories would correspond to massive reweighting factors which were several orders of magnitude larger than what was seen in the case of no stretching of time. In other words these trajectories were highly improbable under the faster pulling speed and while being highly probable under the slower pulling speeds. This means that the accelerated conditions would be inefficient for sampling these highly probable trajectories on the slower potential. To realize this goal, an alternative action that describes the relationship between trajectories at longer time gaps between successive steps, may need to be considered.

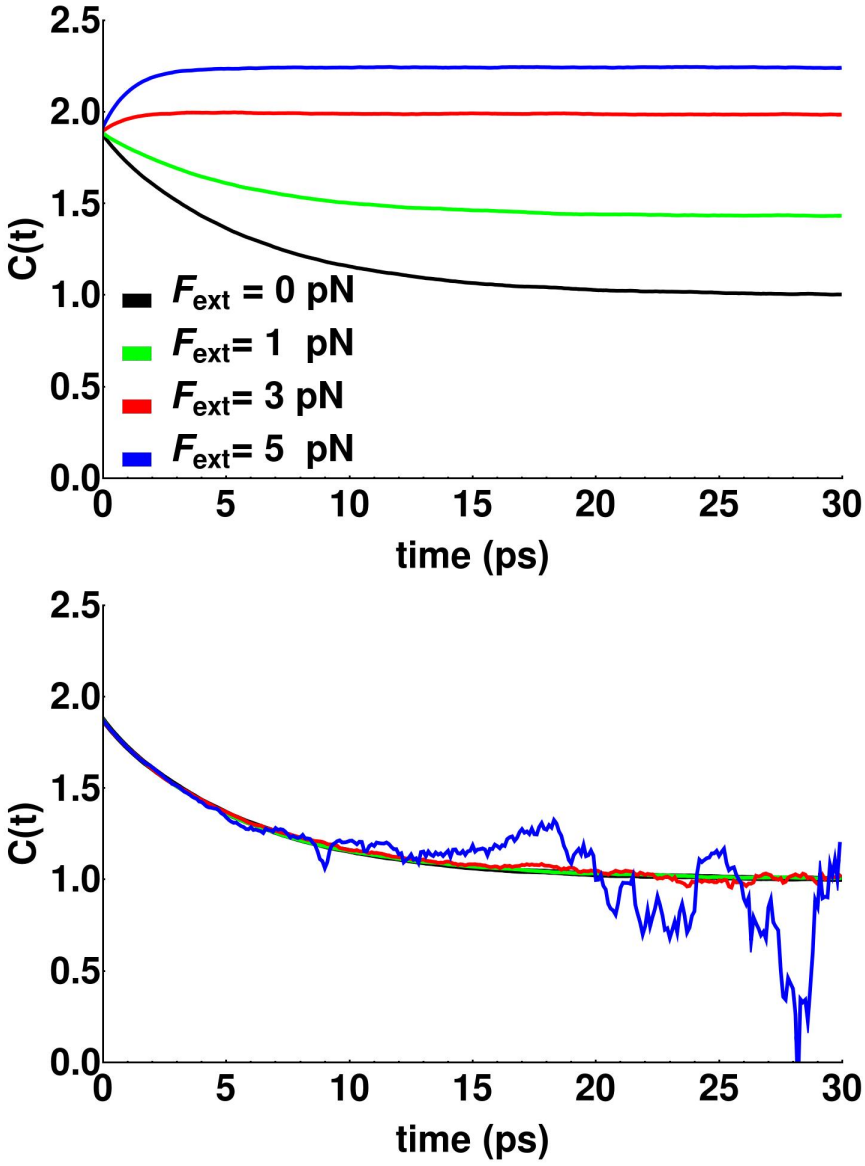

Figure 5.3: Time-correlation functions calculated at various strengths of accelerating force with no re-weighting (top) and re-weighted to zero force (bottom) on the example two well potential. The time correlation functions shown correspond to constant perturbing forces F_{ext}=0,1,3,5 in black, green,red,and blue respectively.

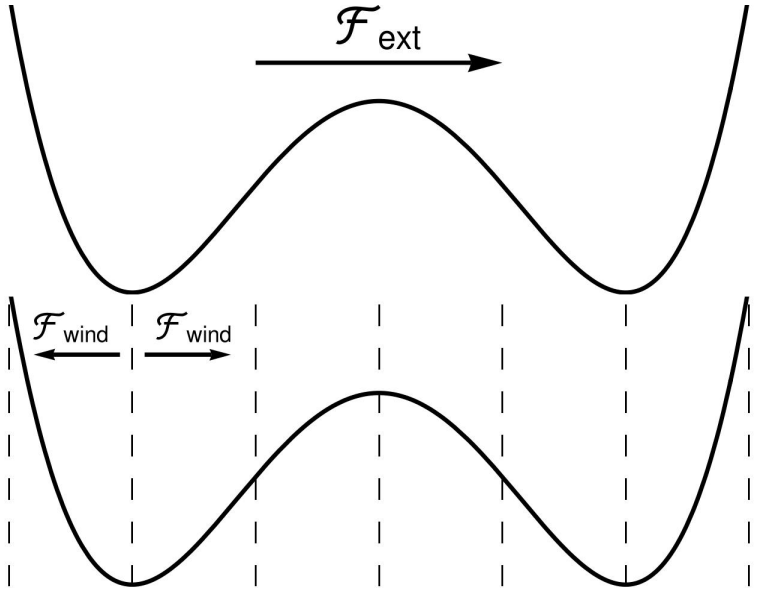

Figure 5.4: (Top) Schematic of applying an external force overall to an entire potential energy surface. (Bottom) Upon placing the milestones (dashed lines), a wind force is pushing trajectories away from each milestone and towards its neighboring milestones, this gives an accelerated directionality that is applied piece-wise to specific segments rather than overall to the entire energy surface.

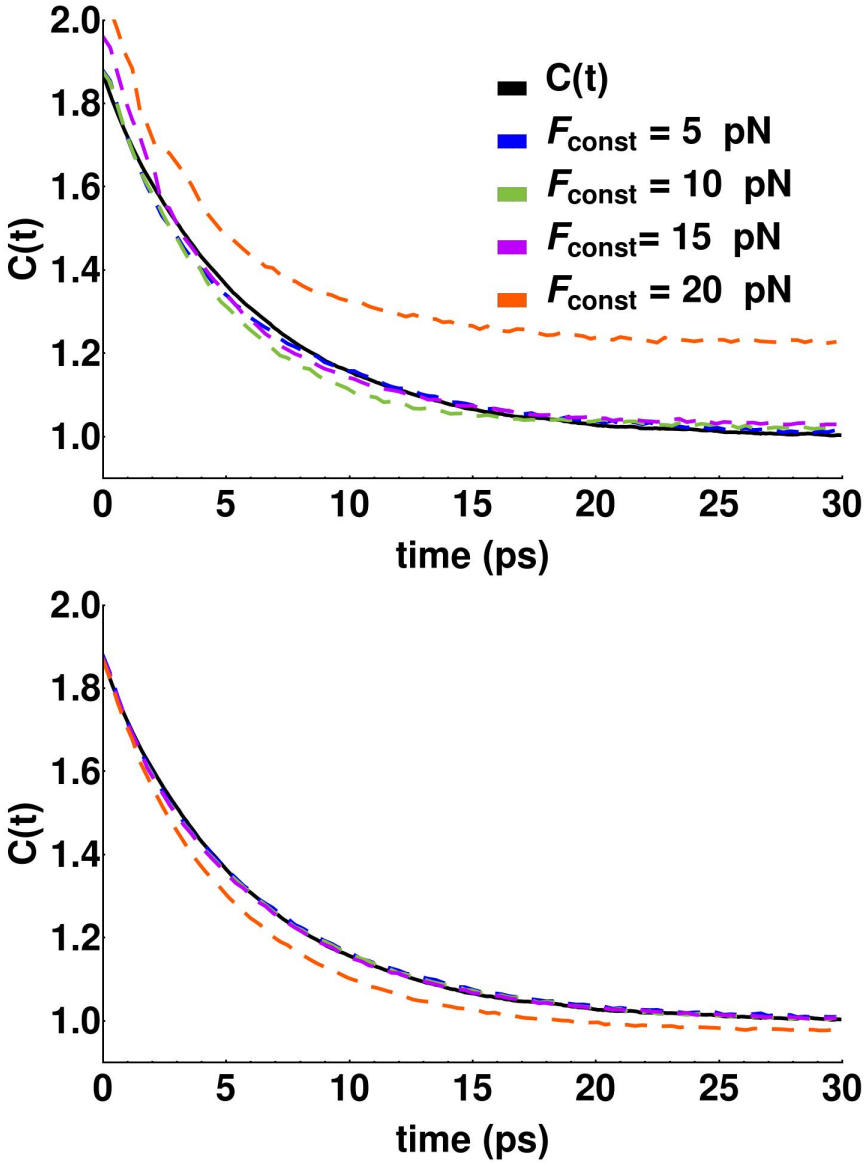

Figure 5.5: Time-correlation functions calculated at various strengths of accelerating with 7(top) and 9(bottom) milestones with constant force winds F_{const}= 5, 10, 15, and 20 pN in blue, green, purple,and orange respectively. In black is the time correlation function calculated without milestoning or accelerating forces

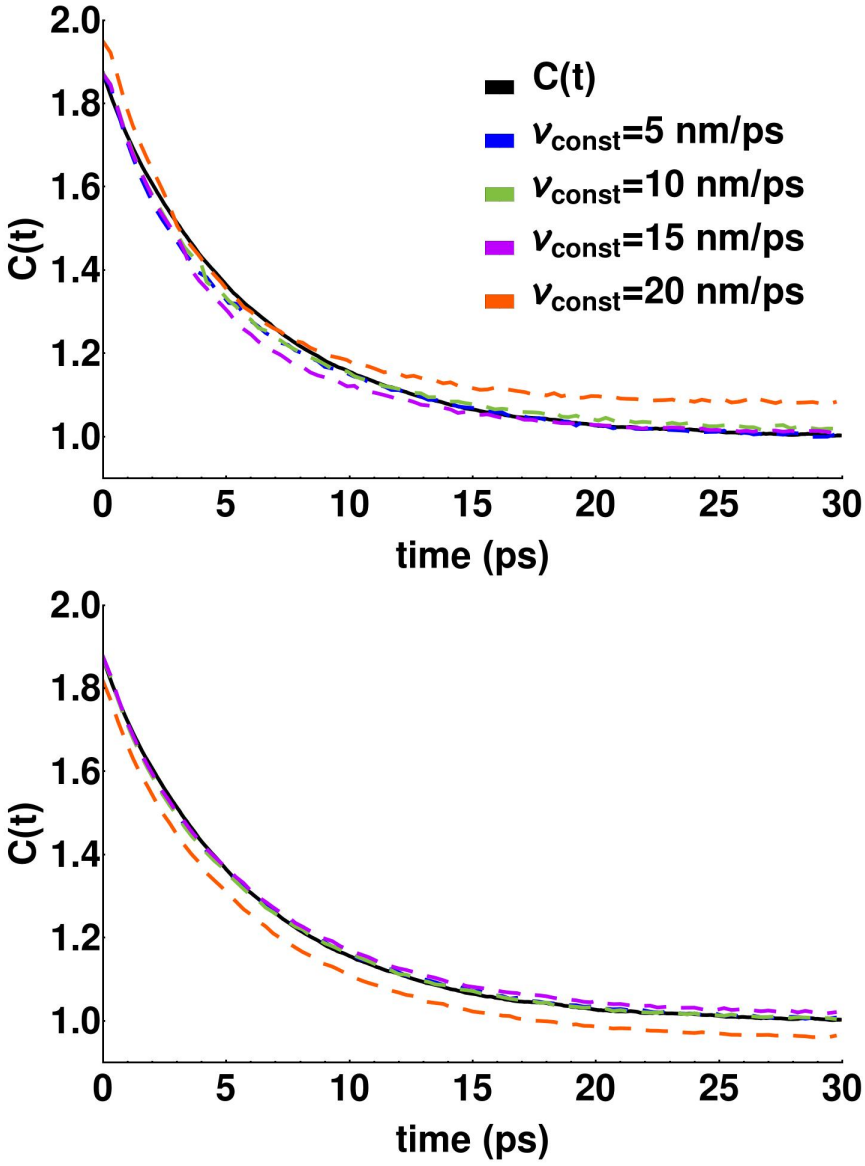

Figure 5.6: Time-correlation functions calculated at various strengths of accelerating with 7 (top) and 9 (bottom) milestones with constant velocity pulling winds with velocities V_{cons}= 5, 10, 15, and 20 nm/ps in blue, green, purple,and orange respectively, and $k = 1/2$. In black is the time correlation function calculated without milestoning or accelerating forces

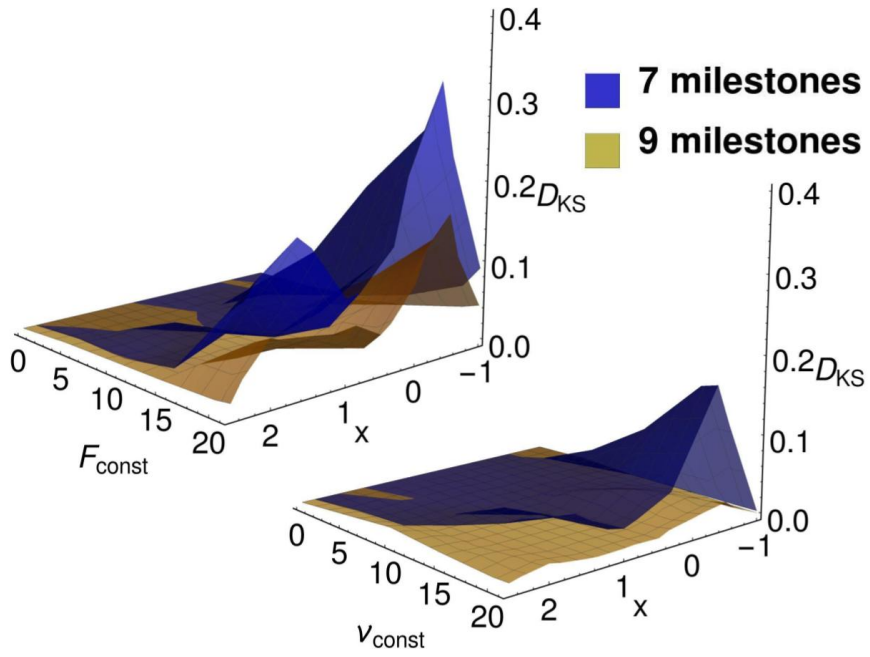

Figure 5.7: The Kolmogorov-Smirnov statistic is plotted for each forward transition distribution by its corresponding starting milestone x value, for simulations with 7 (blue) and 9(gold) milestones. (Top) KS statistics are shown for various constant force winds. (Bottom) KS statistics are shown for various constant velocity pulling wind speeds.

Bibliography

[1] H. S. Bernhardt. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)(a). *Biol. Direct*, 7:23, Jul 2012.

[2] P. G. Higgs and N. Lehman. The RNA World: molecular cooperation at the origins of life. *Nat. Rev. Genet.*, 16(1):7–17, Jan 2015.

[3] J. A. Yeates, C. Hilbe, M. Zwick, M. A. Nowak, and N. Lehman. Dynamics of prebiotic RNA reproduction illuminated by chemical game theory. *Proc. Natl. Acad. Sci. U.S.A.*, 113(18):5030–5035, May 2016.

[4] S. Wasik, N. Szostak, M. Kudla, M. Wachowiak, K. Krawiec, and J. Blazewicz. Detecting life signatures with RNA sequence similarity measures. *J. Theor. Biol.*, 463:110–120, Feb 2019.

[5] Michael L. Goldberg Ann E. Reynolds Lee M. Silver Leland H. Hartwell, Leroy Hood. *Geneitics from Gene to Genomes*. McGraw Hill, 2011.

[6] Omid Khakshoor and Eric T Kool. Chemistry of nucleic acids: impacts in multiple fields. *Chemical Communications*, 47(25):7018–7024, 2011.

[7] Siddhesh D Patil, David G Rhodes, and Diane J Burgess. Dna-based therapeutics and dna delivery systems: a comprehensive review. *The AAPS journal*, 7(1):E61–E77, 2005.

[8] Subgroup 'Assessment of Pathogens Transmissible by Blood' German Advisory Committee Blood (Arbeitskreis Blut). Human immunodeficiency virus (hiv). *Transfus Med Hemother*, 43(3):203–222, May 2016. 27403093[pmid].

[9] Geraldine Aubert and Peter M Lansdorp. Telomeres and aging. *Physiological reviews*, 88(2):557–579, 2008.

[10] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[11] Olivier Fisette, Patrick Lagüe, Stéphane Gagné, and Sébastien Morin. Synergistic applications of md and nmr for the study of biological systems. *BioMed Research International*, 2012, 2012.

[12] Markus Granz, Nicole Erlenbach, Philipp Spindler, Dnyaneshwar B Gophane, Lukas S Stelzl, Snorri Th Sigurdsson, and Thomas F Prisner. Dynamics of nucleic acids at room temperature revealed by pulsed epr spectroscopy. *Angewandte Chemie International Edition*, 57(33):10540–10543, 2018.

[13] David Ban, Colin A Smith, Bert L de Groot, Christian Griesinger, and Donghan Lee. Recent advances in measuring the kinetics of biomolecules by nmr relaxation dispersion spectroscopy. *Archives of biochemistry and biophysics*, 628:81–91, 2017.

[14] Thomas E Cheatham III and David A Case. Twenty-five years of nucleic acid simulations. *Biopolymers*, 99(12):969–977, 2013.

[15] Ignacio Tinoco Jr. Victor A. Bloomfield, Donald M. Crothers. *Nucleic Acids Structures, Properties, and Functions*. University Science Books, 2000.

[16] Alexander Rich. The double helix: a tale of two puckers. *Nature Structural & Molecular Biology*, 10(4):247, 2003.

[17] Karst Hoogsteen. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallographica*, 16(9):907–916, 1963.

[18] E. N. Nikolova, H. Zhou, F. L. Gottardo, H. S. Alvey, I. J. Kimsey, and H. M. Al-Hashimi. A historical account of Hoogsteen base-pairs in duplex DNA. *Biopolymers*, 99(12):955–968, Dec 2013.

[19] H. Zhou, B. J. Hintze, I. J. Kimsey, B. Sathyamoorthy, S. Yang, J. S. Richardson, and H. M. Al-Hashimi. New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. *Nucleic Acids Res.*, 43(7):3420–3433, Apr 2015.

[20] M. Kitayner, H. Rozenberg, R. Rohs, O. Suad, D. Rabinovich, B. Honig, and Z. Shakked. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.*, 17(4):423–429, Apr 2010.

[21] P. A. Rice, S. Yang, K. Mizuuchi, and H. A. Nash. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell*, 87(7):1295–1306, Dec 1996.

[22] G. Ughetto, A. H. Wang, G. J. Quigley, G. A. van der Marel, J. H. van Boom, and A. Rich. A comparison of the structure of echinomycin and triostin A complexed to a DNA fragment. *Nucleic Acids Res.*, 13(7):2305–2323, Apr 1985.

[23] Evgenia N. Nikolova, Eunae Kim, Abigail A. Wise, Patrick J. O/'Brien, Ioan Andricioaei, and Hashim M. Al-Hashimi. Transient hoogsteen base pairs in canonical duplex dna. *Nature*, 470(7335):498–502, Feb 2011.

[24] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor,

R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[25] Jing Huang and Alexander D. MacKerell. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of Computational Chemistry*, 34(25):2135–2145, 2013.

[26] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 1987.

[27] J. Nummela and I. Andricioaei. Exact low-force kinetics from high-force single-molecule unfolding events. *Biophys. J.*, 93:3373–3381, 2007.

[28] R. Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, USA, 2001.

[29] H.A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284 – 304, 1940.

[30] Carey Phelps, Brett Israels, Morgan C Marsh, Peter H von Hippel, and Andrew H Marcus. Using multiorder time-correlation functions (tcfs) to elucidate biomolecular reaction pathways from microsecond single-molecule fluorescence experiments. *The Journal of Physical Chemistry B*, 120(51):13003–13016, 2016.

[31] Junichi Ono, Shoji Takada, and Shinji Saito. Couplings between hierarchical conformational dynamics from multi-time correlation functions and two-dimensional lifetime spectra: Application to adenylate kinase. *The Journal of chemical physics*, 142(21):06B601_1, 2015.

[32] Sayantan Mondal, Saumyak Mukherjee, and Biman Bagchi. Origin of diverse time scales in the protein hydration layer solvation dynamics: A simulation study. *The Journal of chemical physics*, 147(15):154901, 2017.

[33] David E. Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[34] Carlo Laing and Gabriel J Lord, editors. *Stochastic Methods in Neuroscience*. Oxford University Press, September 2009.

[35] Huiqing Zhou, Isaac J. Kimsey, Evgenia N. Nikolova, Bharathwaj Sathyamoorthy, Gianmarc Grazioli, James McSally, Tianyu Bai, Christoph H. Wunderlich, Christoph Kreutz, Ioan Andricioaei, and Hashim M. Al-Hashimi. m1A and m1G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs. *Nat Struct Mol Biol*, advance online publication, Aug 2016. Article.

[36] H. Yang, Y. Zhan, D. Fenn, L. M. Chi, and S. L. Lam. Effect of 1-methyladenine on double-helical DNA structures. *FEBS Lett.*, 582(11):1629–1633, May 2008.

[37] Dan Dominissini, Sigrid Nachtergaele, Sharon Moshitch-Moshkovitz, Eyal Peer, Nitzan Kol, Moshe Shay Ben-Haim, Qing Dai, Ayelet Di Segni, Mali Salmon-Divon, Wesley C. Clark, Guanqun Zheng, Tao Pan, Oz Solomon, Eran Eyal, Vera Hershkovitz, Dali Han, Louis C. Doré, Ninette Amariglio, Gideon Rechavi, and Chuan He. The dynamic n1-methyladenosine methylome in eukaryotic messenger rna. *Nature*, 530(7591):441–446, Feb 2016. Article.

[38] A. G. Palmer. Chemical exchange in biomacromolecules: past, present, and future. *J. Magn. Reson.*, 241:3–17, Apr 2014.

[39] A. J. Baldwin and L. E. Kay. NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.*, 5(11):808–814, Nov 2009.

[40] Y. Xue, D. Kellogg, I. J. Kimsey, B. Sathyamoorthy, Z. W. Stein, M. McBrairty, and H. M. Al-Hashimi. Characterizing RNA Excited States Using NMR Relaxation Dispersion. *Meth. Enzymol.*, 558:39–73, 2015.

[41] H. Yang, Y. Zhan, D. Fenn, L. M. Chi, and S. L. Lam. Effect of 1-methyladenine on double-helical DNA structures. *FEBS Lett.*, 582(11):1629–1633, May 2008.

[42] E. Paci and M. Karplus. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J. Mol. Biol.*, 288(3):441–459, May 1999.

[43] G. Goldsmith, T. Rathinavelan, and N. Yathindra. Selective Preference of Parallel DNA Triplexes Is Due to the Disruption of Hoogsteen Hydrogen Bonds Caused by the Severe Nonisostericity between the G*GC and T*AT Triplets. *PLoS ONE*, 11(3):e0152102, 2016.

[44] J. C. Delaney and J. M. Essigmann. Mutagenesis, genotoxicity, and repair of 1-methyladenine, 3-alkylcytosines, 1-methylguanine, and 3-methylthymine in alkB Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 101(39):14051–14056, Sep 2004.

[45] T. J. Macke and D. A Case. Molecular Modeling of Nucleic Acids. *ACS Symposium Series*, 682(Ch. 24):379–393, 1997.

[46] M. S. Lee, M. Feig, F. R. Salsbury, and C. L. Brooks. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J Comput Chem*, 24(11):1348–1356, Aug 2003.

[47] Y. Xu, K. Vanommeslaeghe, A. Aleksandrov, A. D. MacKerell, and L. Nilsson. Additive CHARMM force field for naturally occurring modified ribonucleotides. *J Comput Chem*, 37(10):896–912, Apr 2016.

[48] Y. Xu, J. McSally, I. Andricioaei, and H. M. Al-Hashimi. Modulation of Hoogsteen dynamics on DNA recognition. *Nat Commun*, 9(1):1473, Apr 2018.

[49] D. E. Gilbert, G. A. van der Marel, J. H. van Boom, and J. Feigon. Unstable Hoogsteen base pairs adjacent to echinomycin binding sites within a DNA duplex. *Proc. Natl. Acad. Sci. U.S.A.*, 86(9):3006–3010, May 1989.

[50] Jose A. Cuesta-Seijo and George M. Sheldrick. Structures of complexes between echinomycin and duplex DNA. *Acta Crystallographica Section D*, 61(4):442–448, Apr 2005.

[51] MF Brana, M Cacho, A Gradillas, B de Pascual-Teresa, and A Ramos. Intercalators as anticancer drugs. *Current pharmaceutical design*, 7(17):1745–1780, 2001.

[52] Hypoxia-Inducible Factor. Echinomycin, a small-molecule inhibitor of. *Cancer Res*, 65:9047–9055, 2005.

[53] MS Jaguar. Schrdinger release 2017-2. *(Schrdinger LLC, New York, NY, 2017)*.

[54] Wenbo Yu, Xibing He, Kenno Vanommeslaeghe, and Alexander D. Jr MacKerell. Extension of the charmm general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J Comput Chem*, 33(31):2451–2468, Dec 2012. 22821581[pmid].

[55] K. Vanommeslaeghe, E. Prabhu Raman, and A. D. MacKerell. Automation of the charmm general force field (cgenff) ii: Assignment of bonded parameters and partial atomic charges. *Journal of Chemical Information and Modeling*, 52(12):3155–3168, Dec 2012.

[56] K. Vanommeslaeghe and A. D. Jr MacKerell. Automation of the charmm general force field (cgenff) i: bond perception and atom typing. *J Chem Inf Model*, 52(12):3144–3154, Dec 2012. 23146088[pmid].

[57] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Jr Mackerell. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *J Comput Chem*, 31(4):671–690, Mar 2010. 19575467[pmid].

[58] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[59] W. G. Hoover. Canonical dynamics—equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.

[60] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[61] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, 1995.

[62] Pål Ø Falnes, Rune F Johansen, and Erling Seeberg. Alkb-mediated oxidative demethylation reverses dna damage in escherichia coli. *Nature*, 419(6903):178, 2002.

[63] Cai-Guang Yang, Chengqi Yi, Erica M Duguid, Christopher T Sullivan, Xing Jian, Phoebe A Rice, and Chuan He. Crystal structures of dna/rna repair enzymes alkb and abh2 bound to dsdna. *Nature*, 452(7190):961, 2008.

[64] Atul Rangadurai, Huiqing Zhou, Dawn K Merriman, Nathalie Meiser, Bei Liu, Honglue Shi, Eric S Szymanski, and Hashim M Al-Hashimi. Why are hoogsteen base pairs energetically disfavored in a-rna compared to b-dna? *Nucleic acids research*, 46(20):11099–11114, 2018.

[65] AA Travers. The structural basis of dna flexibility. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 362(1820):1423–1438, 2004.

[66] Eunae Kim Changwon Yang and Youngshang Pak. Free energy landscape and transition pathways from watsoncrick to hoogsteen base pairing in free duplex dna. *Nucleic Acids Research*, 43(16):7769–7778, Aug 2015.

[67] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11):1859–1865, aug 2008.

[68] Christopher T. Lee, Jeffrey Comer, Conner Herndon, Nelson Leung, Anna Pavlova, Robert V. Swift, Chris Tung, Christopher N. Rowley, Rommie E. Amaro, Christophe Chipot, Yi Wang, and James C. Gumbart. Simulation-Based Approaches for Determining Membrane Permeability of Small Compounds. *Journal of Chemical Information and Modeling*, 56(4):721–733, apr 2016.

[69] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, feb 1996.

[70] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, dec 2005.

[71] Haohao Fu, Hong Zhang, Haochuan Chen, Xueguang Shao, Christophe Chipot, and Wensheng Cai. Zooming across the Free-Energy Landscape: Shaving Barriers, and Flooding Valleys. *The Journal of Physical Chemistry Letters*, 9(16):4738–4745, aug 2018.

[72] Giacomo Fiorin, Michael L. Klein, and Jérôme Hénin. Using collective variables to drive molecular dynamics simulations. *Molecular Physics*, 111(22-23):3345–3362, dec 2013.

[73] Kun Song, Arthur J. Campbell, Christina Bergonzo, Carlos de los Santos, Arthur P. Grollman, and Carlos Simmerling. An Improved Reaction Coordinate for Nucleic Acid Base Flipping Studies. *Journal of Chemical Theory and Computation*, 5(11):3105–3113, nov 2009.

[74] Changwon Yang, Eunae Kim, Manho Lim, and Youngshang Pak. Computational Probing of WatsonCrick/Hoogsteen Breathing in a DNA Duplex Containing N1-Methylated Adenine. *Journal of Chemical Theory and Computation*, 15(1):751–761, jan 2019.

[75] Ioan Andricioaei and Martin Karplus. On the calculation of entropy from covariance matrices of the atomic fluctuations. *The Journal of Chemical Physics*, 115(14):6289–6292, 2001.

[76] Bernard R Brooks, Dušanka Janežič, and Martin Karplus. Harmonic analysis of large systems. i. methodology. *Journal of computational chemistry*, 16(12):1522–1542, 1995.

[77] O. Allner, N. Foloppe, and L. Nilsson. Motions and entropies in proteins as seen in NMR relaxation experiments and molecular dynamics simulations. *J Phys Chem B*, 119(3):1114–1128, Jan 2015.

[78] O. H. S. Ollila, H. A. Heikkinen, and H. Iwai. Rotational Dynamics of Proteins from Spin Relaxation Times and Molecular Dynamics Simulations. *J Phys Chem B*, 122(25):6559–6569, Jun 2018.

[79]

[80] R. Elber. Long-timescale simulation methods. *Current Opinion In Structural Biology*, 15:151–156, 2005.

[81] Serdal Kirmizialtin and Ron Elber. Revisiting and computing reaction coordinates with directional milestoning. *The Journal of Physical Chemistry A*, 115(23):6137–6148, 2011. PMID: 21500798.

[82] B. Isralewitz, J. Baudry, J. Gullingsrud, D. Kosztin, and K. Schulten. Steered molecular dynamics investigations of protein function. *Journal of Molecular Graphics & Modelling*, 19:13–25, 2001.

[83] Ron Elber and Anton Faradjian. Computing time scales from reaction coordinates by milestoning. *Journal of Chemical Physics*, 120(23):10880–9, March 2004.

[84] Anthony M. A. West, Ron Elber, and David Shalloway. Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide. *The Journal of Chemical Physics*, 126(14):–, 2007.

[85] Juan M. Bello-Rivas and Ron Elber. Exact milestoning. *The Journal of Chemical Physics*, 142(9):–, 2015.

[86] G. Grazioli and I. Andricioaei. Advances in milestoning. I. Enhanced sampling via wind-assisted reweighted milestoning (WARM). *J Chem Phys*, 149(8):084103, Aug 2018.

[87] G. Grazioli and I. Andricioaei. Advances in milestoning. II. Calculating time-correlation functions from milestoning using stochastic path integrals. *J Chem Phys*, 149(8):084104, Aug 2018.

[88] L. Onsager and S. Machlup. Fluctuations and irreversible processes. *Physical Review*, 91:1505–1512, 1953.

[89] S. Machlup and L. Onsager. Fluctuations and irreversible process .2. systems with kinetic energy. *Physical Review*, 91:1512–1515, 1953.

[90] H. Kleinert. *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets, 3rd Ed.* World Scientific, 2004.

[91] Frank J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[92] Gary A Huber and Sangtae Kim. Weighted-ensemble brownian dynamics simulations for protein association reactions. *Biophysical journal*, 70(1):97–110, 1996.

[93] E. Suarez, S. Lettieri, M. C. Zwier, C. A. Stringer, S. R. Subramanian, L. T. Chong, and D. M. Zuckerman. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. *J Chem Theory Comput*, 10(7):2658–2667, Jul 2014.