UNIVERSITY OF CALIFORNIA,
IRVINE

Simulating Nucleic Acids from Nanoseconds to Microseconds

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry, with a specialization in Theoretical Chemistry

by

Gavin Dennis Bascom

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

2014

DEDICATION

To my parents, my siblings, and to my love, Lauren.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost I would like to thank Ioan Andricioaei for inviting me to join him in his move to the University Of California, Irvine in the summer of 2006 nearly eight years ago. Since then he has served as an ideal advisor, supporting and guiding my work with patience and expertise. Additionally I would like to thank the current and past Andricioaei group members who have influenced and guided my work through many fruitful conversations. I would particularly like to thank Aaron Frank who worked with me for many years and continues to provide insight regularly. I also need to thank Dr David Busath who introduced me to molecular simulation so many years ago, and patiently set me on the course that led to the completion of this work.

Finally I am indebted to the University of California, Irvine for Graduate Teaching Fellowship and Dissertation Fellowship funding, without which I would not have been able to complete this work. Particularly the hard workers at National Resource for Scientific Computing, the Pittsburgh Supercomputer Center, DeShaw Research, and local UCI supercomputing resources for generous allocations of processing time used to generate and analyze the trajectories reported here.

Appendix A is a reprint of 'Probing Sequence-Specific DNA Flexibility in A-tracts and Pyrimidine-Purine Steps by Nuclear Magnetic Resonance ^{13}C Relaxation and Molecular Dynamics Simulations' published by the American Chemical Society in 2014 with permission of authors Evgenia Nikolova, Hashim-al Hashimi, and Ioan Andricioaei, who directly supervised or carried out the work along with myself. Finally, a quote from an insightful paper by Alan Cooper in 1984 is reproduced in chapter 6 under the fair use clause.

CURRICULUM VITAE

Gavin Dennis Bascom

EDUCATION

Doctor of Philosophy in Chemistry **2014**
University of California Irvine *Irvine, CA*

Bachelor of Science in Biological Sciences **2010**
University of California, Irvine *Irvine, CA*

RESEARCH EXPERIENCE

Graduate Research Fellow **2010–2014**
University Of California, Irvine *Irvine, CA*

Junior Research Specialist **2007–2010**
University of California, Irvine *Irvine, CA*

Undergraduate Lab Assistant **2006–2008**
Brigham Young University *Provo, UT*

Summer Research Opportunity Fellow **2006**
University Of Michigan *Ann Arbor, MI*

TEACHING EXPERIENCE

Graduate Teaching Fellow **2010-2014**
University Of California, Irvine *Irvine, CA*

Undergraduate Teaching Assistant **2007**
Brigham Young University *Provo, UT*

REFEREED JOURNAL PUBLICATIONS

A General Method for Constructing Atomic-Resolution RNA Ensembles using NMR Residual Dipolar Couplings: The Basis for Interhelical Motions Revealed. **2013**
Journal Of American Chemical Society

Probing Sequence-specific DNA Flexibility in A-tracts and Pyrimidine-purine Steps by NMR 13C Relaxation and MD Simulations. **2012**
Biochemistry

REFEREED CONFERENCE PUBLICATIONS

Ab-Initio Protein Folding and Small Molecule Dissociation by Potential Energy Based Biased Molecular Dynamics **March 2011**
Biophysical Society Cell Press

Simulated Single-Molecule FRET Trajectories: A Comparative Analysis Between Three Telomeric G-quadruplexes **March 2009**
Biophysical Society Cell Press

ABSTRACT OF THE DISSERTATION

Simulating Nucleic Acids from Nanoseconds to Microseconds

By

Gavin Dennis Bascom

Doctor of Philosophy in Chemistry, with a specialization in Theoretical Chemistry

University of California, Irvine, 2014

Professor Ioan Andricioaei, Chair

Nucleic Acids, despite being among the most important macromolecules involved in biological life, remain poorly understood in terms of atomistic resolution dynamics at biologically relevant timescales. Due to recent advances in computational power and high resolution structure elucidation we are able to investigate the dynamics of four important nucleic acid structures, namely $5-CGAT_6GGC-3$, $5-CGCGAT_4GGC-3$, $5-GCATCGAT_2GGC-3$ (referred to as A_6, A_4, and A_2 DNA respectively) and the TAR HIV-1 RNA molecule on the nanosecond and microsecond timescales. The trajectories are numerically characterized by the NMR order parameter S^2 which provides a quantitative measure of motion comparable to experiment, from nanosecond based ensembles in the case of A_6, A_4, and A_2DNA, and microsecond based ensembles for A_6DNA and TAR RNA. Specifically, this comparison suggests that while DNA exhibits saturated motions at the nanosecond-microsecond timescale, HIV-1 TAR RNA exhibits motions seemingly correlated across timescales suggesting it has not yet full saturated motion at the microsecond timescale. Effects of internally correlated, temporally correlated, and diffusively continual motions for nucleic acids are discussed. Finally, the potential of mean force (PMF) of one such smooth transition, the A $\Leftrightarrow$ B transition, is reported in the presence of a Single Walled Carbon Nanotube (SWNT) for DNA of GC and AT rich sequences.

Chapter 1

Introduction

1.1 Nucleic Acids in a Larger Context

The history of life on earth begins most likely not with DNA, but with RNA [1, 2]. DNA, despite Francis Crick's dogmatic emphasis, is far from the "center" of physically mechanistic properties that we can discern; in fact it is one of the most stable macromolecular entities we know about [3]. Interestingly, it seems it is this very stability that makes DNA such a prime candidate for long term macromolecular information storage, and it is this unique property that likely allows for it to reside at the "center" of our general concept of life as opposed to the execution of various cellular functions [4]. RNA and proteins carry out this larger vision and therefore require much more diversity in shape, size, and morphology. The question then of DNA centrality becomes something of conceptual convenience as opposed to organizational understanding being dictated by physical parameters or functional focus, which this is well orchestrated by how seldom we here about the actual discovery of nucleic acids in 1868 by Friedrich Miescher [5]. We find this in stark juxtaposition with the more ubiquitously discussed contribution by Watson and Crick [6, 4], by which *the function* of DNA came to

reside at the "center" of our concept of life, not it's shape or morphology necessarily. The idea is tantalizingly simple; DNA encodes genetic information and it is by some interaction between this genetic information and the physical nature of chemical/energetic movements that we are bestowed with the gift of life, comprehension, control, and the passing on of our traits to offspring. While this simple observation was only derived from careful analysis and deduction of systemic interactions, it is hard to argue that the single most important factor came by elucidation of structure. Once the structure of DNA was available, it was clear that it could store and transmit information, a simple fact that was then and is still now held as a sort of holy grail for biologists and armchair philosophers alike. Regardless of the amount of information that can be gathered about a thing, it seems a deep conceptual basis is difficult for us humans to construct until one can start to see how that thing is shaped, and how that shape changes with time. Only once Watson and Crick could see how information could be stored and kept, and identify that information with genetic transmission of traits to offspring, were the masses ready to accept the central dogmatism of the process.

Furthermore, appears that to the best of our current knowledge the wonderfully convoluted interaction between information and structure that gives rise to cellular function is ultimately governed by the two simple properties position and time. The vast array of diversity of life moves around particularly fast, and if those movements cease the organism ceases to function. It must be true then, that movement at the atomic scale has some sort of *fundamental* importance to our continued existence. In this strange (and relatively young) vision of life it would appear that a large portion of the attributes that we identify with "us" are governed entirely by the wholly uncontrollable jiggles (positions changing with time) of these molecules, but more specifically the history of jiggles, and how the various species interconvert during these seemingly random sets of movements. Subsequently it becomes almost natural to assume, or even assert, that the quest to understand how life works at the cellular level, the need to increase the resolution of our tools, to sharpen our conceptual understanding, to reverse maladies and cure sickness, and enhance capabilities through

technological intervention is essentially a question of figuring out, at the most fundamental level, *where* and *when* things happen, or *where* and *when* they have happened in the past, and then inculcating that knowledge into some form of intervention at the cellular level.

It is hardly a surprise, then, to consider that most disciplined approaches involving nucleic acids are attempting, in some way, shape, or form, to detect the movement, the *where* and *when* of the most fundamental molecules of life, and it is unfortunate for us that resolving a fully accurate, fully atomistic non-equilibrium trajectory of fundamental particles is far from simple. Clearly it stands to reason that easy detection of the details of the movement of large numbers of molecules would make our conceptual framework about life look very different than it does currently. Upon further reflection however, we may realize that it is not the *actual* history of these atomic movements that we primarily pay attention to, (indeed we can barely elucidate them in any real detail or certainty) but instead the history of our *knowledge* of generalized, hypothetical DNA and RNA molecules that primarily governs our investigation of said molecules, and we have only recently come to be able to scratch the surface of the complicated rules that govern elucidation of the *actual* molecular "jiggles" [7, 8]. In fact, upon further consideration we have also just begun to understand what we *cannot* know regarding the trajectories of these molecules, and it is only by careful consideration of these factors that we can be sure we are moving towards the "center" of increased understanding of life, and increased interventional capabilities in the cell in the most fundamental way.

And so let us first note that the central dogma of molecular biology, despite being seductively simple, is an incomplete picture of the processes that allow us to receive and make use of genetic information. Since its inception the genetic code, or the knowledge that simple purine/pyrimidine base pairing can give rise to efficient information storage, has served as a nexus for thought regarding the cellular processes to young investigators and seasoned scientists alike. The details of transcription, replication, and translation, however, clearly

spell out that a sequence of code alone is not enough to create a living organism [9]. Nucleic acids are dynamic entities, and anything with such dynamic character inevitably breeds (or perhaps is the product of) complicated interactions with its environment. In short there exists a "second tier" genetic code, a set of necessary conditions that allow the genome to be flagged, altered, folded, cut, twisted, and deformed in almost impossible ways in order to maintain a healthy living cell [10]. RNA, while even more diverse in function, similarly exhibits large diffusive motions that govern many of the fundamental cellular processes either directly or indirectly, including reading DNA for the production of proteins and information storage of its own flavor [11].

Perhaps then we should be weary (or simply cautious) of the possibility of over application of both central dogmatism and the need to assign exact certainties to the *where* and *when* of biological macromolecules. This alone cannot govern our understanding of the processes; indeed it has been stated with great clarity that there is a *thermodynamic* correlate to the famous uncertainty principle (a clean and neat example of the power of pointing out what we cannot know) involving the exact motions that we can assign to elucidated structures [12]. Furthermore, any attempt in describing how to represent those motions *exactly* will differ wildly in conception depending on whether you ask a quantum mechanic, [13], a statistical mechanic [14], or a classical mechanic [15]. This is an important point to keep in mind; regardless of the accuracy and sophistication of our thermodynamic techniques (non-equilibrium or otherwise) we are still obliged to stomach some uncertainty about the fundamental trajectories involved in molecular motion, and no matter the accuracy and sophistication of our simulations, they only represent one *possible* set of motions available to the molecule. We have entered an age in which the unification of both single molecule observable attributes and ensemble averaged thermodynamic variables is clearly the most powerful avenue available to us as structural biophysicists; we must use simulation to reasonably deduce possible trajectories, and we must use thermodynamics to reasonably interpret those trajectories. It is only by the careful application of both of these systems of inquiry

that we arrive at a novel and exciting understanding; *certain motions of flexible nucleic acids are inherently unresolvable, even when the motions of those molecules are explicitly solved in some single molecule sense.* In short, there are many possible trajectories which would yield identical thermodynamic averages, and in order to solve a coherent mathematical picture of any large biological system analytically (or even approximate it with any real accuracy) either thermodynamically or classically or quantum mechanically, we must settle with using *correlated* sets of continuous *probable* motions to fully understand the behavior of these molecules in any meaningful way. This is not a new idea, but rather an idea that has been both suggested and brought to implementation at length in the arena of the protein dynamics [16, 12, 17, 18], which, when folded, are much less flexible than nucleic acids by nature [19].

But first, we must begin our study of this interesting concept as applied to nucleic acid dynamics with the basics of nucleic acid structure, after which we will briefly discuss the history of Newtonian based molecular dynamics methods and experimental techniques which help elucidate said structures. Finally, we will define our larger scope and aim for this study, neatly laying out the road by which we will arrive at carefully reasoned conclusions regarding molecular motion at the atomic level.

1.2 Nucleic Acid Structure

1.2.1 DNA Structure/Motion Basics

Canonical B-form DNA is comprised of two intertwined strands that wrap each other tightly, stabilized by Watson-Crick base pairing and stacking potentials (see figure 1.1). There can be defined a large set of parameters which characterize the motions available to backbone and base atoms alike. Base parameters that are commonly discussed involve the relative

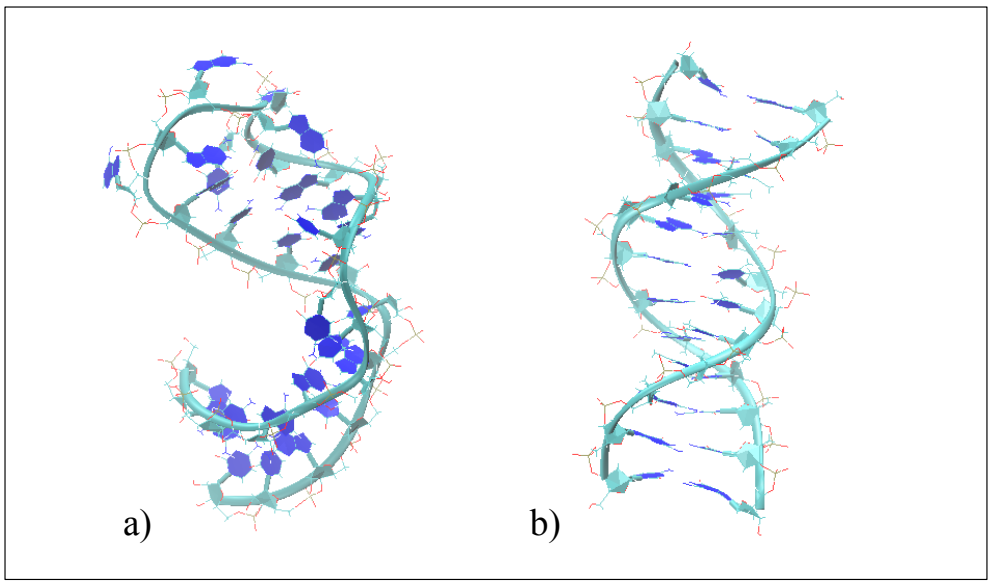

Figure 1.1: Two examples of the major nucleic acids to be considered for this study, a) the TAR HIV-1 RNA sequence and b) $5-CGAT_6GGC-3$, (referred to as A6DNA). Structures shown are starting configurations either built in house or taken from the protein data bank (access code 1ANR) [20].

motions of bases to each other, which are nicely summarized in [21]. In short buckle, shear, stagger, stretch, propeller, slide, roll (etc) describe inter-base distances and angles relative to each other and the backbone. They are among many good indicators of overall DNA form, which generally can adopt one of three broad classifications termed A, B and Z form DNA. In liquids at physiological conditions DNA tends to adopt B form while RNA tends to adopt A form and stretched or pulled DNA can sometimes adopt Z forms [22, 23, 24]. Backbone parameters include sugar puckers, or relative orientations of atoms in the 5-ring sugar that links base moieties to phosphate backbone atoms [25] and overall rigidity, which involves the mechanical ability to deform the double-stranded helix around itself or around histones [26, 8, 27].

What is known about unbound B form dsDNA dynamics involves a rich and vibrant field of study, revealing motions which are quite intricate and subtle, although dsDNA generally doesn't deviate too far from its initial B form without quite a bit of force or energy applied. The more "canonical" motions however are more or less centered around B form, A form, or Z form duplexes in physiological conditions [23] with the exception of some more exotic folds which are also sequence and environment dependent (for some examples see triplexes or quadruplexes [28, 29]). Base stacking and pairing forces seem to have the added benefit of an entropic penalty for base unzipping [30], and forces required for said unzipping have been shown to be quite intricate and high. Initially, stretching away from B form dsDNA into whats called "S form" (or stretched form) by AFM requires around 65pN of force, after which about 150pN of force is required to fully unzip the DNA [31]. Despite these high penalties for dsDNA denaturation, there does exist temperature driven local "melting," which occurs at medium to high temperatures and possibly at lower temperatures with lower frequency and smaller bubble size [32, 30]. Finally it should be noted that both DNA structure and dynamics are at least partly dependent on sequence, with AT rich sequences having propensity to be locked into B form DNA [33] with high propeller twist and a progressive narrowing of the minor groove from the 5' to 3' direction. Furthermore A-tract DNA exhibits anharmonic

torsional stiffness, which can be elucidated by small torques on the molecule [34]. Further effects of specific types of sequence dependence is discussed further in the introductions to chapter 2 and 3.

1.2.2 RNA Structure/Motion Basics

RNA, on the other hand, has a much more diverse family of structures, despite demonstrating relatively stable double stranded helical regions which function more or less like B DNA [35]. At the ends of these regions, however, we find smaller, floppier regions with characteristic bulges, loops, hairpins, mismatches and other motifs that display a wide variety of motion, structure, and function (see figure 1.1 for an example of bulge and loop region on the HIV-1 TAR RNA we will be primarily focused on for this document). The RNA backbone differs from the DNA backbone only by one small addition of an oxygen atom at the C2' position [7, 36], and this seems to be enough to make A form helices lower energy than B form helices in solution at physiological pH [23]. The structure of RNA also has a large dependence on sequence, however, and mismatches and hairpin loops mentioned above seem to be a regular part of sequence conservation in RNA motifs [37, 38]. These sequence dependent motifs tend to fold in a predictable way [39], and also typically demonstrate semi-predictable tertiary motifs. For example, hairpin loops are extremely fluid and act as nucleation sights for folding [40, 41], while bulge mismatches can act as global "hinges" around which the duplex folds [42], but these cover only a few of the diverse functions that small RNA molecules exhibit. Actually, due to its wide variety of fluid like secondary structures, RNA is able to execute and is involved in a broad class of cellular processes, both friendly [43] and not so friendly [44] to human cells. RNA holds and transfers information, acts as a sieve to extraneous information in the genome, effects transcription, translation, and acts as an enzyme to name just a few more functions [45, 37, 46].

Furthermore, relatively little is still understood about the dynamics of the floppier parts of these structures [35], but almost all motions have been shown to be important for both small molecule ligand binding [47], viral life cycles [44] and genetic manipulations [48]. Difficulty in designing small molecule RNA ligands provides only one of many poignant examples of the uncertainty that still surrounds the physical movements of RNA molecules both internally and throughout the cell [48, 39, 49].

1.2.3 Experimental Techniques for Nucleic Acid Structure Elucidation

While a broad range of experimental techniques can capture information allowing us to infer the details discussed above, two major techniques will be focused on in this document. We can resolve static pictures of macromolecular structure through x-ray crystallography, which involves analyzing diffraction patterns of scattered x-rays through solid crystallized samples of interest [50]. This technique, while enjoying a firm place at the heart of structural biology, also boasts the elucidation of some of the most important macromolecules, such as nucleic acid structure and proteins [6]. While a complete discussion of the techniques is not necessary for our current study, we should at least recognize that very little of the current body of knowledge could have existed without the advent of structural biology through X-ray crystallography [51].

In order to look at molecular scale objects while moving, however, one could use Nuclear Magnetic Resonance, which involves small variations in magnetic fields to evaluate structure and dynamics [52]. In other words by looking at not only the *light* spectrum (such as microscopes, spectral analyses, or the aforementioned X-ray diffraction patterns) but also the *magnetic* spectrum of molecules, we can elucidate a picture of not just structure, but equilibrium ensembles of molecular motions as well. Typically, taking an NMR measurement

requires the inculcation of an isotopic version of nitrogen or carbon into a specific residue on the macromolecule, which will then couple to an external magnetic field. This results in the controlled manipulation and measurement of specifically placed $^{13}C-^{1}H$ or $^{14}N-^{1}H$ bond vectors typically referred to as $\hat{\mu}(t)$. Among many different measurements available to NMR spectroscopists, two relevant order parameters include the S^2 measurement, and Residual Dipolar Couplings (RDCs). The S^2 order parameter measures the correlation of motion for rapid internal fluctuations of a bond vector position, and can be attained by relaxation experiments (further discussed in chapter 4) [53]. A simple method for computing this metric from molecular dynamics simulations involves the parameterization of the cartesian coordinate bond vector autocorrelation function:

$$C(t) = \langle P_2[\hat{\mu}(0) \cdot \hat{\mu}(t)] \rangle \tag{1.1}$$

as the sum of two exponentials (a simple but robust approximation which has yet to be improved upon significantly, called the model free approach [16]). Lipari and Szabo first described this approach in detail in 1982 [54] and the following formulation (used in this study) first appears in a paper by Clore and Szabo in 1990 [53] where S^2 is the generalized order parameter which describes the amplitude of the function and τ is the exponential decay time for fast and slow motions:

$$C(t) = S^2 + (1 - S_f^2)e^{-t/\tau_f} + (S_f^2 - S^2)e^{-t/\tau_s} \tag{1.2}$$

Where $S^2 = S_f^2 S_s^2$ and subscripts f and s refer to fast and slow motions respectively, which are assumed to be uncorrelated (this is an important point to which we will return in some detail in chapter 4). It should also be noted that further parameterization could be applied as

more sets of motions are deemed separable, resulting in different forms with more exponential terms [55].

Residual Dipolar Couplings (RDCs) are not explicitly time dependent (but rather time averaged) but dependent on the angle the given bond vector $\hat{\mu}(t)$ makes with an external magnetic field, defined here as θ. They can be extracted from MD simulations easily according to the form [56]:

$$D_{ij} = \frac{\mu_o \gamma_i \gamma_j h}{8\pi^3 \langle r_{ij}^3 \rangle} \left\langle \frac{3\cos^2(\theta) - 1}{2} \right\rangle \tag{1.3}$$

Where ij refers to atoms i and j, γ_i and γ_j refers to the gyromagnetic ratio of the ith or jth atom respectively, r_{ij} is the inter-atom distance or bond vector length, angular brackets denote time averaging and θ is the angle between the bond vector and an outside magnetic field. We cannot, however, directly observe the movements of molecules in real time at the molecular level without the help of computer simulation.

1.3 Simulating Trajectories by Molecular Dynamics

In order to resolve a fully atomistic picture of the movements of atoms, we must use an approach that combines all of the knowledge we have been able to acquire over the history of nucleic acid structure into a mechanistic simulation, only recently implementable due to increasing power of computational algorithms. By reproducing the trajectories of molecules *in silico*, we can begin to elucidate not just structure, but some of the details involved in the inter-conversion between various structures.

1.3.1 Integrating Newtonian Equations of Motion and Force Fields

A fundamental and yet synergistic divide exists between quantum mechanical approaches and classical, or Newtonian approaches to the equations of motion. The latter provides a robust formalism around which we can treat all classical objects while the former allows us (among other things) to calculate important parameters necessary for accurate Newtonian based simulation, which can then be reinforced by experimentally derived parameters. Newtonian mechanics has its nexus in the simple observation that the force applied to an object f_i scales with mass m_i and the second derivative of the three cartesian position coordinates $\ddot{\mathbf{r}}_i$. In the more compact form we write the well known equation as:

$$\mathbf{f}_i = m_i \ddot{\mathbf{r}}_i \tag{1.4}$$

and it should be noted that this equation must be solved specifically for each atom i separately, for each of the three cartesian coordinates every time an atom moves, or every time step computed. In order to do so however, we need to do further define force as the gradient of potential energy, or $\mathbf{f}_i = -\nabla_{\mathbf{r}\,i}\mathbf{V}$. While a deeper discussion of potential energy is beyond the scope of this document, it is sufficient to say that potential energy can be derived from a variety of sources, and parameterized for bio-macromolecules in a semi-empirical way to the potential energy function :

$$\mathbf{V} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \tag{1.5}$$

$$+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] + \sum_{i<j} \left[\frac{q_i q_j}{\epsilon R_{ij}}\right] + \sum_{hbonds} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right] \tag{1.6}$$

which is typical of molecular simulation currently. Variables r_{eq} and θ_{eq} refer to equilibrium bond lengths and equilibrium angles respectively while K_r and K_θ are bond and angular force constants, V_n is barrier height around bond rotations, while A, B C, and D are parameterized constants (which differ from atom to atom) and finally q_i or q_j is partial charge per atom. As one can see, the potential is decomposed into several non-interacting parts, some of which are based on assumptions or approximations, and all of which need to be extensively parameterized giving rise to the concept of a force field [57]. Specifically angles, dihedrals and bonds are treated with simple spring like potentials [58], with a different version for periodic and non-periodic (improper) dihedrals involving trigonometric functions where appropriate [59]. A so called non-bonded potential, which involves atoms that are imparting forces on each other at a distance, is comprised of a coulombic term and a lennard jones 12-6 term (the 12-6 refers to the powers of the repulsive and attractive distance terms respectively) appropriately parameterized for the condensed phase [60, 61]. Hydrogen bonds are treated similarly with a 12-10 term, although current force fields often include an angle dependent term as well [59].

Parameterizing the potential energy function, while difficult to say the least, only provides us with a general form to substitute into Newtons equations of motion, which must be integrated numerically for any non-idealized system. Many forms of integrators exist, but some of the most commonly used are finite-difference integration schemes [62] that assume that there must be a discontinuous calculation of particle position. The verlet algorithm, for example, was named after french mathematician Loup Verlet, although in a brilliant paper in 1993 he and D Levesque note that the algorithm has been in use since as early as 1791 [63]. This apparently very old numerical algorithm defines position as a function of the *change* in time $(t + \Delta t)$ as opposed to time directly, given by:

$$\mathbf{r}_i(t + \Delta t) = 2\mathbf{r}_i(t)_i - \mathbf{r}_i(t - \Delta t) + \Delta t^2 \frac{\mathbf{f}_i(t)}{m_i} \tag{1.7}$$

where particle position $\mathbf{r}_i$ is advanced at each integration step from the previous two steps, advancing each time step numerically, while an exact solution to the equation of motion is never found. This has the added benefit that temperature, pressure, volume, and other simulation parameters can be calculated and modified directly as the simulation is carried out, without changing the potential energy function. Additionally, we can use this added ability to construct ensembles (in the statistical mechanical sense) aligning with a large body of knowledge surrounding statistical analysis of energy and motion as they are defined by thermodynamic variables [14]. It should be noted that many variations on this theme have been introduced successfully, in an attempt to model not only macromolecules, but simple and complex liquids alike.

Although the above described force field, which has now been in active development for over thirty years, has received criticism when compared to potentials of pure chemicals or other condensed phase analytical models, the accuracy of the CHARMM like potential as shown above with extensively parameterized coefficients has proven itself over and over again to the scientific community. Among its successes include small molecule ligand binding to molecular targets, elucidating structure and function of various processes, and predicting thermodynamic properties which correlate to kinetics, just to name a few [18, 64]. Of course its record is not without blemish, and the question of force field accuracy has been an ongoing discussion since its inception. The two force fields most widely used for nucleic acids, AMBER and CHARMM, have required several major updates over the last few decades [65, 66, 67]. While force field generation and refinement is an active field which draws from quantum mechanics and experimental procedures alike, it seems that many of the nucleic acid

force field problems arise from the charged nature of these molecules. Recent updates have specifically targeted updating the difficult dihedral angle force constant parameter [65, 68], which provides considerable difficulty in parameterization for proteins and nucleic acids alike. While essential to the folding and maintenance of secondary structure by backbone movement, dihedral angles are one of the few parameters that cannot be measured directly by experiment, rendering them an ongoing question in the further development of molecular simulation.

1.3.2 Treating Non-bonded Interactions

Some extra time should be devoted to how we treat long-range electrostatic interactions, seeing as a physiologically relevant environment often dictates a more or less infinite space in which the species live, and furthermore treating charged species such as nucleic acids requires that the coulombic distance term be treated with utmost care in order to avoid artifacts in simulations [69]. Upon careful consideration one notices that evaluating the aforementioned electrostatics requires solving N(N-1) sets of coupled equations per integration step, a number which becomes very large and cumbersome for large systems (in other words the memory required scales with the square of the number of atoms in the system, written $\emptyset(N^2)$).

One simple feature that has been implemented with great success to avoid this blow up of needed memory is the particle mesh Ewald (PME) summation [70], in which long range and short range interactions are treated separately. In this formulation the short range electrostatic interaction potential is given by:

$$E_{sr} = \sum_{ij} \varphi_s r(\mathbf{r}_i - \mathbf{r}_j) \tag{1.8}$$

and the total long range electrostatic interaction is given by

$$E_{lr} = \sum_{ij} \tilde{\Phi}_{lr}(\mathbf{k})|\tilde{\rho}(\mathbf{k})| \tag{1.9}$$

Where $\mathbf{k}$ is a lattice of points in mesh, $\tilde{\Phi}_{lr}(\mathbf{k})$ is the Fourier transform of the potential, and $\tilde{\rho}(\mathbf{k})$ is the Fourier transform of the charge density. By virtue of the fact that both summations converge quickly in Fourier space they can be evaluated in a very memory efficient way by use of simple fast Fourier transform algorithms (FFT), along with reducing the number of equations needed to solve for electrostatics drastically. The resulting calculation of long range electrostatics requires symmetry in long range space, which is remedied by the implementation of periodic boundary conditions, in which a periodic crystal is defined for atoms which wander outside of the simulation box [71]. The resulting calculation scales with $\emptyset(N \log N)$, as opposed to the cumbersome $\emptyset(N^2)$, allowing for longer and more accurate electrostatics to be implemented. Furthermore, the use of periodic boundary conditions has the added benefits of controlling the number of atoms in our pre-defined volume as a constant, a necessary condition for many statistical ensembles to be maintained accurately. Use of PME long range electrostatics also necessitates that the total system charge be neutral, a condition which is generally remedied by addition of charged ions.

1.4 Defining Our Scope

As is hopefully now apparent, the central focus of this work will involve, at the very least, examining and characterizing *motion*, namely attempting to close the gap between the *terra incognita* of thermodynamic averaging and single molecule type direct observations. Specifically we will operate under the assumption that molecular dynamics trajectories, while spelling out the most plausible fully atomistic motions of macromolecules, *cannot, and will not ever,* fully resolve all possible trajectories available to these molecules in any cognitively

graspable way. We will similarly operate under the thermodynamic uncertainy principle, namely that thermodynamics has within it a hard limit to the reaches of its inquiry. We will specifically maintain that thermodynamics alone *cannot, and will not ever,* be able to posit any absolutely certain conclusions beyond a *probabilistic* description of energy for a given set of molecules. Appropriately we must remember that we cannot hastily posit *causitive* hypotheses about the relationship between the potential energy function and the motions of molecules, but simply observe them and validate those motions within the appropriate thermodynamic framework.

With this in mind our focus can shift away from the absolute prediction of trajectories in any deterministic way and instead favor a probabilistic interpretation, looking for *correlations* between groups of motions at different parts of the molecule or across various timescales. From this vantage point we will begin our investigation with DNA and RNA at the nanosecond timescale, using the S^2 order parameter as a way of relating our simulated (single molecule) results to thermodynamic averages (or ensemble based experimental) motions (chapter 2). We will then do a similar analysis for microseconds (chapter 3), and discuss at length the implications which can be gleaned from comparing the two (chapter 4). We will pay special attention to the instance in which *smooth, continual* motions are observed (or implied) whether it is across the molecule or across timescales. Specifically we will use this approach to point out a major difference between DNA and RNA, namely that while DNA moves diffusively from A to B like conformers along subtle backbone vibrational modes, conformational changes seem unlikely at these two timescales unless environmental factors drive the conformational change. In contrast, our results will suggest that small functionally active RNA molecules have quite a unique dynamical character spanning from the nanosecond to microsecond timescale, a result which is in good agreement with previous experimental and computational data [72, 47, 73].

Finally we will demonstrate, through pointed application, the power and efficacy of elucidating smooth continual motions and how our modes of inquiry should be aligned with goals to design "embodied" cellular agents which can interact with, or better said, subtly influence, the vibrational structure of nucleic acids in solution (chapter 5).

Bibliography

[1] Eörs Szathmáry. The origin of the genetic code: amino acids as cofactors in an rna world. *Trends in genetics*, 15(6):223–229, 1999.

[2] Sven Siebert. Common sequence structure properties and stable regions in rna secondary structures. 2006.

[3] Alberto Pérez, F Javier Luque, and Modesto Orozco. Dynamics of b-dna on the microsecond time scale. *Journal of the American Chemical Society*, 129(47):14739–14745, 2007.

[4] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[5] Ralf Dahm. Friedrich miescher and the discovery of dna. *Developmental Biology*, 278(2):274–288, 2005.

[6] James D Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[7] Ignacio Tinoco Jr and Carlos Bustamante. How rna folds. *Journal of molecular biology*, 293(2):271–281, 1999.

[8] TE Cloutier and Jonathan Widom. Dna twisting flexibility and the formation of sharply looped protein–dna complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3645–3650, 2005.

[9] FHe Crick. Linking numbers and nucleosomes. *Proceedings of the National Academy of Sciences*, 73(8):2639–2643, 1976.

[10] M Nirenberg, P Leder, M Bernfield, R Brimacombe, J Trupin, F Rottman, and C O'neal. Rna codewords and protein synthesis, vii. on the general nature of the rna code. *Proceedings of the National Academy of Sciences of the United States of America*, 53(5):1161, 1965.

[11] John S Mattick. The hidden genetic program of complex organisms. *Scientific American*, 291(4):60, 2004.

[12] Alan Cooper. Protein fluctuations and the thermodynamic uncertainty principle. *Progress in biophysics and molecular biology*, 44(3):181–214, 1984.

[13] Cécile Morette DeWitt. Feynman's path integral. *Communications in Mathematical Physics*, 28(1):47–67, 1972.

[14] DA McQuarrie. Statistical mechanics, 1976. *Happer and Row, New York*.

[15] Jerry B Marion and Stephen T Thornton. *Classical dynamics of particles and systems*. Saunders College Pub., 1995.

[16] Shinji Sunada, Nobuhiro Go, and Patrice Koehl. Calculation of nuclear magnetic resonance order parameters in proteins by normal mode analysis. *The Journal of chemical physics*, 104(12):4768–4775, 1996.

[17] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Current opinion in structural biology*, 9(2):164–169, 1999.

[18] Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Current opinion in structural biology*, 10(2):165–169, 2000.

[19] J Andrew McCammon and Stephen C Harvey. *Dynamics of proteins and nucleic acids*. Cambridge University Press, 1988.

[20] Fareed Aboul-ela, Jonathan Karn, and Gabriele Varani. Structure of hiv-1 tar rna in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic acids research*, 24(20):3974–3981, 1996.

[21] Xiang-Jun Lu and Wilma K Olson. 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic acids research*, 31(17):5108–5121, 2003.

[22] Jeff Wereszczynski and Ioan Andricioaei. On structural transitions, thermodynamic equilibrium, and the phase diagram of dna and rna duplexes under torque and tension. *Proceedings of the National Academy of Sciences*, 103(44):16200–16205, 2006.

[23] Richard E Dickerson, Horace R Drew, Benjamin N Conner, Richard M Wing, Albert V Fratini, Mary L Kopka, et al. The anatomy of a-, b-, and z-dna. *Science*, 216(4545):475–485, 1982.

[24] Chun Yoon, Gilbert G Privé, David S Goodsell, and Richard E Dickerson. Structure of an alternating-b dna helix and its relationship to a-tract dna. *Proceedings of the National Academy of Sciences*, 85(17):6332–6336, 1988.

[25] Uli Schmitz, Gerald Zon, and Thomas L James. Deoxyribose conformation in [d (gtatatac)] 2: evaluation of sugar pucker by simulation of double-quantum-filtered cosy cross-peaks. *Biochemistry*, 29(9):2357–2368, 1990.

[26] Jason D Kahn, Elizabeth Yun, and Donald M Crothers. Detection of localized dna flexibility. 1994.

[27] Horace R Drew and Andrew A Travers. Dna bending and its relation to nucleosome positioning. *Journal of molecular biology*, 186(4):773–790, 1985.

[28] JY Lee, Burak Okumus, DS Kim, and Taekjip Ha. Extreme conformational diversity in human telomeric dna. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):18938–18943, 2005.

[29] Donald M Gray, Su-Hwi Hung, and Kenneth H Johnson. Absorption and circular dichroism spectroscopy of nucleic acid duplexes and triplexes. *Methods in enzymology*, 246:19–34, 1994.

[30] Thierry Dauxois, Michel Peyrard, and AR Bishop. Entropy-driven dna denaturation. *Phys. Rev. E*, 47(1):R44–R47, 1993.

[31] Ioulia Rouzina and Victor A Bloomfield. Force-induced melting of the dna double helix 1. thermodynamic analysis. *Biophysical journal*, 80(2):882–893, 2001.

[32] Helen G Hansma, Kenichi Kasuya, and Emin Oroudjev. Atomic force microscopy imaging and pulling of nucleic acids. *Current opinion in structural biology*, 14(3):380–385, 2004.

[33] Haran T. E. The unique structure of a-tracts and intrinsic dna bending. *Q. Rev. Biophys.*, 42:41, 2009.

[34] A. K. Mazur. Anharmonic torsional stiffness of DNA revealed under small external torques. *Phys. Rev. Lett.*, 105:018102, Jun 2010.

[35] Kathleen B Hall. Rna in motion. *Current opinion in chemical biology*, 12(6):612–618, 2008.

[36] Paul G Higgs. Rna secondary structure: physical and computational aspects. *Quarterly reviews of biophysics*, 33(03):199–253, 2000.

[37] Thomas Hermann and Dinshaw J Patel. Rna bulges as architectural and recognition motifs. *Structure*, 8(3):R47–R54, 2000.

[38] Aurelie Lescoute, Neocles B Leontis, Christian Massire, and Eric Westhof. Recurrent structural rna motifs, isostericity matrices and sequence alignments. *Nucleic Acids Research*, 33(8):2395–2409, 2005.

[39] Philippe Brion and Eric Westhof. Hierarchy and dynamics of rna folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997.

[40] Neocles B Leontis and Eric Westhof. Analysis of rna motifs. *Current opinion in structural biology*, 13(3):300–308, 2003.

[41] Jan Ferner, Alessandra Villa, Elke Duchardt, Elisabeth Widjajakusuma, Jens Wöhnert, Gerhard Stock, and Harald Schwalbe. Nmr and md studies of the temperature-dependent dynamics of rna ynmg-tetraloops. *Nucleic acids research*, 36(6):1928–1940, 2008.

[42] Martin Zacharias and Paul J Hagerman. Bulge-induced bends in rna: quantification by transient electric birefringence. *Journal of molecular biology*, 247(3):486–500, 1995.

[43] Tamás Kiss. Small nucleolar rnas: an abundant group of noncoding rnas with diverse cellular functions. *Cell*, 109(2):145–148, 2002.

[44] David P Bartel, Maria L Zapp, Michael R Green, and Jack W Szostak. Hiv-1 rev regulation involves recognition of non-watson-crick base pairs in viral rna. *Cell*, 67(3):529–536, 1991.

[45] Jung C Lee and Robin R Gutell. Diversity of base-pair conformations and their occurrence in rrna structure and rna structural motifs. *Journal of molecular biology*, 344(5):1225–1249, 2004.

[46] A Lescoute and E Westhof. The interaction networks of structured rnas. *Nucleic acids research*, 34(22):6587–6604, 2006.

[47] Michael F Bardaro, Zahra Shajani, Krystyna Patora-Komisarska, John A Robinson, and Gabriele Varani. How binding of small molecule and peptide ligands to hiv-1 tar alters the rna motional landscape. *Nucleic acids research*, 37(5):1529–1540, 2009.

[48] Gary D Stormo and Yongmei Ji. Do mrnas act as direct sensors of small molecules to control their expression? *Proceedings of the National Academy of Sciences*, 98(17):9465–9467, 2001.

[49] Nils G Walter. Structural dynamics of catalytic rna highlighted by fluorescence resonance energy transfer. *Methods*, 25(1):19–30, 2001.

[50] Jamie H Cate, Marat M Yusupov, Gulnara Zh Yusupova, Thomas N Earnest, and Harry F Noller. X-ray crystal structures of 70s ribosome functional complexes. *Science*, 285(5436):2095–2104, 1999.

[51] Christopher Hammond and Christopher Hammond. *Basics of crystallography and diffraction*, volume 214. Oxford, 2001.

[52] Joseph P Hornak. Basics of nmr, 1997.

[53] G Marius Clore, Attila Szabo, Ad Bax, Lewis E Kay, Paul C Driscoll, and Angela M Gronenborn. Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *Journal of the American Chemical Society*, 112(12):4989–4991, 1990.

[54] Eric R Henry and Attila Szabo. Influence of vibrational motion on solid state line shapes and nmr relaxation. *The Journal of chemical physics*, 82(11):4753–4761, 1985.

[55] Catherine Musselman, Qi Zhang, Hashim Al-Hashimi, and Ioan Andricioaei. Referencing strategy for the direct comparison of nuclear magnetic resonance and molecular dynamics motional parameters in rna. *The Journal of Physical Chemistry B*, 114(2):929–939, 2009.

[56] Eike Brunner. Residual dipolar couplings in protein nmr. *Concepts in Magnetic Resonance*, 13(4):238–259, 2001.

[57] Alexander D MacKerell, Bernard Brooks, Charles L Brooks, Lennart Nilsson, Benoit Roux, Youngdo Won, and Martin Karplus. Charmm: the energy function and its parameterization. *Encyclopedia of computational chemistry*, 1998.

[58] Scott J Weiner, Peter A Kollman, David A Case, U Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984.

[59] Lennart Nilsson and Martin Karplus. Empirical energy functions for energy minimization and dynamics of nucleic acids. *Journal of computational chemistry*, 7(5):591–616, 1986.

[60] Alexander D MacKerell Jr, Joanna Wiorkiewicz-Kuczera, and Martin Karplus. An all-atom empirical energy function for the simulation of nucleic acids. *Journal of the American Chemical society*, 117(48):11946–11975, 1995.

[61] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[62] Mike P Allen and Dominic J Tildesley. Computer simulation of liquids. 1989.

[63] Dominique Levesque and Loup Verlet. Molecular dynamics and time reversibility. *Journal of Statistical Physics*, 72(3-4):519–537, 1993.

[64] Dennis C Rapaport. *The art of molecular dynamics simulation*. Cambridge university press, 2004.

[65] Elizabeth J Denning, U Priyakumar, Lennart Nilsson, and Alexander D Mackerell. Impact of 2-hydroxyl sampling on the conformational properties of rna: Update of the charmm all-atom additive force field for rna. *Journal of computational chemistry*, 32(9):1929–1943, 2011.

[66] Nicolas Foloppe and Alexander D MacKerell Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, 21(2):86–104, 2000.

[67] Alexander D Mackerell and Nilesh K Banavali. All-atom empirical force field for nucleic acids: Ii. application to molecular dynamics simulations of dna and rna in solution. *Journal of Computational Chemistry*, 21(2):105–120, 2000.

[68] Elzbieta Kierzek, Anna Pasternak, Karol Pasternak, Zofia Gdaniec, Ilyas Yildirim, Douglas H Turner, and Ryszard Kierzek. Contributions of stacking, preorganization, and hydrogen bonding to the thermodynamic stability of duplexes between rna and 2-o-methyl rna with locked nucleic acids. *Biochemistry*, 48(20):4377–4387, 2009.

[69] Darrin M York, Tom A Darden, and Lee G Pedersen. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: a comparison of the ewald and truncated list methods. *The Journal of chemical physics*, 99(10):8345–8348, 1993.

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

[71] Simon W de Leeuw, John W Perram, and Edgar R Smith. Simulation of electrostatic systems in periodic boundary conditions. i. lattice sums and dielectric constants.

Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences, 373(1752):27–56, 1980.

[72] Catherine Musselman, Hashim M Al-Hashimi, and Ioan Andricioaei. ired analysis of tar rna reveals motional coupling, long-range correlations, and a dynamical hinge. *Biophysical journal*, 93(2):411–422, 2007.

[73] Tim Zeiske, Kate A Stafford, Richard A Friesner, and Arthur G Palmer. Starting-structure dependence of nanosecond timescale intersubstate transitions and reproducibility of md-derived order parameters. *Proteins: Structure, Function, and Bioinformatics*, 81(3):499–509, 2013.

Chapter 2

The Nanosecond

2.1 Introduction

Perhaps one of the most interesting phenomena about biological simulation is the way that it marches forward on the heels of computational efficiency while asking similar questions regardless of large advances in timescale access [1, 2]. Fundamentally the question of biological relevance is one of relating *where* and *when* to meaningful conceptual labels while validating those motions by experimental trial and error. Despite so many advances, biology is still constrained to begin many investigations in simple observation due to its relative infancy as a discipline. As a result, biological investigations nest themselves hierarchically in the complicated and fractal like organization of systems of the body as we slowly elucidate them across several disciplines. This complicated result ensures that we will be searching for ways to "look" more efficiently at the body and cellular mechanisms for the foreseeable future. This is unlike other disciplines that have settled on a set of canonized frames through which they view static timescale windows of interest. It is the strange and unique opportunity that we share with a few other disciplines who to study the *act* of scaling time.

In order to do this however, we must first carefully understand the time regime in which we start the scaling operations, and subsequently the scale in which we finish the operation, before we can discuss the differences between the two. We will almost certainly see, however, that carefully and meticulously investigating how we think about the subject will result in greater clarity and understanding (and by extension, embodied biological action) when we endeavor to understand the underlying principles at play here. We start the discussion with the nanosecond, or an increment of time about 1,000,000,000x faster than a normal second.

2.1.1 Biological Processes of Nucleic Acids at the Nanosecond Timescale

Atomic fundamental vibrational frequencies can now be reliably calculated by quantum mechanical calculations, and we can squarely detect and measure with incredible accuracy those movements that fall within certain timescales [3], although this is a complicated field in its own right, and we can hardly to it justice here. Hydrogen bonds move on the order of femtoseconds, which provides the maximum for time resolution that can be reliably calculated using Newtonian based mechanics [4]. Local energetic relaxation seems to happen on the order of picoseconds, although this clearly depends on starting conditions and thermodynamic variables such as temperature [5]. Nucleic acid simulations began in the picosecond regime [6], which seems to match the resolution needed to reproduce certain physical properties of liquids (although this too can become very complicated in its own right), and the reparameterization of the force field for nucleic acids in the condensed phase (as opposed to the gas phase) is when this effect was accounted for [7]. Statistical mechanics provides analytical solutions to idealized problems that show non-equilibrium dynamics of liquids can be reliably calculated at picoseconds, and early simulations were in agreement with this [8, 9]. In short, picoseconds and subsequently 1000s of picoseconds (nanoseconds) were the first biologically significant results our simulations provided.

Cellular processes on this timescale have a rich and full zoology, but in general enzyme catalysis happens on the picosecond range [10, 11], while diffusion across liquids to cellular targets seems to take much longer [12, 13]. In general anything that can be tracked macroscopically (by microscope or by GFP for example) takes at the minimum milliseconds (100,000x longer). The details of allosteric modification are largely still unclear but seem to fall somewhere between picoseconds and microseconds depending on the cellular entity and whether there are chaperones or other cellular members involved. Furthermore it appears that the complicated action of allosteric structural interconversion involves correlated fast and slow timescale movements, further complicating the issue [14]. Transcription and replication processes have sub-processes that seem to happen very fast compared to other cellular signaling processes (nanoseconds to microseconds) but in general the overall processes take much longer than nanoseconds [13, 15]. Molecular species folding and inter-conversion events takes somewhere longer than nanoseconds, although this problem has its own subtlety due to entropic considerations that are difficult to tackle using simulation, particularly for nucleic acids [16].

A study by Fisette et al at the university of Quebec sum up the timescales which have become accessible to simulation along with timescales available to experiment by NMR and biological processes [2]. Figure 2.1, which was adapted from the aforementioned study, shows clearly those processes discussed earlier with the current uncertainties in place. While these guides are practical for navigating the tricky world of what can be known with current NMR techniques, it should be noted that non-equilibrium techniques can address many if not all of the timescales (from femtosecond to millisecond) but introduce significant limitations due to the loss of information and perturbation of the potential. It should also be noted that regardless of how clever the non-equilibrium techniques are, they are only as good as their force field (or experimental data being used to filter the results). For this reason we should be careful when calling these timescales accessible in that many questions remain unanswered regarding their implementation and accuracy for various systems.

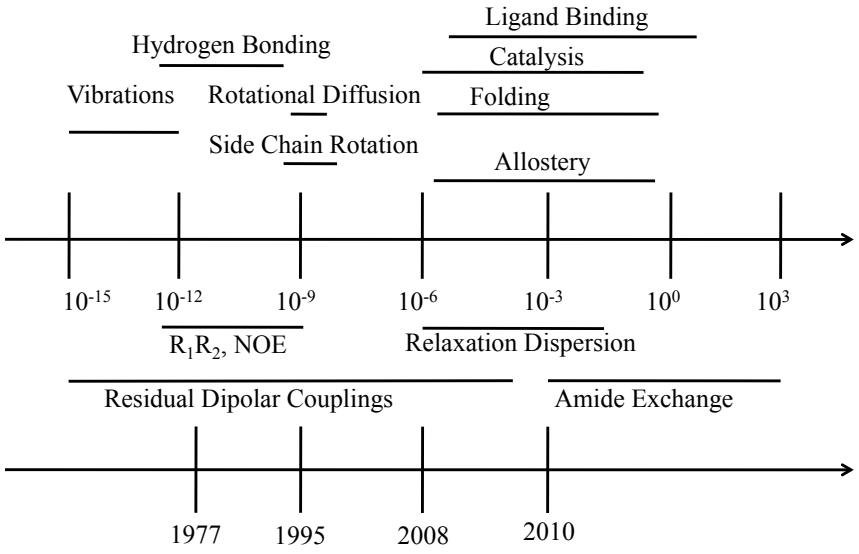

Figure 2.1: Demonstrating which motions occur at which timescales and the current access we have in probing at those timescales. On the bottom row is shown the year that each timescale became accessible to MD simulation. Figure adapted from Fisette et al [2].

2.1.2 DNA Motions on the Nanosecond Timescale

While DNA is generally thought of as a relatively rigid molecule, it is actually a very dynamic entity with a variety of motions available to it on all timescales. While macroscopically there has been success modeling it as a piece of string or solving certain motions analytically, [17, 18, 19, 20] the motions available to it on the nanosecond timescales are actually quite intricate. Small bubble formation is an active area of study thought to be important to transcription and replication that can happen naturally by thermal fluctuations [21], although the specifics of timescales involved are still somewhat difficult to discern. Base extrusion up until recently was thought to happen only by protein mediation, but surprising results suggest some base extrusion and even syn to anti-syn rotation about the glycosidic bond could result in stable hoogsteen paired bases, which have been seen bound to p53 an important protein in cancer research [22, 23]. While all of these motions are available to DNA, it seems that most of them happen on somewhat longer timescales in the absence of increased temperature or external force. The A/B transition, however, may fall on these scales if abrupt change in chemical environment favors one or the other [24].

One of the more interesting aspects of DNA dynamics is its inter-dependence between sequence asnd overall structure. While much work has been done in this area, the basics pertaining to nanosecond motion involve increased rigidity and global bending near AT rich sequences, but the specific biophysical reasons are still controversial [25, 26]. In experiment, most of the current body of knowledge regarding DNA structure and movement results from some form of force puling, which measures internal biophysical parameters important for various cellular engineering and methodology refinement, although most of these assume a fairly rigid molecule being physically separate by size (length), orientation (overall tumbling) or sequence (force pulling) [27]. For a full treatment see the introduction to this dissertation or the discussion of this chapter.

2.1.3 RNA Motions on the Nanosecond Timescale

RNA, however, seems to move around quite a bit more on the nanosecond timescale. RNA is an active participant in enzyme catalysis that requires picosecond dynamics at the very least, and RNA's many roles in the cell require that it must have access to many (cellular) biological timsescales, if not all. A study by Zhang et al lab provides one of the most directed studies in this area, clearing showing picosecond motions vs nanosecond motions or HIV-1 based RNA molecules by domain elongation strategy. Domain elongation strategy involves elongating the helix to which the bio-macromolecule can be attached to a micelle that is then oriented relative to an external magnetic field [28]. This clever approach effectively "freezes" overall tumbling motions separating global tumbling from internal fluctuations. By squeezing out these slower motions, the Hashimi lab is able to resolve many molecular motions of both DNA and RNA at much faster timescales than previously accessible [22]. In general, this mode of inquiry has led to the understanding that while DNA base motions are likely occurring either very sparsely or at the microsecond timescale, RNA base extrusion and possible flipping (Watson crick to Hoogsteen transition) happens much more seldom if at all and has not been observed in solution to date. Base motions for RNA do, however, intercalate or extrude relatively quickly, on the order of nanoseconds near bulge or loop regions [29]. It should also be noted that the elongated RNA duplexes are commonly found to be retained in terms of sequence, and the elongated duplexes are relatively stable and resemble DNA dynamics, with added flexibility and possible local melting [30]. It should also be mentioned that like DNA, major conformational changes and folding events can be observed following abrupt changes in the liquid environment, such as temperature and salt concentration [30].

2.2 Methods

2.2.1 Choosing Structures to Simulate

Despite their ubiquitous nature and importance to biology vast differences in the structure, function, and dynamics of nucleic acids make it difficult to choose the most affective structures/sequences for general study. We have chosen idealized B-form DNA of sequence $5-CGAT_6GGC-3$, $5-CGCGAT_4GGC-3$, and $5-GCATCGAT_2GGC-3$ (referred to as A6, A4, and A2 DNA respectively) and the TAR HIV-1 RNA sequence as the focus of this study. The A-tract is a feature common to DNA that has been shown to be of vital importance to gene expression and DNA dynamics [31]. It has been shown that heterogeneous DNA sequences tend to exhibit characteristics close to idealized B-form DNA, but that increasing AT tract length leads to DNA rigidity and progressive decrease in inter-helical distance [32], along with increased global bending [33]. The HIV-1 Transactivating Response Element (TAR) is a heavily studied RNA sequence that adopts an A-like conformation with characteristic bulge and loop features [34]. The TAR sequence is a vital therapeutic target critical in the progression of the HIV pathology [35].

All systems considered in this study (with the exception of chapter 5) were chosen such that there would be NMR data available. Specifically, Residual Dipolar Couplings provide an ensemble based time-independent measure of structures fluctuating all across various timescales, and S^2 measure fluctuations constricted to a much more confined timescale (see above). RDCs for DNA are unfortunately not available to my knowledge, but they are available for RNA, which will be used as evidence of concept later in the study. In this section we are primarily concerned with the generation of S2 for both DNA and RNA. For full derivation and details of the measurements the reader is referred to Musselman et al for TAR RNA and Nikolova et al for A-tract DNA. At the end of the chapter we report the data from these measurements in table form with permission from authors [36, 37].

2.2.2 Simulation Parameters

Nanosecond based simulations were carried out on A2,4,6DNA and HIV-1 TAR RNA. Specifically in the case of HIV-1 TAR RNA, initial coordinates were obtained from the Protein Data Bank (Access Code 1ANR) [34]. The system was solvated in VMD [38] with TIP3 water and 27 sodium ions to neutralize charge in a 64x64x64 $\mathring{A}$ cube for initial heating, which was carried out in the CHARMM Molecular Dynamics package(CHARMM) [39] or NAMD [40]. CHARMM36 force field parameters for ribonucleic acids were used, including recent changes made in 2011 by Mackarell et al [41]. All systems were heated gently to 300K with harmonic constraints on backbone atoms for 100ps until restraints were gradually released over another 100ps and the system was equilibrated for 5ns. All DNA structures were similarly built in Nucleic Acid Builder NAB, part of AMBERTOOLS [42] and solvated in Visual Molecular Dynamics [42]. DNA simulations were heated carefully to minimize fraying effects, simulated with full water and ions sufficient to neutralize charge with periodic boundary conditions in CHARMM (CHARMM) with the most recent CHARMM DNA force fields [43]. All A-tract trajectories were started from canonical B-form DNA, and all simulations were carefully checked for fraying and maintenance of B-form throughout the trajectories. Figures from the trajectories are included below.

2.2.3 Calculating Nanosecond based S^2 Order Parameters

For a more complete pedagogical treatment of the calculation of S^2 order parameters the reader is referred to either the dissertation introduction, chapter 4, appendix A (which is a reprint of [36]). For the purposes of this section however we can simply define the S^2 as a collective variable which represents the magnitude of motion of a given bond vector. It is given in short form by the Lipari-Szabo model free formalism in which the autocorrelation function is given as a function of the bond vector $\hat{\mu}$:

$$C(t) = \langle P_2[(\hat{\mu}(0).\hat{\mu}(t))]\rangle \tag{2.1}$$

where P_2 refers to the second order legendre polynomial and angled brackets denote time averaging. From here we can use the parameterized version of the autocorrelation function defined by Clore, Szabo, Lipari and Henry [44, 45]:

$$C(t) = S^2 + (1 - S^2)e^{-t/\tau_f} \tag{2.2}$$

Where the subscript f refers to fast motions (internal motions as opposed to tumbling), and S^2 is the plateau value that the $C(t)$ converges to. This makes for a very straightforward calculation from MD trajectories, seeing as the P_2 autocorrelation function is easily calculated, and the long time limit can be calculated as a simple average of the tail value of $C(t)$. In practice the averaging window tends to be placed somewhere after significant relaxation and before the $C(t)$ becomes unstable due to sparse data (around $1/10^{th}$ of the total trajectory).

2.3 Results

2.3.1 DNA Nanosecond Simulations and Calculated S^2 Order Parameters

Ten molecular dynamics simulations of A_2, A_4, and A_6 DNA were carried out in CHARMM with care taken to ensure they were distinct trajectories (see methods for details of simu-

lations). Results of S^2 values calculated from MD trajectories are given in Figure 2.4. In general increased A-tract length showed larger S^2 (large periodicity) in C1'-H1 bond vectors indicating decreased sugar mobility for these residues, which is in good agreement with experimental values. Terminal residues showed decreased stability in poor agreement but it is well known that this is common for DNA simulation due to fraying effects [46]. Figure B shows the agreement with experimental values calculated in collaboration with the Hashim Al- Hashimi lab by Evgenia Nikolova [36]. Again good agreement is shown, with a few exceptions. Places were simulation overestimated S^2 (too much periodicity) were seldom seen except for A2-DNA, where C1 cytosine residues are over-stable when compared to experiment; along with AT_n flanking GC steps, which in experiment show decreased stability but in simulation showed stability similar to GC steps without AT_n flanking regions. In short it would seem that the simulations can recreate with good agreement the majority of bond vector motions (even subtle motions not explicitly parameterized) but that subtle affects of sequence are not captured. For full discussion of differences and derivation of S^2 values, the fully published manuscript is provided in the appendix A or in [36]. Examples of structural motifs are given as snapshots shown below in figure 2.2.

2.3.2 RNA Nanosecond Simulations

Many nanosecond-based simulations have been carried out on RNA, and specifically for HIV-1 TAR RNA, but the majority of them were carried and data published prior to the genesis of this dissertation [35]. There is a large amount of data characterizing what makes a good nanosecond RNA simulation. Primarily one should look for proper A-form maintenance, which will include C3-endo sugar pucker [47]. Furthermore loop residues should be far from bulge residues, and loop and bulge residues should interact minimally with domain I or domain II base residues as the simulation begins. For a more complete characterization of good nanosecond RNA simulations and the methods in which they are screened to fit

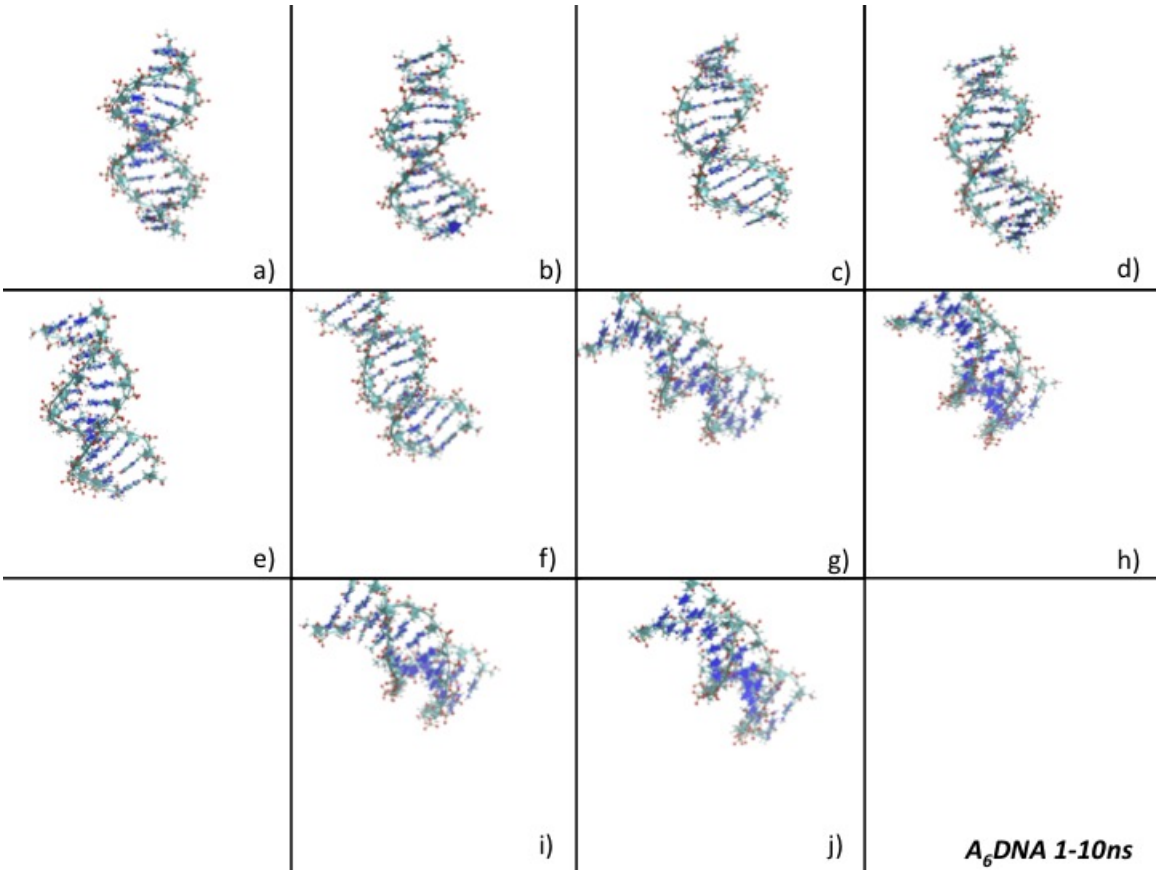

Figure 2.2: Examples of A_6DNA nanosecond simulations. Ten independent trajectories for each A_2DNA, A_4DNA, A_6DNA were generated totaling thirty trajectories (300ns in total), and order parameters were calculated for all trajectories and then averaged by sequence. Additionally, structures were heated gently in order to decrease fraying effects, which was observed to have significant effects on final S^2 order parameters (data not shown). Here is shown A_6DNA from 1 to 10 ns. Global tumbling (rigid body rotation) has not been removed in order to note that the overall tumbling is slow compared to the fast vibrations of individual molecules and bases. This is not the case for TAR RNA even at nanosecond times, as we see in figure 2.5

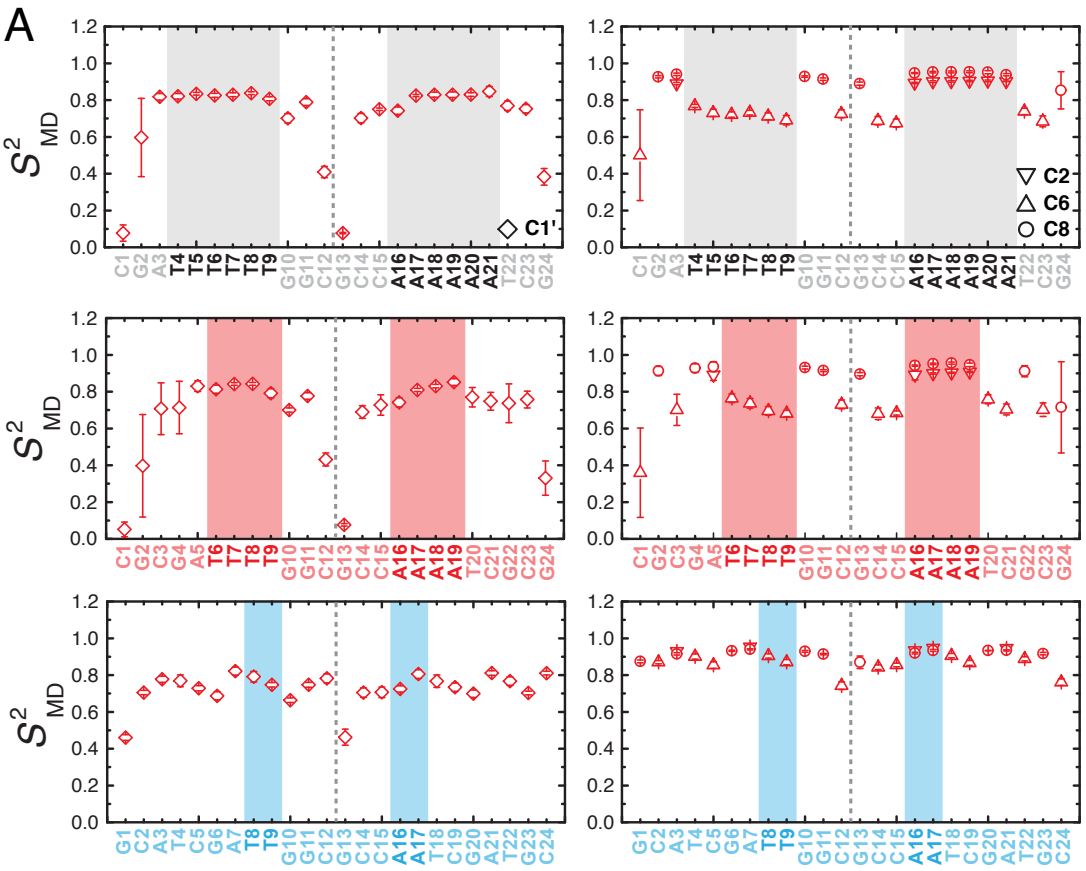

Figure 2.3: S2 order parameters obtained by MD simulations for base (C2-H2, C6-H6, C8-H8, shown in right column) and deoxyribose (C1'-H1', shown in left column) sites in A_6DNA, A_4DNA, and A_2DNA (top, middle, and bottom rows respectively). A-tract regions are shown in varying colors (A_6DNA, A_4DNA, and A_2DNA in grey, red, and blue respectively) to demonstrate increased stability in thymine residues. We see the characteristic S^2 profile, with stable members inside the structure, but end residues showing increased instability (lower S^2 measurements). Averages were taken across 10 independent trajectories and shown here, while error bars were calculated by variance across the ensemble of trajectories. Data taken from ref [36]. A copy of the manuscript from which the figures were taken is also provided in appendix A.

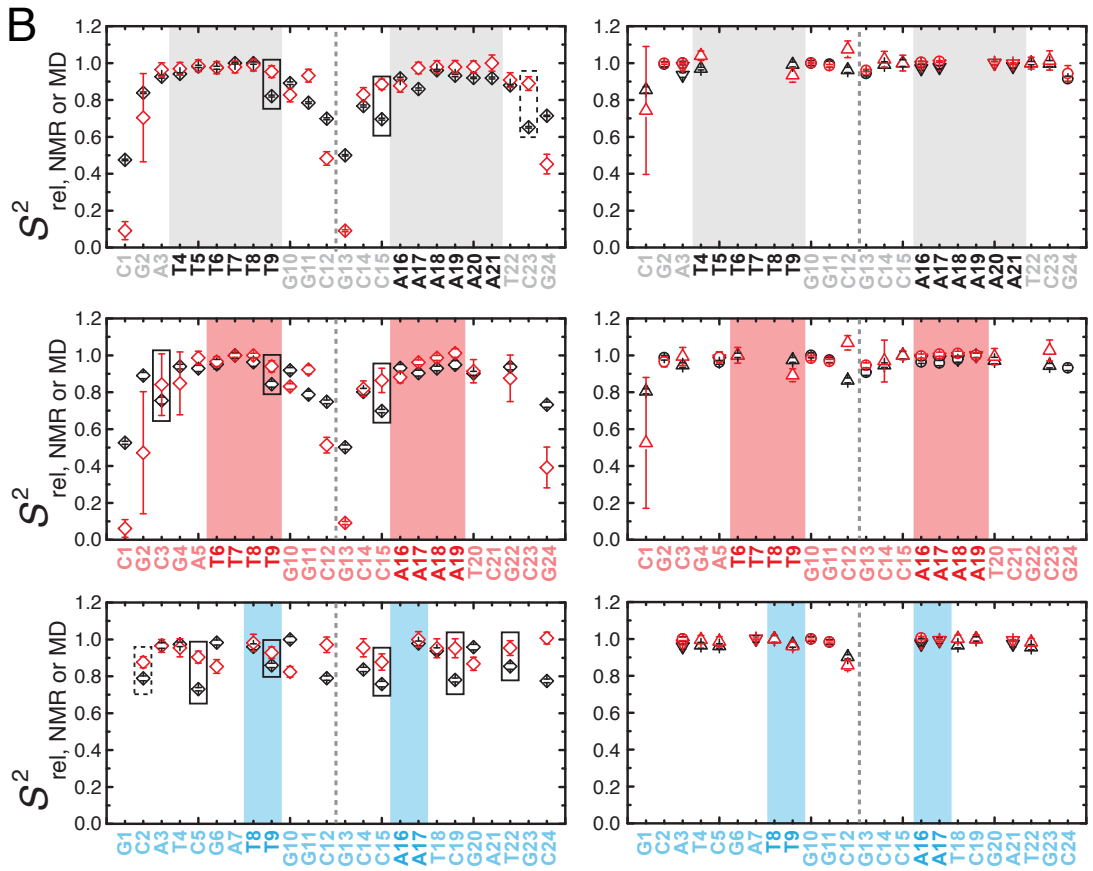

Figure 2.4: Comparison between order parameter S^2 obtained by NMR ^{13}C spins relaxation (red) and MD simulations (black). Left three plots show values for C1'-H1' bonds, where the right column shows C2-H2, C6-H6, and C8-H8 as inverted triangle, triangle and circle respectively (same as previous figure, except experimental is now shown in red). Good agreement is shown with the exception of end-fraying effects and increased instability at AT_n flanking GC steps. NMR S^2 values taken with permission from [36], a copy of which is available in appendix A.

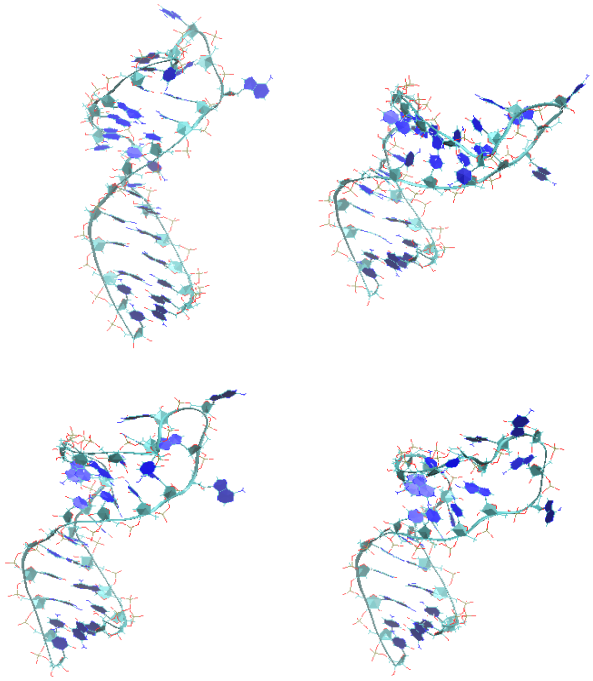

Figure 2.5: Examples of typical TAR RNA motions during a 5 nanosecond simulation. These simulations are common place and were checked against the body of literature pertaining to RNA nanosecond simulations. Notice that in 5 nanoseconds the RNA has not yet had a chance to intercalate or extrude any base residues near the bulge or loop regions, but major conformational change is seen between the left and the right structures. Also notice the hinge action of the bulge region, around which the secondary structure fluctuates.

experimental results, the reader is referred to Frank et al [35]. In chapter 4 we will more fully consider nanosecond based RNA movements in attempt to compare them to microsecond dynamics, but several examples of nanosecond based simulation snapshots with good overall form are given in figure 2.5.

2.3.3 Residual Dipolar Couplings and Nanosecond Simulations

Interestingly, RDCs computed from a three-microsecond trajectory computed on the Anton supercomputer shows little to no improvement in agreement from nano second based RDCs, which are also in poor agreement (data is shown explicitly in chapter 4). While it is understood that nanosecond trajectories give poor RDC agreement [48], it was suggested that this inconsistency arose at least in part due to incompatibility of timescales. In other words, many assumed that upon lengthening said trajectories sufficiently would lead to good agreement of all NMR order parameters. It is later shown in this document (chapter 4) that even upon extending the simulation length by orders of magnitude there is little improvement in terms of agreement of calculated RDCs with experimentally measured RDCs. While disconcerting, this is a common thread for MD simulation. It also raises a familiar question, have we yet to saturate fully the states and therefore do not see good agreement with respect to RDCs, or is there a systematic problem with force fields? An elegant attempt to answer this question by even longer microsecond simulation screening results is discussed further in chapter 4, and the reader is referred to [29] for further reading on the subject.

2.4 Discussion

For a complete discussion of the possible conclusions which can be determined from the experimental S^2 data, the reader is referred to the discussion by Nikolova et al [36] given in this document as appendix (). In terms of the ability of MD simulations to recreate said S^2 order parameters the agreement is quite excellent and provides us with at least some assurance of the ability of MD to capture the essential motions of both backbone and base dynamics of A246DNA from picoseconds to nanoseconds. It is particularly important to note that the deviations from experimental data are seen primarily in reference to tertiary effects of sequence. This is actually somewhat expected, in that the bases in question were

not explicitly parameterized to reflect homogenous sequences, but more so heterogenous sequences of various bases. This gets at the heart of the one of the major theses of this dissertation, *that there is more to genetic encoding than the code itself.* Specifically at AT rich sequences the DNA tends to bend itself in a global way necessitating increased flexibility at adjacent dinucleotide steps [33]; and it is the effects of this tertiary process that is poorly modeled in DNA simulation, while other subtle motions seem to be well modeled even when not explicitly parameterized.

Furthermore there has been much discussion in recent years regarding the specific mobility of purine/pyrimidine dinucleotide steps near relatively homogenous sequences, particularly adjoining AT rich sequences. While this discussion is important in the context of transcription factor binding [49] and further speculation regarding the overall effects on structural dynamics, it is the opinion of this author that such subtleties are not well addressed by intuition, and the simulations are clearly biased or sparse in this area. While there is clearly *some* effects of sequence on tertiary structure visible in experiment, and those effects are likely correlated in some way with flexibility, the S^2 data does not explicitly describe why the increased motion is observed, or characterize the motion in any specific way. It is clear that MD simulations are still suffering some need of increased accuracy before we can attribute specific causal relationships regarding this observation. Additionally, it is our duty as scientists to explicitly model when possible and carefully compare the results with experiment and only then conclude about the labels appropriate to assign to causation and prediction. At the risk of sounding a bit harsh, after careful consideration of the evidence the aformentioned criteria for certainty (or even speculation) is *not* met with the current body of data. These are largely questions which cannot be answered analytically (this has been exhaustively shown [50] and most extensions from the currently sparse and data sets amount to simple extensions of *desired* or *synergistic* outcomes as opposed to carefully reasoned conclusions. We must not be hasty in our desire to assign subtle tertiary effects to the motions of DNA on *any* timescale (*in situ* or *in silica*, we should instead form careful

hypotheses and test them methodically as we have been encouraged to do by our mentors and previous scientists for some time now. If any conclusion can be drawn about the specific tertiary effects of sequence on dynamics, it is that we do not yet have accurate enough simulations to draw conclusions that reflect full chemical (or kinetic) accuracy, a tool which we desperately need if we are to increase our grasp on DNA (or our knowledge of DNA) as a tool.

2.5 Conclusion

It is still early in our careful examination of nucleic acids and the assignment of features to their dynamics, but there are many factors to consider before we can move on to draw conclusions about the concepts we have begun to analyze. In short we have considered the biological processes that lie at the heart of our existence, and begun to sketch some of the complexities that cloud our understanding of their movement with time. Although many observations have been made and we have glimpsed the complexity leading to these observations, we have primarily arrived at a single negative conclusion. Namely we have stated that, although we stand on an exciting transition period in which we are finally able to catch plausible sight the motions of these molecules, we can only cautiously posit that there are major effects which are not captured, and it is this careful and deliberate pace at which we will proceed. Only later will we solidify the focal theses already presented and fully grasp some fundamental insight. From here we move to analyzing motions that make up tens of microseconds, which is roughly akin to allowing motions which persist for 1 second to repeat 10,000x (about 2.5 hours).

Bibliography

[1] David E Shaw, Martin M Deneroff, Ron O Dror, Jeffrey S Kuskin, Richard H Larson, John K Salmon, Cliff Young, Brannon Batson, Kevin J Bowers, Jack C Chao, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

[2] 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.

[3] Takehiko Shimanouchi. Tables of molecular vibrational frequencies consolidated. volume i. Technical report, DTIC Document, 1972.

[4] Vincent Kräutler, Wilfred F van Gunsteren, and Philippe H Hünenberger. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry*, 22(5):501–508, 2001.

[5] CL Brooks III, MW Balk, and SA Adelman. Dynamics of liquid state chemical reactions: Vibrational energy relaxation of molecular iodine in liquid solution. *The Journal of Chemical Physics*, 79(2):784–803, 1983.

[6] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[7] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[8] John E Straub and D Thirumalai. Exploring the energy landscape in proteins. *Proceedings of the National Academy of Sciences*, 90(3):809–813, 1993.

[9] Alan Cooper. Protein fluctuations and the thermodynamic uncertainty principle. *Progress in biophysics and molecular biology*, 44(3):181–214, 1984.

[10] Joseph R Lakowicz and Gregorio Weber. Quenching of protein fluorescence by oxygen. detection of structural fluctuations in proteins on the nanosecond time scale. *Biochemistry*, 12(21):4171–4179, 1973.

[11] Stephen J Benkovic and Sharon Hammes-Schiffer. A perspective on enzyme catalysis. *Science*, 301(5637):1196–1202, 2003.

[12] Dominique Bourgeois, Beatrice Vallone, Friedrich Schotte, Alessandro Arcovito, Adriana E Miele, Giuliano Sciara, Michael Wulff, Philip Anfinrud, and Maurizio Brunori. Complex landscape of protein structural dynamics unveiled by nanosecond laue crystallography. *Proceedings of the National Academy of Sciences*, 100(15):8704–8709, 2003.

[13] Alan S Verkman. Solute and macromolecule diffusion in cellular aqueous compartments. *Trends in biochemical sciences*, 27(1):27–33, 2002.

[14] Bertil Halle. The physical basis of model-free analysis of nmr relaxation data from proteins and complex fluids. *The Journal of chemical physics*, 131(22):224507, 2009.

[15] DICK D Mosser, NICHOLAS G Theodorakis, and RICHARD I Morimoto. Coordinate changes in heat shock element-binding activity and hsp70 gene transcription rates in human cells. *Molecular and cellular biology*, 8(11):4736–4744, 1988.

[16] Richard Martin Ballew, Jobiah Sabelko, and Martin Gruebele. Observation of distinct nanosecond and microsecond protein folding events. *Nature Structural & Molecular Biology*, 3(11):923–926, 1996.

[17] De Witt Sumners. Untangling dna. *The Mathematical Intelligencer*, 12(3):71–80, 1990.

[18] Thierry Dauxois, Michel Peyrard, and AR Bishop. Entropy-driven dna denaturation. *Phys. Rev. E*, 47(1):R44–R47, 1993.

[19] Jiro Shimada and Hiromi Yamakawa. Statistical mechanics of dna topoisomers: the helical worm-like chain. *Journal of molecular biology*, 184(2):319–329, 1985.

[20] Hsiao-Ping Hsu, Wolfgang Paul, and Kurt Binder. Polymer chain stiffness vs. excluded volume: A monte carlo study of the crossover towards the worm-like chain model. *EPL (Europhysics Letters)*, 92(2):28003, 2010.

[21] Grégoire Altan-Bonnet, Albert Libchaber, and Oleg Krichevsky. Bubble dynamics in double-stranded dna. *Physical review letters*, 90(13):138101, 2003.

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

[23] Malka Kitayner, Haim Rozenberg, Remo Rohs, Oded Suad, Dov Rabinovich, Barry Honig, and Zippora Shakked. Diversity in dna recognition by p53 revealed by crystal structures with hoogsteen base pairs. *Nature structural & molecular biology*, 17(4):423–429, 2010.

[24] Nilesh K Banavali and Benoît Roux. Free energy landscape of a-dna to b-dna conversion in aqueous solution. *Journal of the American Chemical Society*, 127(18):6866–6876, 2005.

[25] Tali E Haran and Udayan Mohanty. The unique structure of a-tracts and intrinsic dna bending. *Quarterly reviews of biophysics*, 42(01):41–81, 2009.

[26] Nicholas V Hud, Vladimir Sklenář, and Juli Feigon. Localization of ammonium ions in the minor groove of dna duplexes in solution and the origin of dna a-tract bending. *Journal of molecular biology*, 286(3):651–660, 1999.

[27] Ioulia Rouzina and Victor A Bloomfield. Force-induced melting of the dna double helix 1. thermodynamic analysis. *Biophysical journal*, 80(2):882–893, 2001.

[28] Qi Zhang, Xiaoyan Sun, Eric D Watt, and Hashim M Al-Hashimi. Resolving the motional modes that code for rna adaptation. *Science*, 311(5761):653–656, 2006.

[29] Loïc Salmon, Gavin Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. A general method for constructing atomic-resolution rna ensembles using nmr residual dipolar couplings: the basis for interhelical motions revealed. *Journal of the American Chemical Society*, 135(14):5457–5466, 2013.

[30] Kathleen B Hall. Rna in motion. *Current opinion in chemical biology*, 12(6):612–618, 2008.

[31] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

[32] Wilma K Olson, Andrey A Gorin, Xiang-Jun Lu, Lynette M Hock, and Victor B Zhurkin. Dna sequence-dependent deformability deduced from protein–dna crystal complexes. *Proceedings of the National Academy of Sciences*, 95(19):11163–11168, 1998.

[33] A. K. Mazur. Anharmonic torsional stiffness of DNA revealed under small external torques. *Phys. Rev. Lett.*, 105:018102, Jun 2010.

[34] Fareed Aboul-ela, Jonathan Karn, and Gabriele Varani. Structure of hiv-1 tar rna in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic acids research*, 24(20):3974–3981, 1996.

[35] Andrew C Stelzer, Aaron T Frank, Jeremy D Kratz, Michael D Swanson, Marta J Gonzalez-Hernandez, Janghyun Lee, Ioan Andricioaei, David M Markovitz, and Hashim M Al-Hashimi. Discovery of selective bioactive small molecules by targeting an rna dynamic ensemble. *Nature chemical biology*, 7(8):553–559, 2011.

[36] Evgenia N Nikolova, Gavin D Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. Probing sequence-specific dna flexibility in a-tracts and pyrimidine-purine steps by nuclear magnetic resonance 13c relaxation and molecular dynamics simulations. *Biochemistry*, 51(43):8654–8664, 2012.

[37] Catherine Musselman, Qi Zhang, Hashim Al-Hashimi, and Ioan Andricioaei. Referencing strategy for the direct comparison of nuclear magnetic resonance and molecular dynamics motional parameters in rna. *The Journal of Physical Chemistry B*, 114(2):929–939, 2009.

[38] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[39] Bernard R Brooks, Charles L Brooks, Alexander D MacKerell, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[40] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kale, and Klaus Schul-

ten. Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(16):1781–1802, 2005.

[41] Elizabeth J Denning, U Priyakumar, Lennart Nilsson, and Alexander D Mackerell. Impact of 2-hydroxyl sampling on the conformational properties of rna: Update of the charmm all-atom additive force field for rna. *Journal of computational chemistry*, 32(9):1929–1943, 2011.

[42] David A Case, Thomas E Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.

[43] Alexander D MacKerell, Nilesh Banavali, and Nicolas Foloppe. Development and current status of the charmm force field for nucleic acids. *Biopolymers*, 56(4):257–265, 2000.

[44] Eric R Henry and Attila Szabo. Influence of vibrational motion on solid state line shapes and nmr relaxation. *The Journal of chemical physics*, 82(11):4753–4761, 1985.

[45] G Marius Clore, Attila Szabo, Ad Bax, Lewis E Kay, Paul C Driscoll, and Angela M Gronenborn. Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *Journal of the American Chemical Society*, 112(12):4989–4991, 1990.

[46] Matthew A Young, G Ravishanker, and DL Beveridge. A 5-nanosecond molecular dynamics trajectory for b-dna: analysis of structure, motions, and solvation. *Biophysical journal*, 73(5):2313–2336, 1997.

[47] Richard E Dickerson, Horace R Drew, Benjamin N Conner, Richard M Wing, Albert V Fratini, Mary L Kopka, et al. The anatomy of a-, b-, and z-dna. *Science*, 216(4545):475–485, 1982.

[48] Aaron T Frank, Andrew C Stelzer, Hashim M Al-Hashimi, and Ioan Andricioaei. Constructing rna dynamical ensembles by combining md and motionally decoupled nmr rdcs: new insights into rna dynamics and adaptive ligand recognition. *Nucleic acids research*, 37(11):3670–3679, 2009.

[49] Natalie A Davis, Sangita S Majee, and Jason D Kahn. Tata box dna deformation with and without the tata box-binding protein. *Journal of molecular biology*, 291(2):249–265, 1999.

[50] Ludmila V Yakushevich. *Nonlinear physics of DNA*. John Wiley & Sons, 2006.

Chapter 3

The Microsecond

3.1 Introduction

As mentioned in the previous chapter, access to the microsecond timescale *in silica* is a largely new phenomenon. While we have had ample access to simulated and experimentally elucidated events which take around 10^{-6} seconds to occur, it is still a long and difficult process to characterize them in detail, to capture the *where* and *when* of the goings on of cellular entities at the *nanosecond* timescale, much less the *microsecond* timescale. As stated in the previous conclusion, stating/describing processes that occur for 10s of microseconds in *in terms of* nanoseconds (which in and of themselves encapsulate quite a bit of motion) would be like describing the motions that last 2.5 hours in terms of 1 second intervals. While it is possible, it is a large amount of information and needs careful attention to its implementation. It is therefore natural to investigate the microsecond simulations alone before making explicit comparisons between data derived from the both timescales. Additionally, moving to the microsecond time regime opens up a myriad of biological processes for consideration, which we should briefly discuss. Of course we must remember that it is the primary concern of this

thesis to simply *understand* something about the movement on these timescales, as carefully and meticulously as we can. After doing so, we will then attempt to cast what we have learned about one timescale in terms of the other (as in chapter 4), and then move on to example applications (as in chapter 5).

3.1.1 Microseconds: In Vivo, In Silica

For a full treatment of various biological processes and corresponding time scales the reader is referred to the introduction of the previous chapter or the excellent review by Fisette et al [1], but we will provide a short recapitulation here. We start at the fastest motion, namely the fact that hydrogens vibrate on the order of femtoseconds, which seems to coincide with some of the faster fundamental vibrational modes of small atomistic vibrations. It should be noted however, that most complex molecules (as opposed to atoms) need to be further examined by normal mode analysis or some such complimentary approach for any real accuracy of fundamental vibrational modes (and frequencies) can be established. Freely rotating protein side chains rotate diffusively about as fast as the overall tumbling of a molecule, which is generally on the order of picoseconds [2, 3]. Certain molecules, however, (particularly the heavier ones) can take much longer and the overall rate is highly dependent on solvent and thermodynamic variables [1, 4]. Hydrogen bonds are broken and formed somewhere between femtoseconds and picoseconds, meaning they are largely in some kind of equilibrium for longer processes like allosteric changes or folding events [5]. Processes that require multiple parts acting in concert take much longer, such as ligand binding, enzyme catalysis, folding, and allosteric modulations [6]. It should be noted however, that the *individual* processes making up these slower processes can happen very quickly, but the entire processes taken as a whole can take anywhere from microseconds to seconds [7]. It would behoove us to look more closely at how longer processes are not just happening in and of themselves at these times, but that the longer processes are *composed* of shorter processes separated

by probability driven events. A similar effect is observed in quantum mechanical systems that require very fast processes to wait for quite some time before relaxation or some such event, meaning the probability driven portion of the process governs the overall rate quite effectively [8]. We will return to this concept in more detail in the following chapter, but some introduction is definitely in order when considering the microsecond, and how we set up the simulations regarding them. Experiments elucidating motions on the microsecond timescale are accessible by NMR data, namely relaxation data or Residual Dipolar Couplings (RDCs). As mentioned in the previous chapter, systems were chosen such that experimental data would be available for direct comparison, although ensemble based experimental techniques such as NMR can have difficulty differentiating between members of the ensemble, which may be undergoing different kinds of motion simultaneously [9].

3.1.2 A Brief History of μs Simulations

As was previously mentioned, reliable microsecond simulations of biomacromolecules were not generally available until as recently as 2010, when this author began his time as a graduate student. The history is therefore somewhat short, but already colorful. Not the first, but certainly the most visible character in the story is the independently contracted research firm DEShaw Research who were charged with the task of increasing the computational efficiency of molecular dynamics simulations. They were able to re-envision MD computation from the ground up, taking special care to design every aspect of a specially commissioned processors optimized for MD simulation. By placing the processors in a 3 dimensional grid, they allocated the space physically in a manner similar to the way in which the simulations would be carried out [10]. This ingenuity allowed for efficient memory allocation of long range and non-bonded interaction terms and periodic movement around the simulation box, allowing for simulations to be carried out many orders of magnitude faster than previous processors, including supercomputers and dedicated GPU based simulations. Furthermore,

they optimized an in house MD code in order to eliminate numerical error aggregation during simulation, which demonstrates complete reversibility of trajectories at all times computed [11]. In short, errors which accumulate on anton are entirely errors of force field, and not numerical aggregation problems or time-reversibility problems sometimes associated with finite difference integrators. A copy of the Anton computer was recently given to the Pittsburgh Supercomputing Center, who began awarding allocations to academic researchers. Simultaneously, the team set to work tackling some of the biggest questions available to simulation and soon had a millisecond long trajectory of the BPTI protein, which showed kinetics in agreement with experiment at chemical accuracy [11]. They also computed the S^2 order parameter for the ubiquitin protein, demonstrating that accuracy increases specifically for the unbound loop region when overall tumbling motions are included in the autocorrelation function, a topic which we will return to in some detail in chapter 4 [4]. This provides one of the first moments when the deconvolution of fast and slow motions assumed by the model free approach in NMR relaxation models becomes questionable, a central thesis throughout this study [12, 13]. More specifically in the arena of nucleic acids however, Orozco et al in Spain provided the first microsecond simulations of DNA in 2007 [14]. They computed the Dickerson-dodecamer [15] for about 1 microsecond, and derived impressive classifications of what they called backbone breathing. They demonstrated that the presence of ions near the helix facilitates said breathing and furthermore demonstrated the ability of the DNA force field to maintain proper forms longer than nanosecond simulations, which had not been previously shown. To the best of our knowledge RNA has not been simulated on the microsecond timescales before now, largely because of the difficulty of RNA force fields. Non-equilibrium and experimental based approaches have begun to characterize the microsecond timescale of RNA helices, but in general little has been done directly simulating RNA at timescales longer than nanoseconds [16].

3.2 Methods

3.2.1 Microsecond Simulation Parameters

Microsecond simulations were all carried out on the Anton supercomputer, which we obtained access to by a generous grant through National Resource for Biomedical Supercomputing (NRBSC) and Pittsburgh Supercomputing Center (PSC) and DEShaw Research. HIV-1 TAR crystal structure coordinates were downloaded from the Protein Data Bank (Access Code 1ANR) to provide the starting configuration for the microsecond simulations [17]. The system was solvated in VMD [18] with TIP3 water and 27 Na+ ions to neutralize charge in a 64x64x64 Angstrom cube for initial heating. CHARMM36 force field parameters for ribonucleic acids were used, after testing of several other force fields [19, 20, 21, 22] (for more discussion about problems with RNA force fields see results and discussion of this and the following chapter, along with the introduction). System heating from 0 to 300K was carried out with harmonic constraints on backbone atoms for 100ps until restraints were gradually released over another 100ps and the system was equilibrated for 5ns. Velocities, coordinates, system, and force field parameters were transferred from initial heating run to Anton style formats and the simulation was extended on Anton for 8.2 microseconds, using the Nose-Hoover NVT integrator and with a time step of 2 femtoseconds. Coordinates were saved every 250ps (.25ns). Standard periodic boundary conditions were applied, with long range interactions calculated according to the Particle Mesh Ewald summation [23] with cut off parameter 12.99 A and a RESPA scheme of 1,1,3. To ensure accuracy, the final trajectories used were found to be predominantly in A-form by sugar pucker and inter-helical distances (data not shown). Once an initial 8 μs trajectory was generated, 10 more trajectories were generated from random snapshots within the original 8 μs trajectory, with care taken to ensure that all structures were significantly different conformation by RMSD. Additionally new ISEED values for random number generators were assigned for each new trajectory, en-

suring the resulting ensemble was made up of distinct trajectories. The resulting trajectories, although sometimes showing artifacts of force field issues, were all checked to ensure A-form was maintained for all times incorporated into the data. DNA was similarly prepared using CHARMM and VMD for initial solvation and heating, although allocation restraints did not allow for multiple trajectories to be calculated [24, 18]. As such we used A6DNA of sequence 5-CGAT6GGC-3 which has NMR data readily available and several sequence effects which can be considered due to its long Adenine-tract. Furthermore the A6DNA sequence showed the best agreement with ns based S^2 values, which will be considered in further detail in chapter 4 [25].

3.3 Results

All resulting trajectories showed satisfactory secondary structure for all simulations used except where noted below. While some structural diversity seems to be present in RNA, less is present for DNA despite some fraying effects which will be discussed further in the discussion section. RMSD traces of the initial TAR run and A6DNA trajectories are given in figures 3.1 and 3.2.

3.3.1 DNA Microsecond Simulations

A_6DNA microsecond DNA primarily retained B-form throughout simulation, despite significant fraying events which reached up to two bases at one point during the simulation. Figure 3.3 shows several snapshots of typical structures generated in the A6DNA microsecond simulations with the aforementioned fraying, and figure 3.2 shows the RMSD of the simulation after a least squares fit and periodic wrapping. The trajectory was carefully analyzed but no major rare-events were observed throughout the simulation. Backbone dynamics show

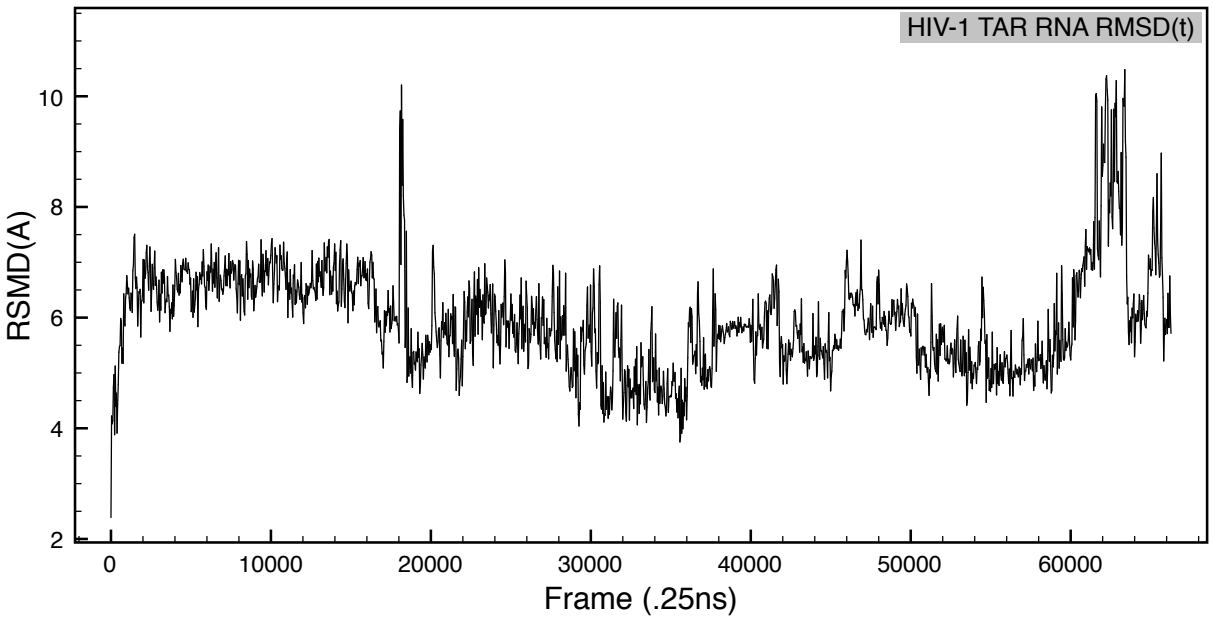

Figure 3.1: The RMSD plot as a function of time for HIV-1 TAR RNA for 15 μs run. RMSD is computed from starting structure, and shows significant structural interconversions on the microsecond timescale. Figure was generated in VMD, with .25ns per frame. Note the region post 50k frames, when the structure loses A form and degrades.

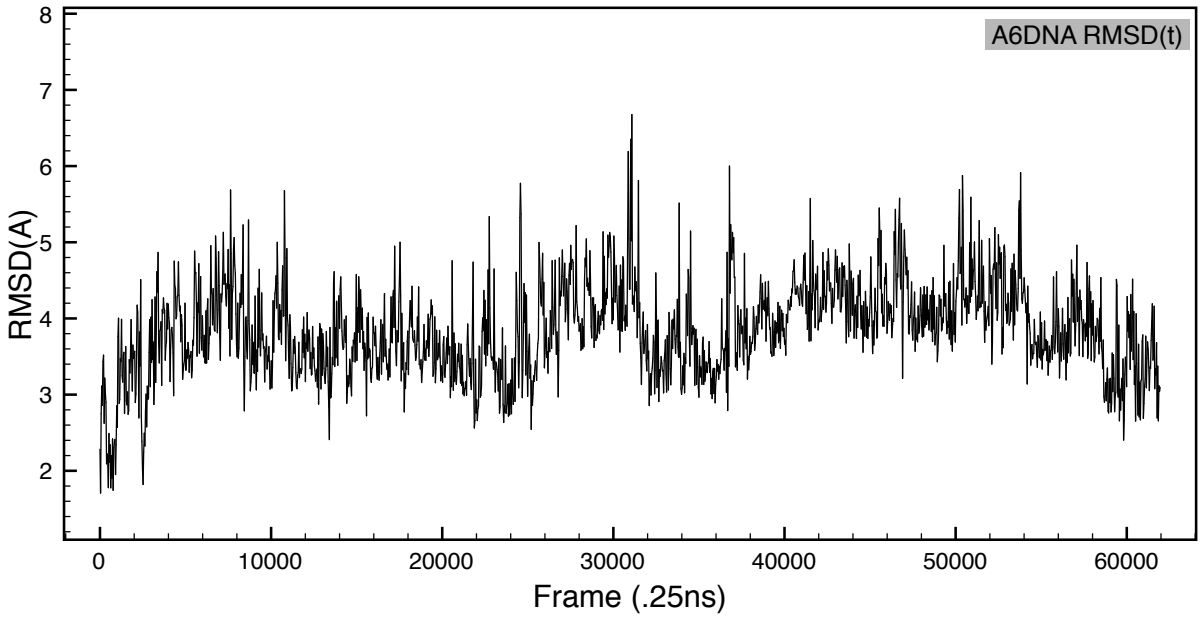

Figure 3.2: The RMSD plot as a function of time for A6DNA for 15 μs run. RMSD is computed from starting structure, and shows little structural interconversions on the microsecond timescale. Figure was generated in VMD, with .25ns per frame.

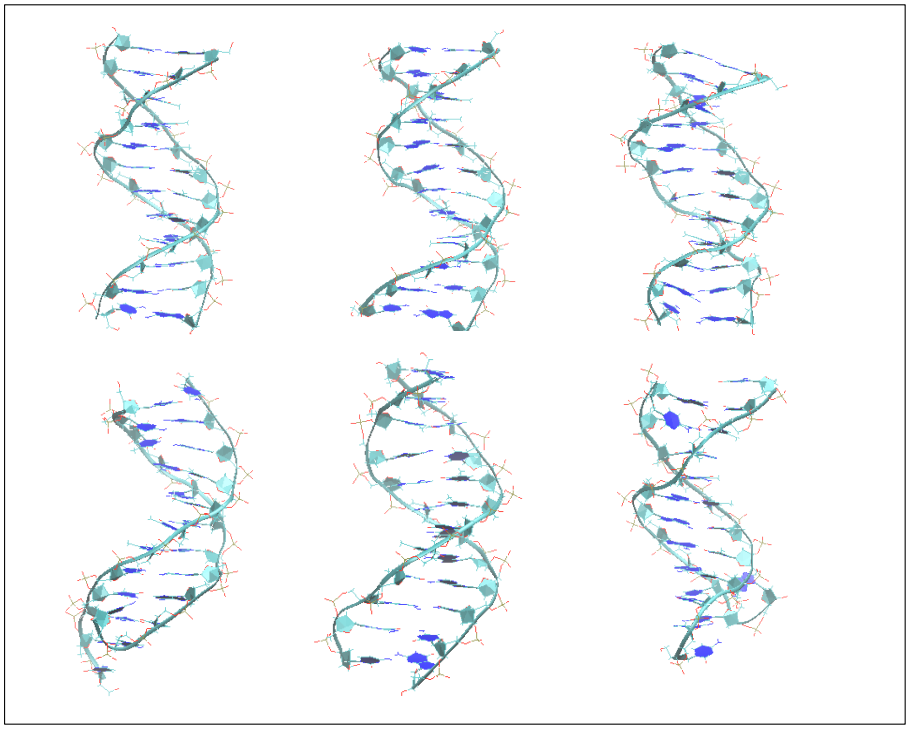

Figure 3.3: Example snapshots of A6DNA during microsecond simulations. Some fraying is observed near the end residues, which is exacerbated at long timescales possibly contributing to errors in subsequent S^2 virtual order parameters.

high fluidity, which may or may not have artifactual contributions at such long times, but S^2 order parameter agreement suggests that at least the magnitude of motions for non-frayed residues are in agreement with thermodynamic averages.

3.3.2 RNA Microsecond Simulations

TAR RNA, despite being a relatively small system, provided considerable difficulty in generation of trajectory. Initial efforts were performed using the CHARMM22 force field that quickly degraded into an unfolded structure without A form. Despite multiple attempts to implement constraints and reverse the unfolding process, no usable trajectories could be

obtained using the CHARMM22 force field. AMBER99 was also run, but within 3 microseconds the simulation adopted the notorious "ladder" structure in which the bases are still stacked and hydrogen bonded, but the backbone is no longer helical but rather linearized [22]. Finally the CHARMM36 parameters were implemented, and the resulting trajectories maintain proper A form for 8-10 μs. Further simulation degraded similarly as before, and as such subsequent simulations were all carried out under 10 μs. Examples of kept RNA simulations are given in figure 3.4.

3.4 Discussion

3.4.1 DNA Simulation Results

The results regarding A_6DNA are promising, to say the least. Never before had DNA been simulated as much as 10 μs to our knowledge, and our results show that it is robust and stable despite some possible effects of fraying which becomes exacerbated here. This allows us to tentatively validate much of the simulation work that is already being carried out regarding the interaction of DNA and proteins at 10s of nanoseconds, for example, but it does bring into question some of the abilities of non-equilibrium techniques which approximate microseconds or milliseconds with nanosecond based motions. As it has been discussed in the previous chapter (and will be discussed at length in chapter 4), there are some subtle effects of sequence that are not captured, particularly for small bases around homogenous sequences, and effects of fraying are exacerbated. The details of these effects as noted are particularly subtle for nanosecond simulations, but will likely be an important issue to scientists in the future regarding DNA [25] simulations at longer timescales. More pressing, however, is the added effects of fraying at long timescales, which seem to be additive instead of in equilibrium for all trajectories attempted. It is likely that the BI $\Leftrightarrow$ BII transition is innacruately represented

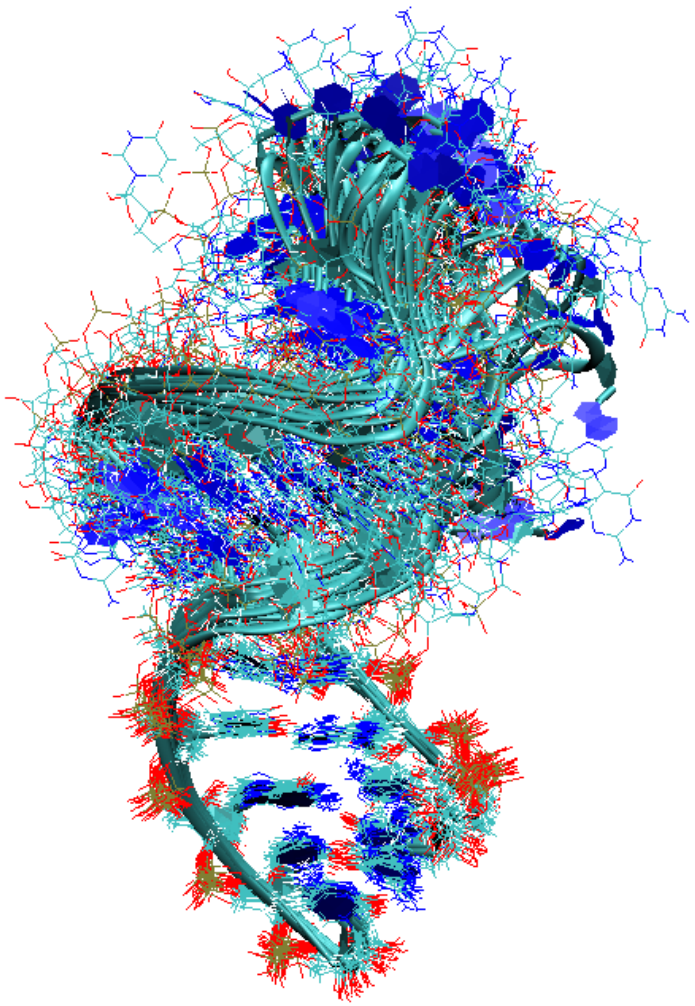

Figure 3.4: Examples of RNA structural motions during μs runs, superimposed into one image. The structure was aligned by least squares fit for the bottom helix (domain II) only. Shown is around 8μs of motion, with around 200ns per frame pictured. Note the significant amount of motion for both the loop, the hinge, and global bend around the hinge. Furthermore note that the bases intercalate and extrude much faster than the frames shown here ($\sim$100ns).

here, seeing as it is not correctly modeled for nanosecond based simulations (see [25] or appendix A for a more detailed discussion regarding this phenomenon.)

3.4.2 RNA Simulation Results

The results regarding force field problems of RNA are not surprising, but concerning. RNA dynamics lie at the heart of structural and cellular biology, and provide some of the most promising models for disease targeting, cellular reprogramming, and manipulation of genetic information on the cellular level [26]. Around the turn of the century major updates were made to the AMBER and CHARMM nucleic acid force fields that rendered them largely adequate for simulating on the nanosecond timescale [19, 20, 21, 22]. The changes that were made involved deriving new constants for backbone or sugar dihedral angles, but when simulations reached the 100s of nanoseconds deformations and problems began to show up. In response changes have been made very recently to both AMBER and CHARMM force fields, again to dihedral angle parameters. Our simulations were carried out using the latest CHARMM force field, and we have not yet tested the latest AMBER force field. It is a difficult question, whether additional parameterizations will be possible or if more detailed potentials need to be developed for highly charged residues such as RNA, but it should be obvious that as a scientific community we should make it a priority. It is quite possible that the effects are one and the same as those seen in DNA, except that errors accumulate driving degradation of the helix more quickly than in DNA simulations. If this is the case, slightly incorrect backbone dynamics are likely the culprit, despite recent changes to dihedral angle parameters. It may be beneficial to consider sequence dependent reparameterization of backbone and ribose sugar dihedrals. There are methods which attempt to overcome these shortfalls however, which rely heavily on experimental data in order to screen results of simulation. In particular, the Sample and Select (SAS) method originally described by Chen et al [27] involves using monte carlo simulated annealing to minimize a cost function

which is designed as a numeric metric for similarity between simulation and experiment. It was shown to work particularly well for elucidating TAR binding small molecules that had not been previously found [16]. In addition, the resulting ensembles can be subjected to further numerical analysis, although it is constrained by initial sample size, accuracy of experimental results, accuracy of simulation results, and saturation of sampling. The following figure shows the agreement of TAR simulation based Residual Dipolar Couplings (RDCs) from a 3ns and a 3us simulation pool, without SAS. The agreement between experimental and simulated RDCs is very poor, and it has been the center of some controversy whether this was due to problems with the force field or lack of saturation of sampling. We have demonstrated here that even though the simulation length has increased by orders of magnitude, the resulting fit decreases very little. It was later shown that the microsecond trajectories we generated here could be used to generate ensembles whose RDCs matched more closely the experimental RDCs, and the resulting ensembles were analyzed to comment on the likely mechanisms underlying the RNA movement [28]. In particular it was suggested by the resulting ensembles that global interhelical motions on the nanosecond-microsecond scale could be at least somewhat correlated to base intercalation at the bulge region, which happens on the nanosecond timescale [29].

3.5 Conclusion

Despite the large difficulties and at times large amount of criticism that nucleic acid force fields has endured, it seems to be producing real, viable results which elaborate and expound the *where* and *when* of these molecules at timescales entirely unavailable to us by any other method at this detail, despite some small changes which need to be made (and likely will be soon). It is an exciting time for simulation and molecular biology, and we are privileged to be sitting at the cusp of its implementation and nascent influence on our thinking. Of course as

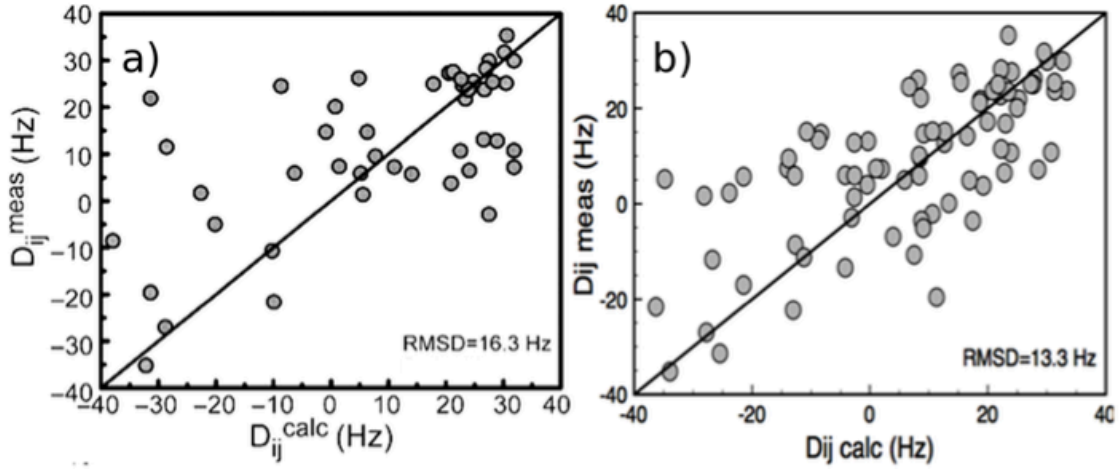

Figure 3.5: Correlation plot of experimental and calculated RDCs from a) a nanosecond based ensemble, and b) a three microsecond based ensemble. It shows clearly that discrepancy between experiment and simulation exists, despite increased simulation length, even if the simulation time is increased several orders of magnitude. Experimental RDCs and nanosecond based ensemble RDCs taken with permission from Aaron Frank and the Al-Hashimi lab [30]. It should be noted, however, that an elegant solution to fitting the above data using novel SAS based techniques has yeilded good insight into ensembles with correct RDCs. In short the above question has largely been solved, and the results are published in [28].

with any new technology or revelation we must not rush to conclusions but instead carefully tread forward resolutely, checking and double-checking the accuracy of each new result as it becomes available. It is in this spirit that we now note; we have thoroughly discussed both the nanosecond and the microsecond in detail, and so we are ready to cautiously address the issue of comparing the two, and deriving conclusions from each in reference to the other.

Bibliography

[1] 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.

[2] Måns Ehrenberg and Rudolf Rigler. Fluorescence correlation spectroscopy applied to rotational diffusion of macromolecules. *Quarterly reviews of biophysics*, 9(01):69–81, 1976.

[3] Hao Hu, Jan Hermans, and Andrew L Lee. Relating side-chain mobility in proteins to rotameric transitions: insights from molecular dynamics simulations and nmr. *Journal of biomolecular NMR*, 32(2):151–162, 2005.

[4] Paul Maragakis, Kresten Lindorff-Larsen, Michael P Eastwood, Ron O Dror, John L Klepeis, Isaiah T Arkin, Morten Ø Jensen, Huafeng Xu, Nikola Trbovic, Richard A Friesner, et al. Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *The Journal of Physical Chemistry B*, 112(19):6155–6158, 2008.

[5] George A Jeffrey and George A Jeffrey. *An introduction to hydrogen bonding*, volume 12. Oxford university press New York, 1997.

[6] Joseph R Lakowicz, Ignacy Gryczynski, Grzegorz Piszczek, Leah Tolosa, Rajesh Nair, Michael L Johnson, and Kazimierz Nowaczyk. Microsecond dynamics of biological macromolecules. *Methods in enzymology*, 323:473, 2000.

[7] Dorothee Kern and Erik RP Zuiderweg. The role of dynamics in allosteric regulation. *Current opinion in structural biology*, 13(6):748–757, 2003.

[8] Steve Guillouzic, Ivan LHeureux, and André Longtin. Rate processes in a delayed, stochastically driven, and overdamped system. *Physical Review E*, 61(5):4906, 2000.

[9] Bertil Halle. The physical basis of model-free analysis of nmr relaxation data from proteins and complex fluids. *The Journal of chemical physics*, 131(22):224507, 2009.

[10] David E Shaw, Martin M Deneroff, Ron O Dror, Jeffrey S Kuskin, Richard H Larson, John K Salmon, Cliff Young, Brannon Batson, Kevin J Bowers, Jack C Chao, et al. Anton, a special-purpose machine for molecular dynamics simulation. *ACM SIGARCH Computer Architecture News*, 35(2):1–12, 2007.

[11] David E Shaw, Ron O Dror, John K Salmon, JP Grossman, Kenneth M Mackenzie, Joseph A Bank, Cliff Young, Martin M Deneroff, Brannon Batson, Kevin J Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*, pages 1–11. IEEE, 2009.

[12] G Marius Clore, Attila Szabo, Ad Bax, Lewis E Kay, Paul C Driscoll, and Angela M Gronenborn. Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *Journal of the American Chemical Society*, 112(12):4989–4991, 1990.

[13] Catherine Musselman, Hashim M Al-Hashimi, and Ioan Andricioaei. ired analysis of tar rna reveals motional coupling, long-range correlations, and a dynamical hinge. *Biophysical journal*, 93(2):411–422, 2007.

[14] Alberto Pérez, F Javier Luque, and Modesto Orozco. Dynamics of b-dna on the microsecond time scale. *Journal of the American Chemical Society*, 129(47):14739–14745, 2007.

[15] Richard E Dickerson and Horace R Drew. Structure of a¡ i¿ b¡/i¿-dna dodecamer: Ii. influence of base sequence on helix structure. *Journal of molecular biology*, 149(4):761–786, 1981.

[16] Andrew C Stelzer, Aaron T Frank, Jeremy D Kratz, Michael D Swanson, Marta J Gonzalez-Hernandez, Janghyun Lee, Ioan Andricioaei, David M Markovitz, and Hashim M Al-Hashimi. Discovery of selective bioactive small molecules by targeting an rna dynamic ensemble. *Nature chemical biology*, 7(8):553–559, 2011.

[17] Fareed Aboul-ela, Jonathan Karn, and Gabriele Varani. Structure of hiv-1 tar rna in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic acids research*, 24(20):3974–3981, 1996.

[18] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[19] Elizabeth J Denning, U Priyakumar, Lennart Nilsson, and Alexander D Mackerell. Impact of 2-hydroxyl sampling on the conformational properties of rna: Update of the charmm all-atom additive force field for rna. *Journal of computational chemistry*, 32(9):1929–1943, 2011.

[20] Nicolas Foloppe and Alexander D MacKerell Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, 21(2):86–104, 2000.

[21] Alexander D Mackerell and Nilesh K Banavali. All-atom empirical force field for nucleic acids: Ii. application to molecular dynamics simulations of dna and rna in solution. *Journal of Computational Chemistry*, 21(2):105–120, 2000.

[22] Elzbieta Kierzek, Anna Pasternak, Karol Pasternak, Zofia Gdaniec, Ilyas Yildirim, Douglas H Turner, and Ryszard Kierzek. Contributions of stacking, preorganization, and hydrogen bonding to the thermodynamic stability of duplexes between rna and 2-o-methyl rna with locked nucleic acids. *Biochemistry*, 48(20):4377–4387, 2009.

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

[24] Alexander D MacKerell, Bernard Brooks, Charles L Brooks, Lennart Nilsson, Benoit Roux, Youngdo Won, and Martin Karplus. Charmm: the energy function and its parameterization. *Encyclopedia of computational chemistry*, 1998.

[25] Evgenia N Nikolova, Gavin D Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. Probing sequence-specific dna flexibility in a-tracts and pyrimidine-purine steps by nuclear magnetic resonance 13c relaxation and molecular dynamics simulations. *Biochemistry*, 51(43):8654–8664, 2012.

[26] Patrick P Dennis, Arina Omer, and Todd Lowe. A guided tour: small rna function in archaea. *Molecular microbiology*, 40(3):509–519, 2001.

[27] Yiwen Chen, Sharon L Campbell, and Nikolay V Dokholyan. Deciphering protein dynamics from nmr data using explicit structure sampling and selection. *Biophysical journal*, 93(7):2300–2306, 2007.

[28] Loïc Salmon, Gavin Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. A general method for constructing atomic-resolution rna ensembles using nmr residual dipolar couplings: the basis for interhelical motions revealed. *Journal of the American Chemical Society*, 135(14):5457–5466, 2013.

[29] Catherine Musselman, Qi Zhang, Hashim Al-Hashimi, and Ioan Andricioaei. Referencing strategy for the direct comparison of nuclear magnetic resonance and molecular

dynamics motional parameters in rna. *The Journal of Physical Chemistry B*, 114(2):929–939, 2009.

[30] Aaron T Frank, Andrew C Stelzer, Hashim M Al-Hashimi, and Ioan Andricioaei. Constructing rna dynamical ensembles by combining md and motionally decoupled nmr rdcs: new insights into rna dynamics and adaptive ligand recognition. *Nucleic acids research*, 37(11):3670–3679, 2009.

Chapter 4

Comparing Microsecond and Nanosecond Motions

4.1 Introduction

As has been stated previously, it is not the intention of this study to elaborate exactly what we *can* know about the movements of RNA and DNA, but what we currently *cannot* know, and where we should focus and flex our efforts for more accurate representations of the jiggles of these life-bearing molecules. We have visited the nanosecond and seen how DNA and RNA simulation can shed light on this timescale, and we have then similarly shown the same data for the microsecond, but have cautiously abstained from comparing the two or drawing any conclusions about *nanosecond* dynamics from microsecond dynamics, or *microsecond* dynamics from *nanosecond* dynamics. We have done so for a very good reason, mostly that one must think carefully about the relationship between these two hierarchically related concepts before deriving information or conclusions about one in reference to the other. These are movements that are entirely off limits to the human eyes, ears, and hands,

and as such we cannot hastily ascribe characteristics to them based on intuition until we have examined our intuition and decided that it is correctly representing the information involved.

4.1.1 Periodicity In Various Papyrii

Perhaps one of the most insightful places to steer a discussion of molecular motion is that of periodicity, or repetition within a function. Periodicity describes predictable repetition, and can be characterized in many ways. In basic physics it is regularly introduced with the concept of waves, noting that water or guitar strings alike move in this manner, repeating similar motions several times a second (or minute, or hour, and while the distinction may seem mundane, it is the point of this document to carefully point out that this very distinction is anything but trivial). In other words, one can think of a periodic motion as being *made up of* smaller sets of motions that then sum to make up the larger motion.

Take for example, a child swinging a small toy around his head by a string. The circular motion of the toy is periodic in that it often returns to its initial place in a predictable fashion. Now let us add some complexity by imagining that the child is swinging the toy *while walking in a circle*. Although the simpler periodic motion from before is still periodic (in reference to the child) the overall motion of the toy now may not return it to its initial position in the same predictable way (in reference to us). We may say in this case that the motion of the toy is still periodic, but the situation has clearly changed. Now let us imagine that our only method of measuring the toy's motion is via a strobe light, which can only flash at some discrete interval. At the slowest time interval allowed (aka the light flashes only once at the beginning and once at the end of the toy's journey), we might reasonably deduce that there is no movement at all! If we were able to then increase the frequency of our strobe so that it flashed three times during the walk, we might deduce that the toy

simply moved the length that the child walked, and then returned to its initial position, meaning that it has only one low frequency period (the time it took the child to get back to his starting position). Upon increasing the periodicity (or frame rate) of the strobe yet again, we will finally begin to see irregularities in the path as opposed to what we could see with the previously visible points, and only with careful analysis might we conclude that the initial period was actually *made up of* smaller periodic motions which gave rise to the final period.

This analogy is very similar to the current picture we have of atoms moving around in solution; experimental and simulated techniques alike are generally constrained to a "snapshot" based resolution which is often much slower (or in other cases much faster!) than the fundamental vibrational modes which likely make up the total motions in question [1, 2]. In order to instill the correct image however, one must note that (while it is definitely correct to say that microseconds are made up of nanoseconds) it is misleading to say that microsecond based motions can be described *efficiently* by nanosecond based motions. There are 10,000 nanoseconds in 10 microseconds (the average length of our microsecond simulations), which is roughly analogous to stating that a motion with a duration of 2.5 hours (let us say for example how long someone might take to walk about 7.5 miles) is made up of 10,000 1 second long moments (a stride lasting about one meter, let's say). While the comparison is a powerful pedagogical tool, interpreting it too literally in anything less than an ideal situation will undoubtedly fall short of proper communication. Namely, it would be overly simplistic (or perhaps just missing the point) to state that you were able to walk 7.5 miles simply by "taking ten thousand steps." The analogy breaks down even more when we realize that much of the motion is periodic, or in other words you could just as well state that you travelled zero *total* miles (remember the boy walking in a circle) by taking ten thousand steps in succession. We would need to know about how the steps are strung together; in what order they were taken, and which other forces were involved to fully describe the situation.

As complicated as the discussion of periodicity has become, there is one factor that complicates it still further. Returning to our child swinging a toy analogy, consider the non-trivial case in which the movement of the toy has a considerable effect on the period of the *childs* motion. While this may not be very likely for any of our examples so far, one could just as easily envision that the child is swinging a much heavier object such as a bowling ball. Additionally lets add that instead of walking calmly in a circle, he is performing a somersault. It is not just that the motion of the bowling ball that affects the rate and way in which the child will perform the tumble, but the *periodicity* of the bowling ball's motion also affects the overall rate and inertia of the somersault. This is actually a much more apt analogy to molecular motion in that a given molecule is subject to both internal and external forces, quantum mechanical forces governing periodicity of bond motions (think of the childs arm or the string being akin to a the virtual spring we use to model bonds or angles) and non-bonded forces (which for all intensive purposes resemble gravity in our little thought experiment). The assumption that the convolution of these fast and slow motions can be ignored, while a good assumption for most globular proteins and highly ordered macromolecules, begins to break down for longer timescales and highly disordered (or unfolded) entities [3, 4, 5]. It is the focus of this thesis to claim that RNA structures in general suffer from this problem, and its effects are currently being felt in the realm of molecular dynamics simulations. Returning to our example involving a 7.5 mile walk, this complicated situation describes the condition in which the time it took you to walk the mile was not only dependent on the time it takes to move your foot one meter, *but the time it takes you to move your foot one meter depends on how long it takes you to walk 7.5 miles.* While this seems counterintuitive for linear motion, it is anything but for periodic motions nested hierarchically inside one another.

4.1.2 The S^2 Revisited

Nuclear Magnetic Resonance has clever ways of exploiting this periodicity, however, and it is to the S^2 order parameter that we now turn our discussion. NMR relaxation data is often characterized by two parameters, the spin-lattice relaxation constant T_1 and the spin-spin relaxation constant T_2; both of which are derived from refined magnetic field pulse experiments. By using a strong magnetic field to perturb a controlled environment of molecules and measuring the time it takes for them to return to their initial state, (plus the de-phasing that occurs) [6], one can reliably probe periodicity as we have described it above (it would be loosely akin to hitting the toy in the childs hand and waiting for it to come back to where it started in order to determine the period). Used in conjunction with the Nuclear Overhauser Effect (NOEs) that determines distances between specific atoms, many important characterizations and structural elucidations have been reported [7] which are regularly used synergistically with molecular simulation techniques. Generally these measurements are carried out in nucleic acids by inserting carbon-13 or nitrogen-15 isotopes into the nucleic acid structures of interest at specific locations, and in this way single bond vectors can be perturbed and measured without introducing artifacts into the structural dynamics of the molecules.

More specifically, the orientation of the external magnetic field $\bar{B}$ is pulsed such that the time it takes for the bond vector $\hat{\mu}$ to return to its initial orientation within the magnets XY plane, denoted as $\bar{M}_{xy}$ (aka perpendicular to the vector describing the magnetic field $\bar{M}_z$) gives the *spin-lattice* relaxation constant T_1. While the spins of the nuclei flip out of the $\bar{M}_{xy}$ plane, timing the resulting dephasing gives rise to the *spin-spin* relaxation constant T_2. In terms of the spectral density function, $J(\omega)$, T_1 and T_2 are given by

$$T_1^{-1} = c[J(\omega_c - \omega_H) + 3J(\omega_c) + 6(\omega_c + \omega_H)] \tag{4.1}$$

$$T_2^{-1} = c[eJ(0) + J(\omega_c - \omega_H) + 3J(\omega_c) + 6J(\omega_H) + 6J(\omega_c + \omega_H)] \tag{4.2}$$

Where the spectral density function is the amount that each frequency contributes to the total spectrum, or:

$$J(\omega) = s \int_0^\infty \cos(\omega t) C(t) dt \tag{4.3}$$

where $C(t)$ is the correlation function for a given bond vector (usually $^{13}C - H$ or $^{15}N - H$ for our purposes). By this method the experimentally derivable quantities T_1 and T_2 can be related to the actual motions of bond vectors of interest by assuming the decoupling of internal and overall motions (tumbling versus internal motions, or swings of the toy versus the walking of the boy). In other words, the fast versus slow motions which we assume must exist can be decoupled such that C(t) can be expressed as:

$$C(t) = C_f(t) C_s(t) \tag{4.4}$$

where f and s subscripts refer to fast and slow motions respectively. For proteins and most globular molecular entities this slow tumbling $C_s(t)$ can be modeled simply by:

$$C_s(t) = (\frac{e^{-t/\tau_M}}{5}) \tag{4.5}$$

and the internal or fast motions are given by the second order legendre polynomial of the bond vector $\hat{\mu}(t)$ which is written:

$$C_f(t) = \langle P_2[\hat{\mu}(0) \cdot \hat{\mu}(t)] \rangle \tag{4.6}$$

Here we use the second order legendre polynomial because of its geometric description of a bond vector carving out a sphere [8]. If instead we were interested in autocorrelation analysis on a single variable time series, we would use $\langle \hat{\mu}(0).\hat{\mu}(t) \rangle$. Lipari and Szabo were the first to develop a formalism in which the above autocorrelation function is parameterized using relaxation data, and assumed no particular motional model in their formalism, called the model free approach [3, 8]. In this approach, which has seen widespread acceptance across several disciplines, $C_f(t)$ is given as :

$$C_f(t) = S^2 + (1 - S^2)e^{-t/\tau_f} \tag{4.7}$$

Which finally introduces the generalized order parameter S^2, the plateau of the internal $C(t)$ function. As it is given in the introduction, S^2 can also encode fast motions in further parameterizations that yield slightly different functional forms. In short, the S^2 measurement tells us about the overall amount of motion of a given bond vector, and is easily calculable from molecular dynamics trajectories for comparison with experimentally derived S^2 order

parameters. It is the focus of this study to show that this simple measure of motion reveals something both novel and interesting about the modes available to RNA and DNA as we move from nanosecond to microsecond timescales, namely that the separation of fast and slow motions is problematic for RNA, but not as much for DNA. A further treatment is in order, however, to understand the autocorrelation function in and of itself. Let us return to the example given above, namely that of the boy spinning a toy above his head. A more explicit form of writing the autocorrelation function (for a discrete process with known variance, as is the case for time series within MD simulations) can also be given as:

$$C(t) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{(n-k)} (X_t - \mu)(X_{(t+k)} - \mu) \tag{4.8}$$

where k is lag time (the length to which we carry out the calculation), n is the number of points in the data set, σ is the variance of the data set and μ is the mean value about which the time series is fluctuating. Typically max lag times are reported for about 1/8 to 1/10 of the total data set, due to instability arising from sparse data sets at long lag times. One can imagine rewriting the motions of our childs toy as a *time series* in which position (or distance to origin perhaps) is plotted as a function of time. If the child was interested in doing a series of complicated dances, the time series arising from it could be averaged with themselves at varying times, leading us to see how many of the motions are only represented as a measure of correlation between periods. Keep in mind that the final place to which the resulting autocorrelation function stabilizes is the measurement we can glean from the NMR data for direct comparison, and so there is an inherent inability to de-convolute the S^2 order parameter from various types of periodicities' contribution to the motion metric.

It is not the intention of this study to give a complete characterization of the effects of different types of periodicities herein, only to point out that several types of convolution can

occur, and the S^2 order parameter only tells us about *aggregate* motions. It is important to keep in mind therefore, that the final S^2 values that we derive are not *direct* measures of magnitude of motion, but *derived* measures of the correlation of motion with itself as a function of time. A stark reminder of this comes from the recent publication by Zieske et al at Columbia University, who demonstrated that S^2 order parameters can overly represent slow conformational changes when they are being accessed on the microsecond timescale, but the analysis was shown only for proteins [4]. It is clear, however, that by comparing motions present at one timescale to another timescale we can deduce something about the kinds of *nanosecond*-based motions that are making up the various *microsecond*-based motions.

4.2 Methods

4.2.1 Calculation of Microsecond S^2 Order Parameter

While the derivation of the relationship of the S^2 to experimental values is complicated and somewhat nuanced (see above), the calculation of the S^2 order parameter from molecular dynamics trajectories is straightforward. After generation of trajectories (for simulation parameters and such see chapters 2-3), the P_2 autocorrelation functions were written in CHARMM [9] and several examples are presented below in figure 4.1. After printing autocorrelation functions and analyzing them for stable parameters within reasonable bounds autocorrelation tails can be averaged to calculate the S^2 as shown in figure 4.2.

Finally all values were averaged for each bond vector across all trajectories where applicable. Furthermore, we also tested the equilibrium version of the S^2 calculation given by:

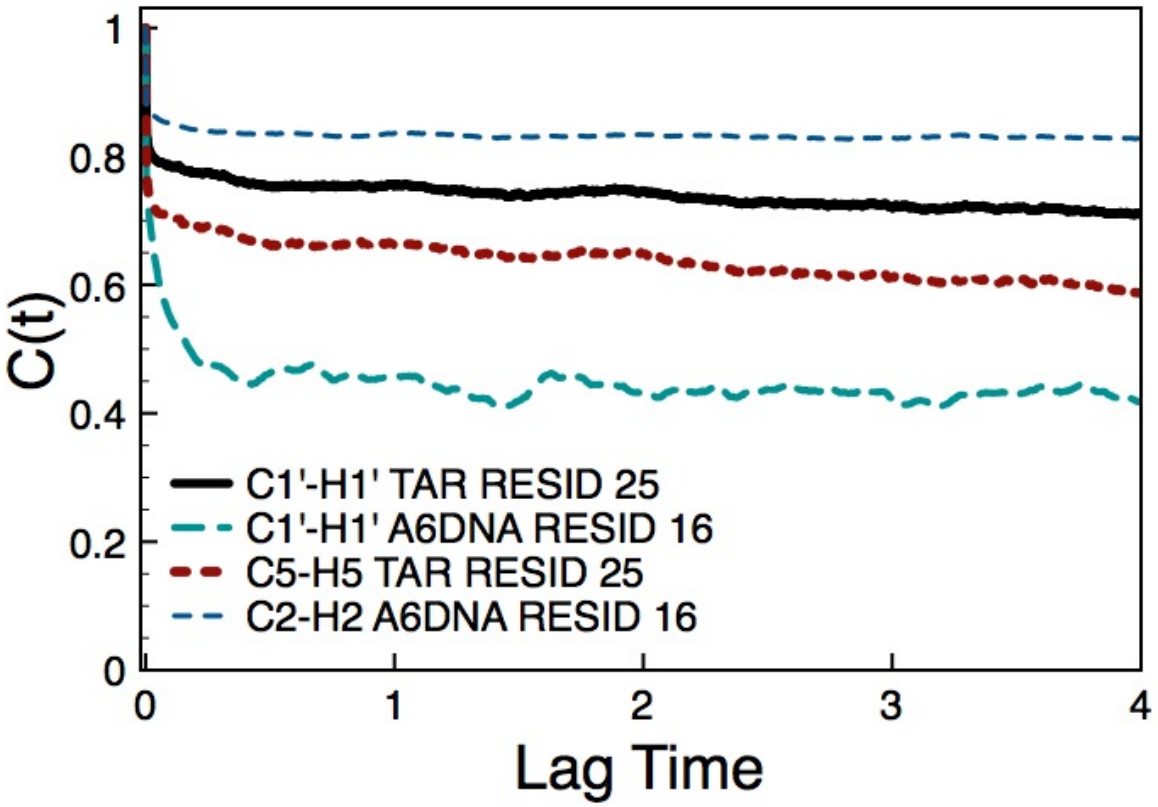

Figure 4.1: Examples of the autocorrelation functions of backbone and bases in A6DNA and TAR RNA. The x-axis is lag time in microseconds. Autocorrelation functions were calculated from residue 25 and 16 from TAR and A6DNA respectively, for the C1'-H1' backbone bond vector and the C5-H5 or C2 -H2 bond vectors from TAR and A_6DNA respectively.

$$S^2_{eq} = \frac{\langle 1/r^3 \rangle^2}{\langle 1/r^6 \rangle} \left[\frac{3}{2} (\langle \hat{\mu}_x^2 \rangle^2 + \langle \hat{\mu}_y^2 \rangle^2 + \langle \hat{\mu}_z^2 \rangle^2) + (\langle \hat{\mu}_x \hat{\mu}_y \rangle^2 + \langle \hat{\mu}_z \hat{\mu}_y \rangle^2) - \frac{1}{2} \right] \tag{4.9}$$

As further elaborated in Musselman et al [10, 11]. It was found that the values were in good agreement with the model-free approach, and so only the model-free approach data is reported here. All experimental values were taken from the previously mentioned data sets [10, 12].

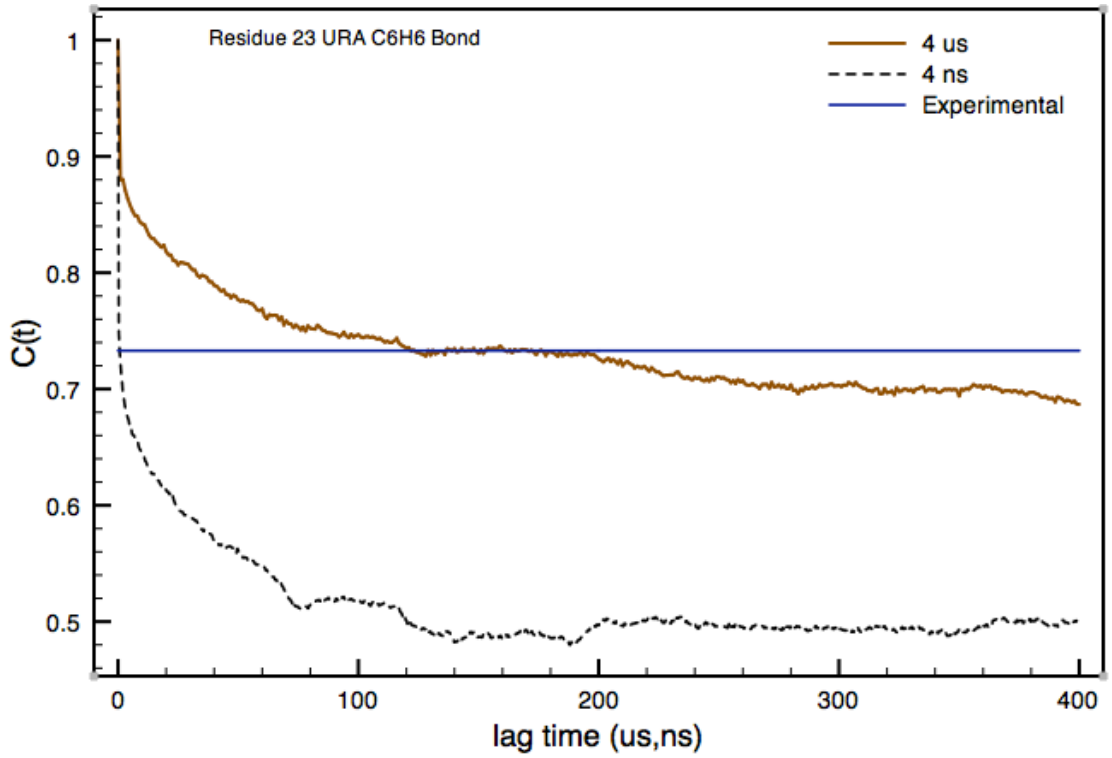

Figure 4.2: Demonstrating the relationship between experimental S^2 and the simulated S^2 by direct averaging of the autocorrelation function tail. Shown is the autocorrelation function for the residue 23 uracil C6-H6 bond vector. The experimental S^2 value is shown as a straight line. Averaging on the autocorrelation function tail is carried out from some value after the decay time and before 1/10th of the data set. Above is shown an example of increased agreement between RNA nanosecond and microsecond autocorrelation functions. Increased RNA autocorrelation coefficient (S^2) indicates that this bond vector was trapped in a local minimum during nanosecond simulations.

4.3 Results

4.3.1 DNA Microsecond S^2 Order Parameter

DNA S^2 order parameter agreement with experiment is reported in figure 4.3, which shows experimental S^2 order parameters in red, with microsecond S^2 order parameters in blue and nanosecond S^2 order parameters in green. The top portion of the figure shows S^2 order parameters for backbone bond vectors C1-H1, while the bottom half of the figure shows S^2 order parameters for bond vectors located on bases; namely C2-H2, C8-H8, and N1-H21 for purines and C5-H5 and C6-H6 for pyrimidines. (For more explicit description of locations of bases see figure 4.4). While the overall "frown" profile is in good agreement with experimental results, the microsecond S^2 order parameters have decreased overall, and fraying events are exacerbated. Particular attention should be paid to the left half of the backbone parameters near adenine rich sequences from residues 20-24, which show significantly more motion that the other residues. Visualizing the simulation shows these residues were affected by a two base level fraying event. This is partially visible in figure 3.3.

4.3.2 RNA Microsecond S^2 Order Parameter

S^2 order parameters for RNA are given in figure 4.5. Similar to figure 4.3 experimental S^2 order parameters are presented in red, while microsecond S^2 order parameters are reported in blue and nanosecond S^2 order parameters in green. The various positions of the residues along the secondary structure are color coded to demonstrate major differences between the loop and bulge regions as opposed to the two A like helices. The top portion of the figure shows S^2 order parameters for backbone bond vectors, while the bottom half of the figure shows bond vectors located on base moieties similar to figure 4.5. Furthermore, in figure 4.6 base located bond vector S^2 order parameters are separated into the larger purines

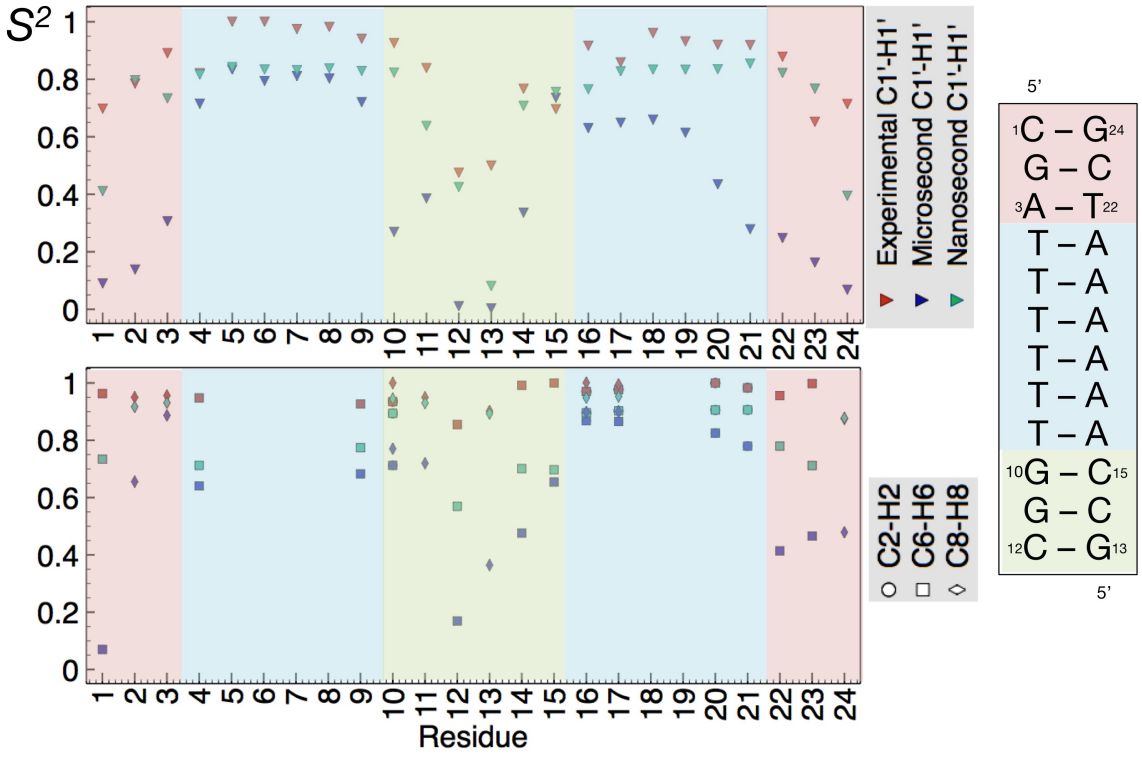

Figure 4.3: Experimental and virtually derived S^2 order parameters for A_6DNA. Experimental results are shown in red, while microsecond S^2 order parameters are in blue, and nanosecond based S^2order parameters are in green. The top plot shows backbone values, namely the C1'-H1' bond vectors, while the bottom shows the order parameter for bond-vectors located on bases. The residue number is shown to the right by schematic and the AT rich region is highlighted in blue while flanking regions are in red or green. In general the parameter shows the overall "frown" profile common to S^2 order parameters for folded macromolecules, but microsecond results show decreased periodicity for adenines, and effects of fraying are exacerbated between microsecond and nanosecond data.

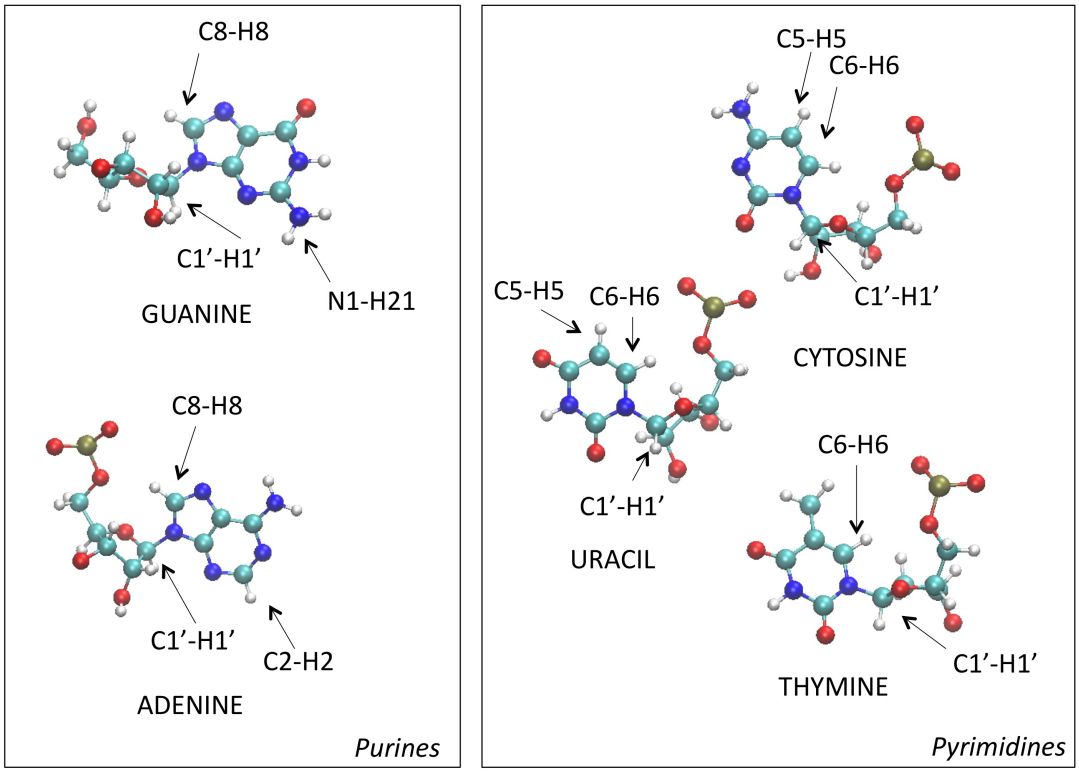

Figure 4.4: Assignment of bond vector labels, separated by purine vs pyrimidine. Labeling of bond vectors and residues follow standard conventions.

adenine and guanine, and the smaller pyrimidines uracil and cytosine. RNA microsecond based S^2 order parameters show significant improvement from the nanosecond based S^2 order parameters on almost every residue, including bases near fraying regions. Finally, calculated and virtually derived S^2 order parameters are given in tabular form for A_6DNA and TAR RNA, separated by structure and ribose/base bond-vectors at the end of the chapter.

4.4 Discussion

4.4.1 Impacts of DNA μs Based S^2 Order Parameter and Implications for DNA Force Fields

As was previously discussed, DNA simulations by MD are relatively good at recreating experimental parameters from femtoseconds to nanoseconds, but are not entirely accurate, particularly around areas of homogenous sequences that exhibit tertiary effects. While this level of accuracy is very useful for a myriad of important applications and even a fundamental understanding of the molecule that encodes our genetic memory, it lacks the ability to describe one of the most fundamental and yet widely misunderstood aspects of the role of DNA in life. Essentially what the effects of indeterminate motion around specific sequences and tertiary effects of sequence implies is that while the code itself seems to be the way that information is encapsulated, it is not sufficient in and of itself to accomplish the task that we so often ascribe to it, as mentioned earlier, *there is more to genetic encoding than the code itself.* What our nanosecond scale data implies is that the code can effect subtle changes in biophysical parameters and therefore global motion, and that global motions can affect the reading of the code. This hints at a second tier genetic code involving the sets of motions that govern DNA interaction with molecular machinery.

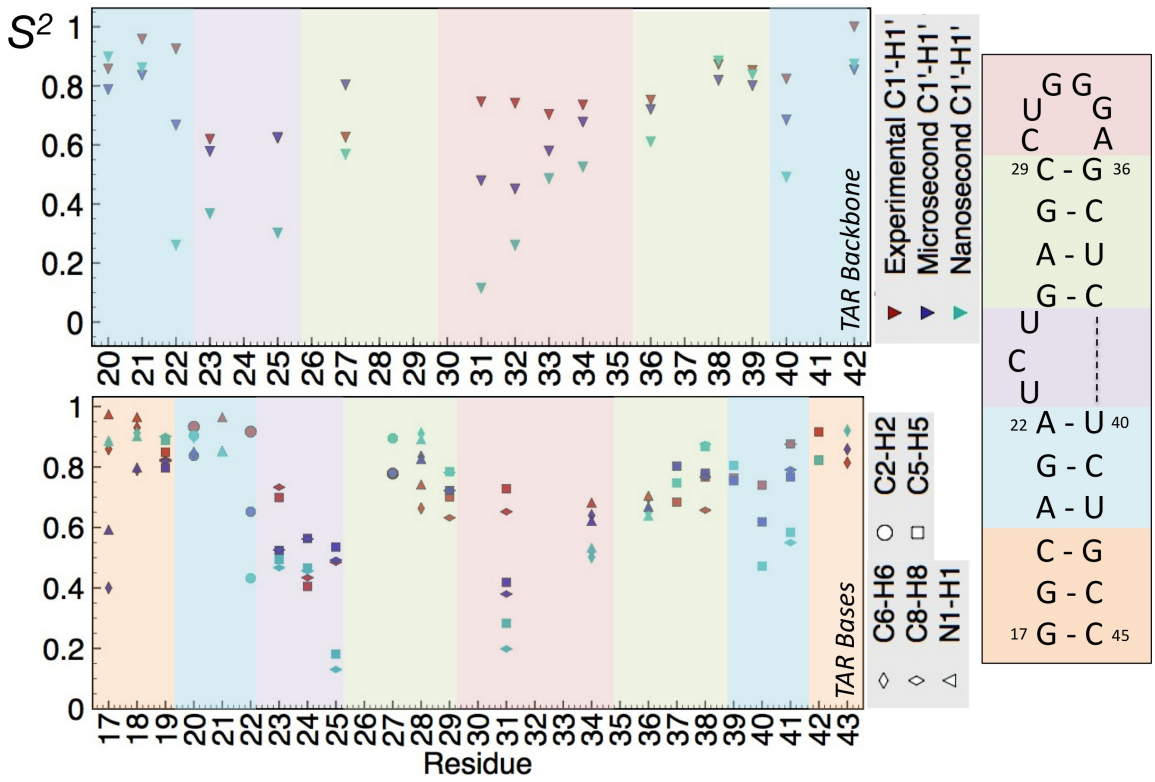

Figure 4.5: Experimental and virtually derived S^2 order parameters for HIV-1 TAR RNA. Experimental results are shown in red, while microsecond S^2 order parameters are in blue, and nanosecond based S^2 order parameters are shown in green. The top plot shows backbone values, namely the C1'-H1' bond vectors, while the bottom shows the order parameter for bond-vectors located on bases. Interestingly, we here see increased agreement between microsecond and nanosecond based virtual S^2 order parameters, but by bi-directional movement of the data from nanosecond to microsecond results. This is evidence of the drastically different types of movement that RNA and DNA are demonstrating through simulation at the microsecond timescale, a central thesis to this study.

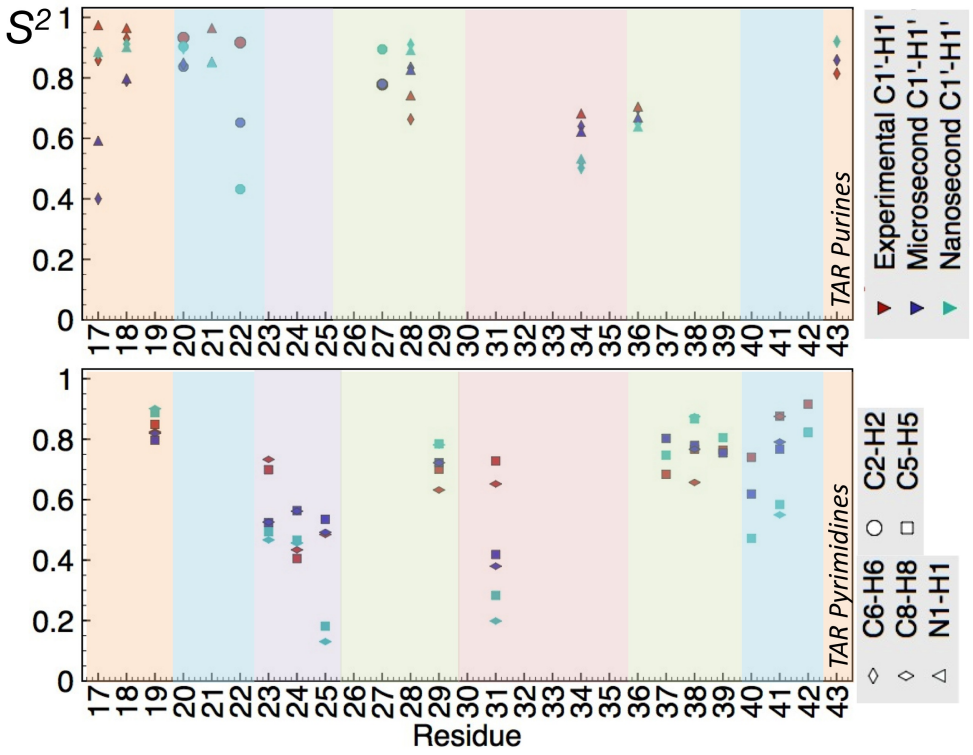

Figure 4.6: Similar to figure 4.5 we here show virtual and experimental S^2 order parameters values derived from nanosecond and microsecond ensembles on base moities, but here we have separated purines and pyrimidines for clarity. The position of each type of bond-vector on the bases themselves is shown in figure 4.4.

This point is even more poignantly driven home when we extend our discussion to the microsecond dynamics. We did not expect to see globally inaccessible motion regimes to become unlocked by moving into longer timescales for DNA, we suspected that DNA would continue to move in ways that were well documented and understood by Orozco et al [13] which would furthermore comment on the *function* of the molecule. What we see here through careful examination of both physical trajectory files and the resulting S^2 order parameters is that while DNA tends to continue moving around in familiar ways, it is sampled much more exhaustively, and hence we see the entire S^2 profile decrease in a coherent way. The main message here is that DNA seems to be relaxing inside a relatively stable well whose relaxation time is somewhere between nanoseconds and microseconds.

The more important aspect of the previous discussion, however, is that the inaccuracy of the DNA force field noted in chapter 2 discussions is exacerbated here. Residues 20-24 show significantly decreased S^2 order parameters, insinuating that they are moving around too much. It may be that some of the effects noticed by Zeiske et al in [4] are at play, namely that adenines are accessing some conformational states slowly and sparsely but that the transition is overrepresented in the final order parameters calculated here, but it seems more likely that the fraying event noted in simulation is manifesting here as increased motion. While the errors in force field associated with fraying may seem small and largely unimportant for processes happening with nanosecond duration, simulating processes that take longer will undoubtedly suffer from the inability of DNA to be simulated accurately around the ends of the duplex. It seems likely that some tweaking to the parameters (particularly the larger purines) needs to occur before we can move forward confidently into microsecond simulation. If so, this seems reasonably accomplishable with some hard work and deeper understanding; while it is not the job nor expertise of the authors herein to suggest specific changes to parameterization it would seem that some investigation into sequence sensitivity should be initiated, perhaps leading to a new set of parameters which accounts for each combinatorial instance of neighboring sequence possibility. It also seems likely that changes

to backbone dihedrals and stacking potentials might benefit the most from explicit sequence dependence, but that is something that will have to be determined by systematic experiment and simulation [14, 15, 16, 17].

It is important to note, however, that we are only commenting on naked ds-B form DNA in a stable environment. One of the difficulties of working at such small distance and time scales is the inherent dependence of each trajectory on a set of wildly fluctuating parameters, and the difficulty of working in the biological regime is that we have a wide array of biologically relevant situations and circumstances which are of interest to varying fields for varying reasons. In short, the previous discussion does not necessarily apply to higher order DNA structures such as triplexes or quadruplexes, Z-form or A-form DNA, or situations in which seemingly small changes have been introduced in the environment of the molecule without enough time for equilibrium to be reached. Perhaps more interesting, however, is to note that large-scale conformational changes do not seem to be *spontaneous* at room temperatures and neutral pH. This is definitely not the case for DNA bound to protein, and it seems likely that the "lock and key" type of thinking for proteins and DNA interaction is somewhat outdated. At the risk of over interpreting the current data, perhaps we should consider that the complicated interaction between DNA and its environment looks like a *correlated* phenomenon between several complicating factors, as opposed to some initial conformational change which is spontaneously adopted that then "recruits" proteins to it. This is only speculation at this point, but it seems an important possibility to consider seeing as this move has happened already in the field of protein dynamics and function [18, 19, 20] with great success.

4.4.2 Impacts of RNA μs Based S^2 Order Parameter and Implications for RNA Force Fields

The results regarding the HIV-1 TAR RNA, while in better agreement than previously reported nanosecond S^2 order parameters [10, 21], provide for a much more difficult discussion (or subsequent insight) than the previous discussion for two primary reasons. Firstly, *RNA simulation inevitably loses A form at times larger than 10 μs*, and secondly, the S^2 profile from nanosecond to microsecond *does not maintain overall form.*

The first point is both troubling and perplexing. In one interpretation, the agreement with experiment is better and the simulations must be providing a genuine picture of microsecond motions, but in another interpretation, how can we trust data derived from simulations that are obviously accruing some sort of error? The question then becomes one of when or where the errors accrue. If the simulation is fundamentally flawed, or in other words if the simulation is wrong due to the inherent inability of the simulation to capture RNA motions, then we shouldnt see any agreement at all, or we should be able to see some *systematic* disagreement from simulation data, whereas it seems that as long as A-form is maintained then we have good agreement whereas if A-form is not maintained we see poor agreement. If one attempts to attribute the degradation event to some erroneous motion that has become available to RNA motion only at long timescales then we run the dangerous course of searching inevitably through an endless search space for culprits which are clearly only symptomatic of larger problems with simulations. In contrast, however, if we do not attribute it to some motion allowed for within the model but to the fundamental nature of the model itself it will feel difficult to trust any of the microsecond simulations; a logical conclusion that is negated by the excellent agreement of the data with experiment. In short, we dont know why the simulated RNA structures fall apart, but it would be in our interest to elucidate the cause as quickly as possible, seeing as it may come at the cost of loss of confidence in our simulations (or aspects thereof). It is the opinion of this author that

the only way to know (or even begin to systematically approach deriving solutions to this problem) lies in simulating more RNA systems at longer timescales.

A more exciting aspect of this difficult discussion however, is that bi-directional adjustment of S^2 order parameters suggests something both novel and exciting for RNA as a molecular species. Recall the introduction to this chapter regarding the example of a child and his precessing toy viewed by a strobe-illuminated camera. There exists many complicated deviations from this simple motion, and we could envision how several of them would result in convoluted and largely un-resolvable trajectories with the same overall magnitude of motion. Upon calculating the relaxation parameters for such motions one might rightfully decide that the problem is not worth the headache of elucidating the specific motions, and search for another metric. If we were to take an ensemble of the toys, however, and calculate all the relaxation parameters we might find something much more interesting, namely that as we incorporate more and more *time* into the ensemble (aka longer and longer complicated precessions per toy) we could look at the how the resulting S^2 coefficients move *relative to each other*. Such an analysis may not tell us the specifics about the internal motions, but it could tell us if the time dependence of the slowest modes of each ensemble member are bigger or smaller than the slowest modes of another member of the ensemble. What we see in our RNA microsecond data as compared to our nanosecond data, is that (unlike DNA) many of the bond vectors seem to be precessing in ways that have different periods *relative to each other,* facilitating the bi-directional movement of the order parameters. in other words, letting the simulation run longer actually *decreases* the total magnitude of RNA motion, a complicated result that only makes sense if we incorporate our careful understanding of the role of periodicity therein. In other words, it may be that DNA has reached ergodicity somewhere between nanoseconds and microseconds, while RNA has not [22].

This analysis is not entirely novel, and stands in good agreement with the scientific literature. Maragakis et al [23] suggested much the same analysis for the amide backbone

loop residues for the ubiquitin protein, and Halle discusses at length a similar analysis for allosteric modulations or any case when local motions considerably alter the shape of the overall molecule (think of a diver tucking for a flip)[2]. Prompers and Bruschweiller designed a method in 1995 which can evaluate the statistical separability of timescales which they called isotropic Reorientational Eigenmode Dynamics (iRED) which was then used to show that another small RNA molecule (the iron response element) exhibited the same effects we are positing here [24], while specifically elucidating *which* motions are non-separable at which timescales. The same method was later applied to TAR RNA elucidating the major structural motions that are the culprit for our strange RNA behavior, specifically that a global "hinge," accounts for the large tertiary fluctuations, a result that agrees well with the microsecond simulations [5]. Furthermore, combined with recent work by Salmon et al [25], it was reported that base intercalation at the bulge is likely responsible for facilitating the large conformational fluctuations around said hinge. It seems quite plausible that small, non-globular RNA molecules with bulges, loops, or denatured regions are likely to exhibit this same character. If this is true it would have profound implications for the "lock and key" type models that have dominated thinking about small ligand binding behavior in the past [26]. Specifically, it would suggest that RNA is likely to visit many conformations fluidly, and that those conformations are being accessed aperiodically on our timescales, meaning that RNA is more the "key than the lock in diffusively governed environments, which again, is not a new idea [27]. Perhaps more importantly is the alternative methods of RNA manipulation this suggests, specifically that we should also be interested in the *correlation* of RNA structural motifs with surrounding environmental variables as opposed to the crude act of designing of agents which can unequivocally bind and lock onto the target. This is reminiscent of many tactics taken for treatment of highly mutagenic diseases such as HIV and cancer in which some of the best efforts have come not from approaches designed at *causally* eradicating the life cycle of the problem (by direct disruption at one single point),

but simply *guiding* it away from its normal life cycle with several points of attack so to speak [28]. This leads nicely to the next topic of discussion regarding RNA diffusive motion.

It should be apparent by now that non-saturation of RNA motions will undoubtedly be important to the field of biomedical description (and by extension biomedical engineering). It has been discussed at length that non-separability of internal and overall motions would introduce systematic errors to the calculated NMR order parameters observed in experiment which are generally treated with the model free approach [19], and it is not impossible that this simple fact explains much of the above discussed paradox; while the NMR order parameters may be systematically introducing biases, using the model free formalism to calculate virtual order parameters would also succumb to this bias, rendering the agreement good but the force field still subject to spontaneous failure if and when said errors accrue enough to bias the overall simulation towards an erroneous structure. We suspect, however, that something a bit more subtle is going on, and only careful investigation will solve the issue unequivocally.

Finally, we should now return to our discussion started in the opening lines of this dissertation, namely the concept of RNA as the "center" of functional life in lieu of DNA. If life truly originated with RNA, then RNA (in some form or another) should be able to functionally enact all the processes necessary for cellular survival (which we have indeed seen demonstrated time and time again), but specifically in such a way as to maintain information transfer simultaneously. This suggests that RNA is literally at the "center" of functional timescale access *and* functional information storage, and we should see continuous motions along all the biologically relevant timescales. Seeing as the nanosecond to microsecond time scaling sits at the center of biologically functional motions (catalysis to diffusion), we should see *smooth, continual* motion (whether adaptive or intrinsic) of RNA along these time-scales, which is precisely what our data suggests.

Residue	Bond Vector	Virtual S^2 (μs)	Virtual S^2 (ns)	Experimental S^2
20	C1'-H1'	0.787115	0.897902	0.857
21	C1'-H1'	0.83553	0.861186	0.958
22	C1'-H1'	0.667192	0.260621	0.925
23	C1'-H1'	0.577929	0.367252	0.619
25	C1'-H1'	0.62499	0.301328	0.622
27	C1'-H1'	0.803416	0.567939	0.626
31	C1'-H1'	0.478479	0.114713	0.746
32	C1'-H1'	0.450564	0.260003	0.741
33	C1'-H1'	0.578925	0.485641	0.703
34	C1'-H1'	0.676695	0.525023	0.735
36	C1'-H1'	0.719559	0.610271	0.752
38	C1'-H1'	0.818932	0.884319	0.872
39	C1'-H1'	0.800672	0.838505	0.852
40	C1'-H1'	0.684617	0.490945	0.823
42	C1'-H1'	0.853886	0.873422	1

Table 4.1: Experimental and virtual (both μs and ns based) S^2 order parameters for HIV-1 TAR RNA ribose moieties, taken with permission from Musselman et al [10] listed by both residue and bondvector type. For residue indices see figure 4.5.

4.5 Conclusion

Now that we have thoroughly examined to the best of our abilities the *where* (position) and *when* (dynamics) of DNA and RNA model systems we can finally turn to application, or the *why*. We will now examine a case in which established techniques help us address a question of application for DNA as an extension of ourselves (our perhaps just our *intentions*) on the molecular level. We will ask how much and which kinds of energy is necessarily involved in maintaining proper B form DNA in the presence of a single walled carbon nanotube (DNA-SWNT). The question will be answered directly and thoroughly, using molecular dynamics simulations and non-equilibrium techniques in illustration of this powerful tool, while highlighting another instance of a smooth structural continuum along which we can construct not just *motional* axes, but also *cognitive* axes of formal inquiry.

Residue	Bond Vector	Virtual S^2 (μs)	Virtual S^2 (ns)	Experimental S^2
20	C2-H2	0.836827	0.903566	0.933
22	C2-H2	0.652108	0.431947	0.917
27	C2-H2	0.779693	0.894532	0.778
19	C5-H5	0.796618	0.887733	0.849
23	C5-H5	0.523697	0.493803	0.699
24	C5-H5	0.563808	0.465806	0.405
25	C5-H5	0.534844	0.181489	0.535
29	C5-H5	0.722233	0.784571	0.701
31	C5-H5	0.418504	0.283325	0.728
37	C5-H5	0.802817	0.747342	0.684
38	C5-H5	0.779247	0.867186	0.767
39	C5-H5	0.754591	0.805209	0.764
40	C5-H5	0.618656	0.472192	0.74
41	C5-H5	0.767271	0.583739	0.876
42	C5-H5	0.822753	0.822422	0.916
44	C5-H5	0.785343	0.867901	0.884
19	C6-H6	0.819641	0.90094	0.824
23	C6-H6	0.525801	0.46684	0.733
24	C6-H6	0.561755	0.456748	0.434
25	C6-H6	0.491904	0.130369	0.485
29	C6-H6	0.722173	0.781123	0.632
31	C6-H6	0.379898	0.198608	0.652
38	C6-H6	0.766674	0.874499	0.657
41	C6-H6	0.790805	0.549784	0.875
17	C8-H8	0.400137	0.882008	0.859
18	C8-H8	0.789814	0.91161	0.93
20	C8-H8	0.846155	0.89856	0.845
28	C8-H8	0.833479	0.91046	0.663
34	C8-H8	0.639578	0.502295	0.524
36	C8-H8	0.698209	0.649526	0.655
43	C8-H8	0.858363	0.920142	0.814
17	N1-H21	0.592508	0.885589	0.974
18	N1-H21	0.798343	0.901512	0.965
21	N1-H21	0.853343	0.851045	0.965
28	N1-H21	0.826535	0.891519	0.742
34	N1-H21	0.621678	0.532792	0.682
36	N1-H21	0.668556	0.639086	50.705

Table 4.2: Experimental and virtual (both μs and ns based) S^2 order parameters for HIV-1 TAR RNA for base moieties, taken with permission from Musselman et al [10] listed by both residue and bond vector type. For residue indices see figure 4.5.

Residue	Bond Vector	Virtual S^2 (μs)	Virtual S^2 (ns)	Experimental S^2
12	C1'-H1'	0.01088	0.425228	0.475
11	C1'-H1'	0.385971	0.63784	0.839
10	C1'-H1'	0.269344	0.823602	0.926
9	C1'-H1'	0.720448	0.829272	0.941
8	C1'-H1'	0.80253	0.837832	0.982
7	C1'-H1'	0.811128	0.831792	0.975
6	C1'-H1'	0.794186	0.833938	1
5	C1'-H1'	0.833665	0.842379	1
4	C1'-H1'	0.715024	0.816323	0.821
3	C1'-H1'	0.306032	0.733786	0.891
2	C1'-H1'	0.138399	0.797191	0.785
1	C1'-H1'	0.08989	0.410954	0.698
13	C1'-H1'	0.003569	0.080983	0.5
14	C1'-H1'	0.33603	0.707657	0.767
15	C1'-H1'	0.73593	0.755918	0.696
16	C1'-H1'	0.630141	0.765404	0.917
17	C1'-H1'	0.648972	0.827771	0.859
18	C1'-H1'	0.659271	0.833594	0.961
19	C1'-H1'	0.614035	0.833189	0.932
20	C1'-H1'	0.434469	0.835281	0.92
21	C1'-H1'	0.278536	0.854433	0.919
22	C1'-H1'	0.247741	0.821861	0.879
23	C1'-H1'	0.162363	0.767542	0.652
24	C1'-H1'	0.068058	0.395044	0.714

Table 4.3: Experimental and virtual (both μs and ns based) S^2 order parameters for A6DNA ribose bond vectors, taken with permission from Nikolova et al [12] listed by both residue and bondvector type. For residue indices see figure 4.3.

Residue	Bond Vector	Virtual S^2 (μs)	Virtual S^2 (ns)	Experimental S^2
10	C2-H2	0.712512	0.894155	0.935
16	C2-H2	0.86841	0.895894	0.97
17	C2-H2	0.865176	0.902717	0.977
20	C2-H2	0.825359	0.905883	1
21	C2-H2	0.779602	0.906125	0.983
12	C6-H6	00.169369	0.569345	0.855
9	C6-H6	00.682788	0.77443	0.927
4	C6-H6	00.640592	0.712244	0.948
1	C6-H6	00.07008	0.733913	0.963
14	C6-H6	00.475992	0.700957	0.992
15	C6-H6	10.654373	0.697078	1
22	C6-H6	00.414055	0.779668	0.956
23	C6-H6	00.465442	0.71187	0.998
11	C8-H8	0.719795	0.929966	0.95
10	C8-H8	0.770541	0.94504	1
3	C8-H8	0.886859	0.930744	0.956
2	C8-H8	0.65568	0.916858	0.95
13	C8-H8	0.364406	0.893297	0.902
16	C8-H8	0.898802	0.948586	1.001
17	C8-H8	0.900458	0.95317	0.995
24	C8-H8	0.479022	0.878048	0.875

Table 4.4: Experimental and virtual (both μs and ns based) S^2 order parameters for A6DNA base moieties, taken with permission from Nikolova et al [12] listed by both residue and bondvector type. For residue indices see figure 4.3.

Bibliography

[1] Florence Tama, Florent Xavier Gadea, Osni Marques, and Yves-Henri Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Structure, Function, and Bioinformatics*, 41(1):1–7, 2000.

[2] Bertil Halle. The physical basis of model-free analysis of nmr relaxation data from proteins and complex fluids. *The Journal of chemical physics*, 131(22):224507, 2009.

[3] Giovanni Lipari and Attila Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *Journal of the American Chemical Society*, 104(17):4546–4559, 1982.

[4] Tim Zeiske, Kate A Stafford, Richard A Friesner, and Arthur G Palmer. Starting-structure dependence of nanosecond timescale intersubstate transitions and reproducibility of md-derived order parameters. *Proteins: Structure, Function, and Bioinformatics*, 81(3):499–509, 2013.

[5] Catherine Musselman, Hashim M Al-Hashimi, and Ioan Andricioaei. ired analysis of tar rna reveals motional coupling, long-range correlations, and a dynamical hinge. *Biophysical journal*, 93(2):411–422, 2007.

[6] Jeffrey W Peng and Gerhard Wagner. Mapping of spectral density functions using heteronuclear nmr relaxation measurements. *Journal of Magnetic Resonance (1969)*, 98(2):308–332, 1992.

[7] Erik RP Zuiderweg. Mapping protein-protein interactions in solution by nmr spectroscopy. *Biochemistry*, 41(1):1–7, 2002.

[8] G Marius Clore, Attila Szabo, Ad Bax, Lewis E Kay, Paul C Driscoll, and Angela M Gronenborn. Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *Journal of the American Chemical Society*, 112(12):4989–4991, 1990.

[9] Alexander D MacKerell, Nilesh Banavali, and Nicolas Foloppe. Development and current status of the charmm force field for nucleic acids. *Biopolymers*, 56(4):257–265, 2000.

[10] Catherine Musselman, Qi Zhang, Hashim Al-Hashimi, and Ioan Andricioaei. Referencing strategy for the direct comparison of nuclear magnetic resonance and molecular dynamics motional parameters in rna. *The Journal of Physical Chemistry B*, 114(2):929–939, 2009.

[11] Eric R Henry and Attila Szabo. Influence of vibrational motion on solid state line shapes and nmr relaxation. *The Journal of chemical physics*, 82(11):4753–4761, 1985.

[12] Evgenia N Nikolova, Gavin D Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. Probing sequence-specific dna flexibility in a-tracts and pyrimidine-purine steps by nuclear magnetic resonance 13c relaxation and molecular dynamics simulations. *Biochemistry*, 51(43):8654–8664, 2012.

[13] Alberto Pérez, F Javier Luque, and Modesto Orozco. Dynamics of b-dna on the microsecond time scale. *Journal of the American Chemical Society*, 129(47):14739–14745, 2007.

[14] Elizabeth J Denning, U Priyakumar, Lennart Nilsson, and Alexander D Mackerell. Impact of 2-hydroxyl sampling on the conformational properties of rna: Update of the charmm all-atom additive force field for rna. *Journal of computational chemistry*, 32(9):1929–1943, 2011.

[15] Nicolas Foloppe and Alexander D MacKerell Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, 21(2):86–104, 2000.

[16] Alexander D Mackerell and Nilesh K Banavali. All-atom empirical force field for nucleic acids: Ii. application to molecular dynamics simulations of dna and rna in solution. *Journal of Computational Chemistry*, 21(2):105–120, 2000.

[17] Elzbieta Kierzek, Anna Pasternak, Karol Pasternak, Zofia Gdaniec, Ilyas Yildirim, Douglas H Turner, and Ryszard Kierzek. Contributions of stacking, preorganization, and hydrogen bonding to the thermodynamic stability of duplexes between rna and 2-o-methyl rna with locked nucleic acids. *Biochemistry*, 48(20):4377–4387, 2009.

[18] Alan Cooper. Protein fluctuations and the thermodynamic uncertainty principle. *Progress in biophysics and molecular biology*, 44(3):181–214, 1984.

[19] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Current opinion in structural biology*, 9(2):164–169, 1999.

[20] Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Current opinion in structural biology*, 10(2):165–169, 2000.

[21] Aaron T Frank, Andrew C Stelzer, Hashim M Al-Hashimi, and Ioan Andricioaei. Constructing rna dynamical ensembles by combining md and motionally decoupled nmr rdcs: new insights into rna dynamics and adaptive ligand recognition. *Nucleic acids research*, 37(11):3670–3679, 2009.

[22] D Thirumalai, Raymond D Mountain, and TR Kirkpatrick. Ergodic behavior in supercooled liquids and in glasses. *Physical Review A*, 39(7):3563, 1989.

[23] Paul Maragakis, Kresten Lindorff-Larsen, Michael P Eastwood, Ron O Dror, John L Klepeis, Isaiah T Arkin, Morten Ø Jensen, Huafeng Xu, Nikola Trbovic, Richard A

Friesner, et al. Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *The Journal of Physical Chemistry B*, 112(19):6155–6158, 2008.

[24] Scott A Showalter, Nathan A Baker, Changguo Tang, and Kathleen B Hall. Iron responsive element rna flexibility described by nmr and isotropic reorientational eigenmode dynamics. *Journal of biomolecular NMR*, 32(3):179–193, 2005.

[25] Loïc Salmon, Gavin Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. A general method for constructing atomic-resolution rna ensembles using nmr residual dipolar couplings: the basis for interhelical motions revealed. *Journal of the American Chemical Society*, 135(14):5457–5466, 2013.

[26] William L Jorgensen. Rusting of the lock and key model for protein-ligand binding. *Science*, 254(5034):954–955, 1991.

[27] Hashim M Al-Hashimi and Nils G Walter. RNA dynamics: it is about time. *Current opinion in structural biology*, 18(3):321–329, 2008.

[28] Bruno Spire, Ségolène Duran, Marc Souville, Catherine Leport, François Raffi, and Jean-Paul Moatti. Adherence to highly active antiretroviral therapies (haart) in hiv-infected patients: from a predictive to a dynamic approach. *Social science & medicine*, 54(10):1481–1496, 2002.

Chapter 5

Applications in Nanotech

5.1 Introduction

While up until this point we have primarily been concerned with the *where* and *when* of molecular motions at particular timescales, there have been several allusions to the biological embodiment that such a knowledge allows for. Although a primary concern for biologists should lie at elucidating *structure* and *function* in analytical forms, the final purpose for such elucidation is in then using that knowledge to design or identify points at which we may effectively intervene. Intervention requires some form of embodied action, however, and as such we must endeavor to build tools of functional access and not just tools of monitoring. Such points can then be canonized such that our future generations also benefit not only from the tools developed but from the knowledge that led to those tools as well.

In the previous chapters we showed that despite the lucrative desire to assign clean categorical breaks in the motions and structures of nucleic acids at various timescales, some of them cannot be de-convoluted in any straightforward way (our primary example throughout this document is the uncoupling of internal motions to overall tumbling and diffusion in RNA).

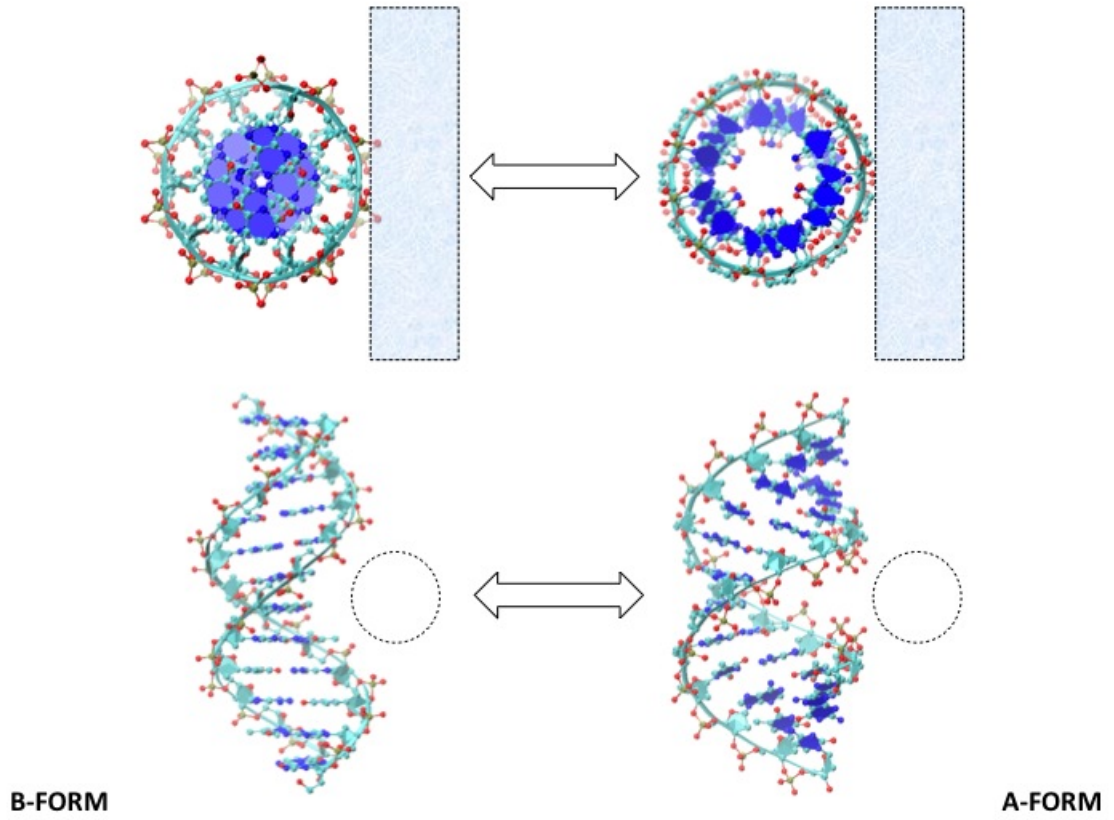

Figure 5.1: Ideal B and A form DNA shown from the top and side, with the outline of the SWNT position during simulations. Notice the widening of the interhelical distance and major groove from B to A form. Double arrows indicate interconversion in a smooth continuous fashion as a response to environmental changes.

Here we show how the inability to cleanly separate groups of DNA motions that are important to overall function and life in the cell can actually be utilized as a tool instead of a hindrance; an axis along which we can inquire and build knowledge for later use. By studying the A $\Leftrightarrow$ B transition of DNA in solution (which represents a *continual, smooth transition*) we can see how the tools and insight we gain from accurate atomistic molecular simulation allow us to design a theoretical construct that numerically assesses free energy. A well defined potential of mean force (PMF) allows us to assess how much energy would be required to maintain nucleic acids in either A or B form, while in the presence of a single walled carbon nanotube (SWNT). Specifically, we here look at a conformation in which a dodecamer ds-DNA of either poly(dGC) or poly(dTA) sequence and a SWNT is bound tightly on the major groove side. As mentioned above, we calculate the potential of mean force for A to B transitions in such a configuration and demonstrate dependence on AT or GC rich sequences.

5.1.1 Single Walled Carbon Nanotubes and DNA

Single Walled Carbon nanotubes (SWNT) are cylindrical carbon fullerenes that vary widely in both structure and application [1]. By wrapping a carbon sheet of specified thickness (single walled nanotubes refer to nanotubes which are one atom thick) around itself one can construct these unusual molecules with relative ease, and these molecular structures exhibit quite amazing properties. Chemical attributes can be controlled by simple alterations to the preparation protocols, and the resulting fibers are extremely strong, accurate in position down to the atomistic detail, and can be altered in terms of electrical conductivity, optical properties, thermal properties and structural attributes. For these reasons they have found widespread use over the last few decades in a myriad of applications, particularly in nanotech [2]. Specifically in the arena of nucleic acids, it has been proposed and even demonstrated that the interactions of DNA with SWNTs could facilitate things like DNA-SWNT nano-engines, [3] ultra-fast gene sequencing [4, 5], gene therapies, protecting DNA from oxidation,

and cancer treatments [6, 7]. The system that inspired the current study however, comes from Gang et al in 2005, [8] in which SWNTs were placed in the major groove of a DNA molecule and energetically assessed by the self consistent-charge density-functional based tight-binding method (SCC-DFTB) [8, 9] to demonstrate the stability and feasibility of such a configuration. Additionally, much has been done both experimentally and in simulation to show that SWNTs interact in various ways with DNA, mostly by adhereing to surfaces or even spontaneous DNA insertion into the interior of nanotubes [10, 11, 12]. Furthermore it should be noted that SWNTs can be expressed in lipid bilayers to constitute a perfect open pore through which water can pass freely [13]. Some consideration should be taken, however, to note that the DNA itself might not maintain a constant structure when in the presence of SWNTs. If the previous chapters of this dissertation have made any solid claims, it is that nucleic acids are highly dynamic entities, and it behooves us to assume that they will continue to be dynamic entities near SWNTs. In the aforementioned studies we are not privy to dynamics in atomistic detail, but can only see collective variables based on measured ensembles that average across time. The only alternative picture of dsDNA bound to a SWNT is the static snapshot published by Gu et al, which gives only energetics of a single structure [8]. Johnson et al demonstrated several key aspects of DNA dynamics in the presence of SWNTs, but only for single-stranded DNA-SWNT hybrids [14], and the ds-DNA-SWNT complex has not been investigated for dynamics to the best of our current knowledge. If DNA is highly dynamic, however, which dynamics might we wish to probe? One of the most common and biologically relevant motions available to DNA is the A to B transition. B form DNA is characterized by highly stable DNA with decreased inter-helical distance and C3 endo sugar pucker, while A form is characterized by increased inter-helical distance but decreased diameter and C2 endo sugar puckering [15]. Despite only about 6 Å difference in root mean squared deviation (RMSD), the A to B transition has been the key aspect by which simulations have been confirmed to be correctly parameterized, and many fruitful efforts have come from this fundamental transition [16, 17]. Furthermore, the A to

B forms of DNA do not mark two simple distinct conformational structures, but instead the extremes of a continuum of states that can evolve diffusively from one to another across an A/B spectrum. Appropriately, this study explores a vivid example of how recognizing and labeling different aspects of a continually diffusive spectrum of states is a powerful tool that allows for clear and concise hypotheses to be investigated systematically. As was discussed at length in the previous chapter, there are many instances regarding nucleic acid conformational structures in which states are only measurable separately, but motions allowed to them suggest continuous diffusive interconversion. Such inter-conversion is often strongly correlated across various timescales and hierarchically juxtaposed functions. This is in opposition to globular proteins that generally demonstrate stable folds and discrete states or motions, which can be separately treated. Hopefully, the following application demonstrates the power of the fully atomistic (and fully sampled) simulation diffusing across a continual separation of states to derive the free energy of transition. By treating the transition as such we can avoid the over eager assignment of causal relationships to various conformational species which actually coexist in complicated and hierarchical relationships, and the corollary systemic interactions facilitated by such subtle motions.

5.2 Methods

5.2.1 Umbrella Sampling and WHAM

Due to difficulties simulating at long time scales and entropic pooling there is a hard limit to brute force sampling (another central focus of this study). At times when the pathway or transition of interest is known, it is wise to use non-equilibrium techniques to describe the transitions, and statistical physics provides us with a myriad of tools to perform analysis on the resulting biased trajectories. One such method is umbrella sampling in conjunction

with the weighted histogram analysis method, in which artificial potentials are posted along a predetermined reaction coordinate to ensure sampling near wells covering the said reaction coordinate. These umbrella potentials force sampling along the area of interest, and the resulting probability distributions can be solved for the energy it took to hold the reaction coordinate there, telling us valuable information about the energy required to traverse that reaction coordinate. Specifically, if we modify our initial Hamiltonian to include the newly placed constraints, we arrive at (we are representing the 3N Cartesian coordinates simply as x for clarity):

$$\hat{H}_\lambda(x) = \hat{H}_0(x) + \sum_{i=1}^{N} \lambda_i \hat{V}_i(x) \tag{5.1}$$

Where N stands for the total number of umbrella potentials (and therefore simulations) indexed by i and given as $\hat{V_i}(x)$, while $\hat{H}_0(x)$ describes the unperturbed potential. By allowing the simulation to then proceed along the "umbrellas" one can measure the probability of being in each window after sufficient sampling $n_i(x)$. By Bayesian statistics and the method of minimizing errors in overlapping $n_i(x)$s (see the appendix in [18] we can arrive at the WHAM equations given as :

$$P(x) = \frac{\sum_{i=1}^{N} n_i(x)}{\sum_{i=1}^{N} N_i exp\left(\frac{[F_i - U_{bias,i}(x)]}{k_B T}\right)} \tag{5.2}$$

$$F_i = -k_B T ln\left[\sum_{x_{bins}} P(x) - exp\left(\frac{-U_{bias,i}(x)}{k_B T}\right)\right] \tag{5.3}$$

where U_{bias} is analogous to $\hat{V}(x)_i$ as given in equation 5.1, F_i is the Helmholtz free energy shift during the simulation and $P(x)$ is the *best estimate* of the unbiased probability distribution

as given by the minimizing errors in window overlap. Both $P(x)$ and F_i are unknowns, and so the WHAM equations must be solved self-consistently by numerical integration. These results provide a powerful formalism with which we can ascertain the potential of mean force (PMF) of keeping the system near the reaction coordinate, provided ample sampling is reached. It has been proven useful for developing an intuitive understanding of relative energy barriers and can even be related to kinetics and ensemble populations given enough experimental information about the system. In short it is the closest microscopic metric we have to correlate to the macroscopic Gibbs free energy, which is commonly used to predict spontaneity and can be given in an alternative formulation:

$$V(x) = \int_o^x \langle F \rangle (x')dx' \tag{5.4}$$

or the sum of the average force necessary to keep the system at position x' relative to x [19] As mentioned above, one major advantage of the technique is that the reaction coordinate can be chosen such that it represents the most interesting transition available to us (whether biological in nature or otherwise). In order to study the A to B transition of DNA at varying sequences in the presence of SWNTs, we have chosen the ΔRMSD measurement as a metric around which to assign umbrella potentials. By defining our reaction coordinate as ΔRMSD we can now write the following potential which we will add to the simulation for umbrella sampling:

$$E = \sum_i k_f^i [(RMSD_B - RMSD_A) - D_{min}]^2 \tag{5.5}$$

Where k_f is a force constant, and D_{min} is the minimum around which the constraint is held. This constraint was used with great success in order to find the potential of mean force for the A to B transition by Banavali et al in 2005 [20], and we have similarly adopted it in order to calculate the potential of mean force (PMF) for a DNA molecule undergoing the A to B transition in the presence of a SWNT.

5.2.2 Simulation Parameters

Poly-(dAT) and poly-(dGC) B form and A form dsDNA dodecamers were built using Nucleic Acid Builder (NAB) as implemented in AMBERTOOLS [21]. The carbon nanotube was built using TubeGen as an achiral (10,0) nanotube [22]. Systems were then solvated and ionized using Visual Molecular Dynamics [23] in a 64x64x64 angstrom cube such that charge is neutralized. Periodic boundary conditions and Particle Mesh Ewald were used for long distance electrostatics [24], and a 2 fs integrator time step with SHAKE was implemented with langevin dynamics [25, 26]. CHARMM36 force fields with standard sp2 hybridized carbon parameters were used for all simulations [27, 28]. 30 simulation windows were run with a distance of .4 Å separation by ΔRMSD reaction coordinate, and histograms of the reaction coordinate at each window were generated to ensure sufficient overlap in WHAM calculations. Furthermore, an additional 30 simulation windows were simulated and then interpolated at .2 Å separation to ensure convergence of PMF. PMF was calculated by the Weighted Histogram Analysis Method (WHAM) as described above, and implemented in a C like script by Alan Grossfield [29].

5.3 Results: The Potential Of Mean Force

Simulations all showed good agreement with expected helical and base parameters, and all trace file and histograms showed good overlap so as to assure confidence in simulations. Furthermore, the PMF generated without the SWNT matches perfectly with that computed by Noy et al [15] as given in figures 5.2 and 5.3. In figure 5.2 we show the poly(dGC) sequence in the presence of the SWNT plotted against the ΔRMSD reaction coordinate. In this formulation, a positive reaction coordinate represents large B like character, while a negative ΔRMSD represents little B character and large A character. Simulations were all inspected visually to ensure B and A forms were indeed well represented by the reaction coordinates. The PMF with the SWNT present shows some markedly different features, although the overall curve has similar overall character (concavity, limits etc). Particularly the SWNT present PMF curve is higher near the A form by as much as 5-10 kcal/mol, a substantial increase. Furthermore, the minimum near B like region has shifted left by as much as 2 Å , indicating the equilibrium poly-(dGC) nucleotide structure may be significantly different than the corresponding nucleic structure without the SWNT present. Similarly the PMF of poly-(dTA) with and without the SWNT present is given in figure 5.3. Many of the poly-(dTA) PMF qualities are similar to those of the poly-(dGC) dodecamer with SWNT present. Finally, a visual schematic of the simulations starting and ending windows are given in figure 5.5. Finally, the average RMSD with respect to either A or B form during simulation windows is shown in figure 5.4 demonstrating the decreased tendency of DNA to fluctuate around A form like parameters when the nanotube is present.

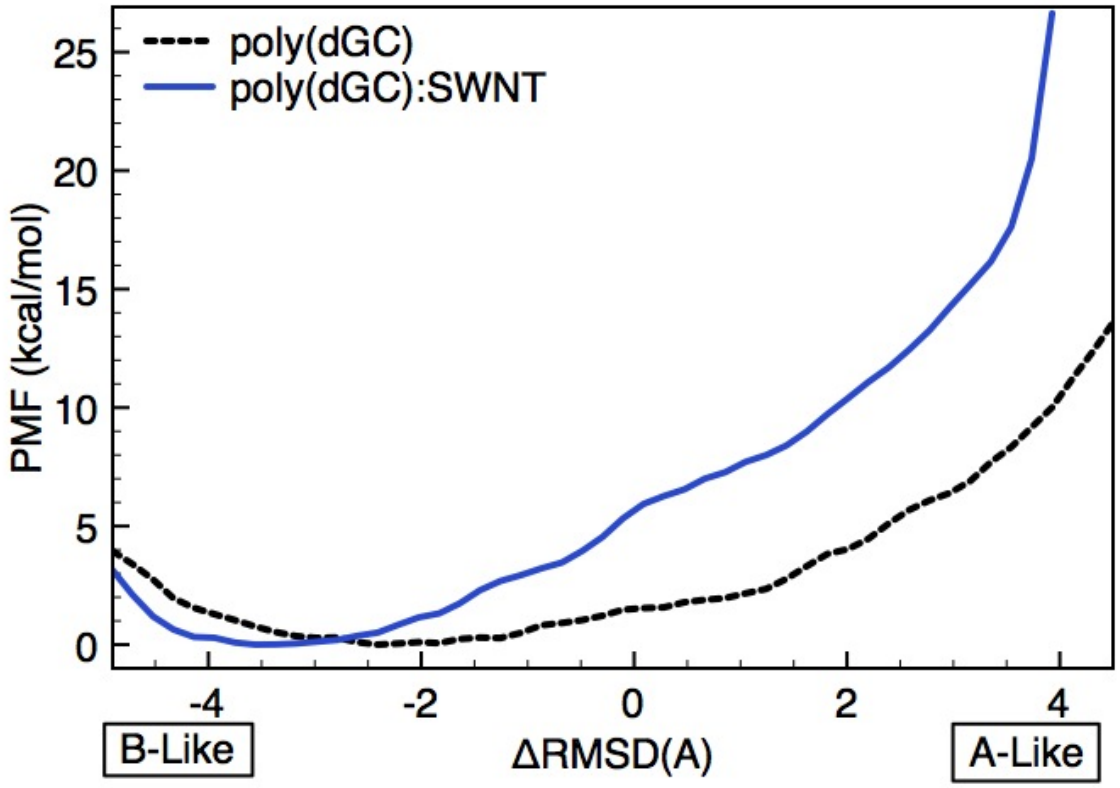

Figure 5.2: Potential of Mean Force for a poly-(dGC) dodecamer with (solid blue line) and without (dashed black line) a SWNT fit into the major groove plotted against ΔRMSD. ΔRMSD is a quantitative measure of B vs A form character, where large negative ΔRMSD indicates B like structure and a large positive ΔRMSD indicates A like structure. Presence of the SWNT shifts the equilibrium position around which the DNA molecule fluctuates about 2 Å closer to B form than without the SWNT.

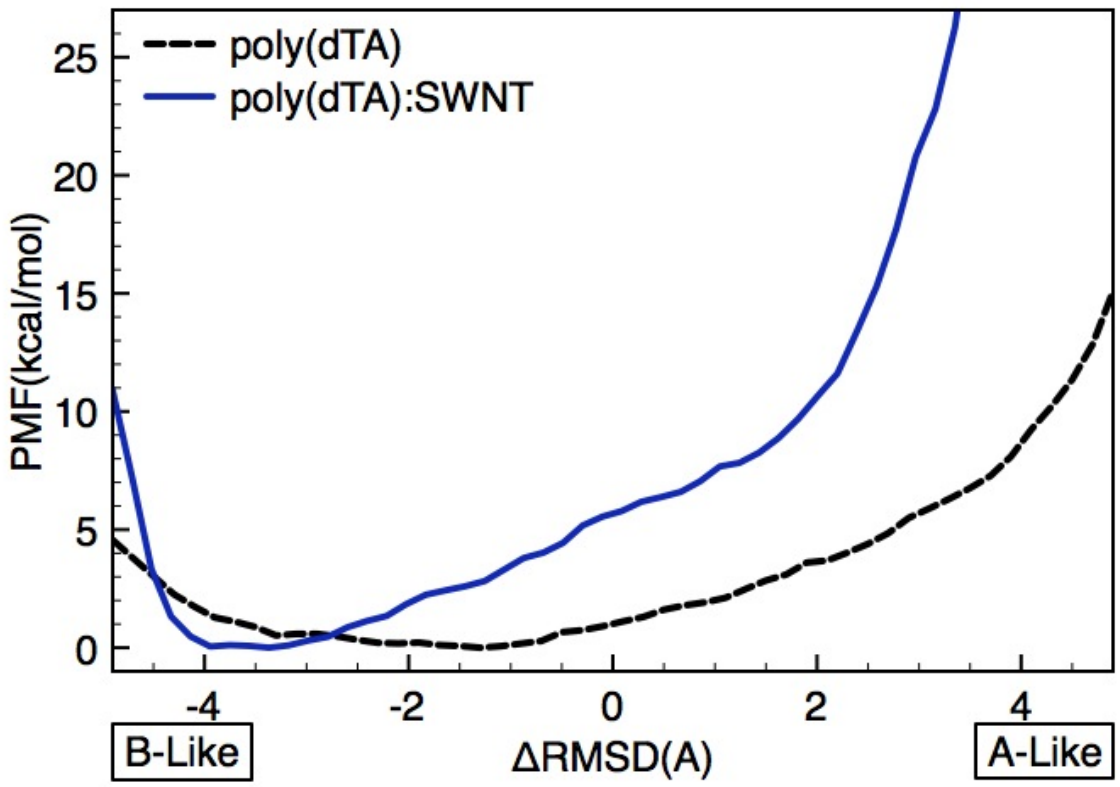

Figure 5.3: Potential of Mean Force for a poly-(dTA) dodecamer with (solid blue line) and without (dashed black line) a SWNT fit into the major groove plotted against ΔRMSD. ΔRMSD is a quantitative measure of B vs A form character, where large negative ΔRMSD indicates B like structure and a large positive ΔRMSD indicates A like structure. Presence of the SWNT shifts the equilibrium position around which the DNA molecule fluctuates about 3 Å closer to B form than without the SWNT, and the penalty for fluctuating away from B form is much sharper than in the poly-(dGC) case given in figure 5.2.

5.4 Discussion

First we should note that the PMFs without a SWNT present are in good agreement with published results involving the A to B transition as calculated by a similar methodology (umbrella sampling with WHAM and ΔRMSD as umbrella constraint) [20]. This establishes with confidence the ability of our methods to reproduce accurate energetics for this transition while using this metric. Upon looking closely at these curves one might correctly deduce that the dodecamer is fluctuating just a bit away from idealized B form in normal solution without the SWNT present, and we assume this is the effect of the shortened length of the DNA molecule. The major finding of the resulting curves involves the shift in global minimum from SWNT present curves versus those without, suggesting that the equilibrium structure of the nucleic sequences has been altered by around 2 Å . This is indeed an important result; seeing as the DNA will not adopt the same fluctuations between A and B form in the presence of SWNT than otherwise. It is important to note however, that we are *not* suggesting that DNA will adopt a relatively static conformation that is B-like, but instead that it will fluctuate *around an equilibrium* which is more B like than A like, and that near AT rich sequences, the fluctuations will be measurably tighter by as much as a few kcal/mol. This is a significant amount of energy for a single molecule calculation; in ensemble equilibrium the effects could definitely be measurable. Furthermore, it is likely that significant applications could be derived from such a measurable difference in energy around different sequences.

If one were to attempt to build a causal accounting of this result, it would have to include both the fact that the largely hydrophobic core of the SWNT strongly attracts the negatively charged DNA backbone and the effects of steric crowding considerations, but we must not be hasty in assigning definite assertions regarding such character. The truth is that the data simply states that DNA is in lower energy when it fluctuates closer to B form when a SWNT is present near the major groove, and that this effect is even more present near AT rich sequences. The sequence dependence of this transition is particularly exciting, in

that SWNTs are already thought to be good candidates for ultra fast genome sequencing techniques [4].

Furthermore, we should consider at length the implication of our result, that the smooth continual nature of the B to A DNA transition can be gently perturbed without introducing fundamental changes to the molecular structure, and without designing a specific "key" like ligand. This hints at a broad direction that biomedical nucleic acid research might take, that of gentle "coercion" of a species into a target fluctuation, as opposed to strict reformation of secondary or tertiary structure. As discussed in the previous chapter, this is a common motif for nucleic acids as opposed to proteins, who also exhibit many of these properties but only when the structure-function relationship is more complicated than simple lock and key type mechanisms [30, 31, 32, 33]. It seems likely that all of the nucleic acid major structural conformational transitions will exhibit these types of motions. This would suggest that some considerable difficulty will arise in interpreting single molecule experiments regarding non-canonical nucleic acid structures, which is already common in the literature [34].

5.5 Conclusion

The question of fundamental vibrational modes in nucleic acids has been a central thesis to this study, and we hope that it is now apparent why such care was taken in exploring the way in which such motions make up a continual smooth transition landscape between seemingly disparate structures, in which all of the intermediates are available for possible adoption under normal circumstances (or for contributing to other effects). By understanding the smooth movement from A to B transitions, and how modulating the surrounding environment of the molecule can significantly alter this landscape, we demonstrate the fundamental difficulty in ascertaining significance from our simulations. Simulated trajectories are fickle and indeterminate, but not because they are fundamentally inaccurate (in many cases they

are most definitely *not*), but because they represent only *one possible* trajectory; there is no guarantee that any given trajectory encapsulates enough information to answer whichever question is driving the inquiry. In almost every case simulation describes trajectories that are likely much more complicated than we initially ascertain (or currently have the faculties to be able to ascertain), and we are just barely scratching the surface of this complexity when we begin to consider the more fluid types of structural interconversion such as the A to B transition. Our previous results show a fundamental shift in the focus of how applications should be derived for such structures, they must be *guided* along with a gentle touch into more defined conformations as opposed to assuming they will visit them naturally or be forced into them by some rigid molecule which is designed to bind somewhere along it.

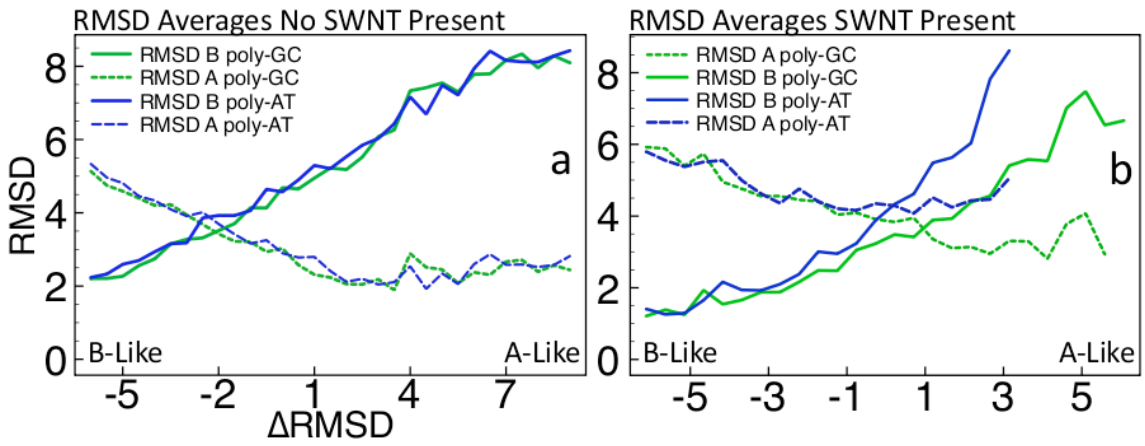

Figure 5.4: Averages of RMSD with respect to A and B structures for each window plotted against the position of the ΔRMSD constraint (B) with SWNT present, and (A) with no SWNT present. A negative ΔRMSD represents large B character and little A character, whereas a positive ΔRMSD represents large A character and little B character. (A) shows a smooth transition to A form without SWNT present, whereas (A) shows the difficulty that DNA has in adopting the A form despite the constraint being applied.

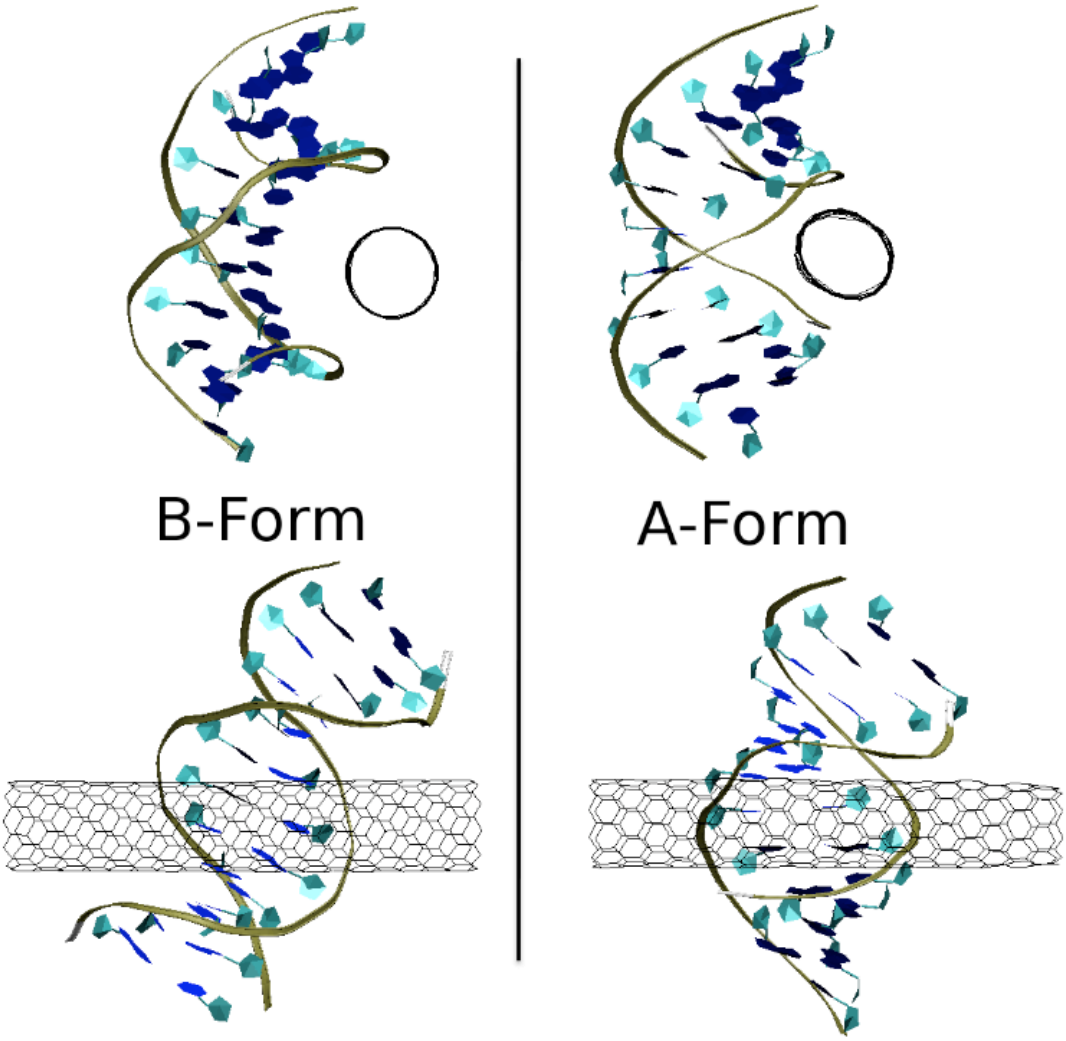

Figure 5.5: DNA molecules during simulation showing A form (right) or B form (left) with SWNT fitted into the major groove. The simulations were constrained along the smooth continuum of structures representing this transition for calculation of Potential of Mean Force (PMF) by umbrella sampling and the Weighted Histogram Analysis Method (WHAM) [18].

Bibliography

[1] Rupesh Khare and Suryasarathi Bose. Carbon nanotube based composites-a review. *Journal of minerals and Materials Characterization and Engineering*, 4:31, 2005.

[2] J Justin Gooding. Nanostructuring electrodes with carbon nanotubes: A review on electrochemistry and applications for sensing. *Electrochimica Acta*, 50(15):3049–3060, 2005.

[3] Suwussa Bamrungsap, Joseph A Phillips, Xiangling Xiong, Youngmi Kim, Hui Wang, Haipeng Liu, Arthur Hebard, and Weihong Tan. Magnetically driven single dna nanomotor. *small*, 7(5):601–605, 2011.

[4] Ming Zheng, Anand Jagota, Michael S Strano, Adelina P Santos, Paul Barone, S Grace Chou, Bruce A Diner, Mildred S Dresselhaus, Robert S Mclean, G Bibiana Onoa, et al. Structure-based carbon nanotube sorting by sequence-dependent dna assembly. *Science*, 302(5650):1545–1548, 2003.

[5] Xiaomin Tu, Suresh Manohar, Anand Jagota, and Ming Zheng. Dna sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature*, 460(7252):250–253, 2009.

[6] Christopher J Gannon, Paul Cherukuri, Boris I Yakobson, Laurent Cognet, John S Kanzius, Carter Kittrell, R Bruce Weisman, Matteo Pasquali, Howard K Schmidt,

Richard E Smalley, et al. Carbon nanotube-enhanced thermal destruction of cancer cells in a noninvasive radiofrequency field. *Cancer*, 110(12):2654–2665, 2007.

[7] Elijah J Petersen, Xiaomin Tu, Miral Dizdaroglu, Ming Zheng, and Bryant C Nelson. Protective roles of single-wall carbon nanotubes in ultrasonication-induced dna base damage. *Small*, 9(2):205–208, 2013.

[8] Gang Lu, Paul Maragakis, and Efthimios Kaxiras. Carbon nanotube interaction with dna. *Nano letters*, 5(5):897–900, 2005.

[9] Marcus Elstner, Dirk Porezag, G Jungnickel, J Elsner, M Haugk, Th Frauenheim, Sandor Suhai, and Gotthard Seifert. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Physical Review B*, 58(11):7260, 1998.

[10] Yuichi Noguchi, Tsuyohiko Fujigaya, Yasuro Niidome, and Naotoshi Nakashima. Single-walled carbon nanotubes/dna hybrids in water are highly stable. *Chemical Physics Letters*, 455(4):249–251, 2008.

[11] Huajian Gao and Yong Kong. Simulation of dna-nanotube interactions. *Annu. Rev. Mater. Res.*, 34:123–150, 2004.

[12] Huajian Gao, Yong Kong, Daxiang Cui, and Cengiz S Ozkan. Spontaneous insertion of dna oligonucleotides into carbon nanotubes. *Nano Letters*, 3(4):471–473, 2003.

[13] Mainak Majumder, Nitin Chopra, Rodney Andrews, and Bruce J Hinds. Nanoscale hydrodynamics: enhanced flow in carbon nanotubes. *Nature*, 438(7064):44–44, 2005.

[14] Robert R Johnson, AT Charlie Johnson, and Michael L Klein. Probing the structure of dna-carbon nanotube hybrids with molecular dynamics. *Nano letters*, 8(1):69–75, 2008.

[15] Richard E Dickerson, Horace R Drew, Benjamin N Conner, Richard M Wing, Albert V Fratini, Mary L Kopka, et al. The anatomy of a-, b-, and z-dna. *Science*, 216(4545):475–485, 1982.

[16] TE Cheatham III and PA Kollma. Observation of the¡ i¿ a¡/i¿-dna to¡ i¿ b¡/i¿-dna transition during unrestrained molecular dynamics in aqueous solution. *Journal of molecular biology*, 259(3):434–444, 1996.

[17] Agnes Noy, Alberto Pérez, Charles A Laughton, and Modesto Orozco. Theoretical study of large conformational transitions in dna: the b a conformational change in water and ethanol/water. *Nucleic acids research*, 35(10):3330–3338, 2007.

[18] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of computational chemistry*, 13(8):1011–1021, 1992.

[19] David Keller, David Swigon, and Carlos Bustamante. Relating single-molecule measurements to thermodynamics. *Biophysical journal*, 84(2):733–738, 2003.

[20] Nilesh K Banavali and Benoît Roux. Free energy landscape of a-dna to b-dna conversion in aqueous solution. *Journal of the American Chemical Society*, 127(18):6866–6876, 2005.

[21] DA Case, TA Darden, TE Cheatham Iii, CL Simmerling, J Wang, RE Duke, R Luo, RC Walker, W Zhang, KM Merz, et al. Amber tools 1.5, 2010.

[22] JT Frey and DJ Doren. Tubegen 3.3 university of delaware newark de. 2005. *Available via web interface http://turin. nss. udel. edu/research/tubegenonline. html. Accessed*, 16, 2010.

[23] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

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

[25] Vincent Kräutler, Wilfred F van Gunsteren, and Philippe H Hünenberger. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry*, 22(5):501–508, 2001.

[26] Mike P Allen and Dominic J Tildesley. Computer simulation of liquids. 1989.

[27] Nicolas Foloppe and Alexander D MacKerell Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, 21(2):86–104, 2000.

[28] Alexander D Mackerell and Nilesh K Banavali. All-atom empirical force field for nucleic acids: Ii. application to molecular dynamics simulations of dna and rna in solution. *Journal of Computational Chemistry*, 21(2):105–120, 2000.

[29] Alan Grossfield. Wham: the weighted histogram analysis method. *Version 2.02 Downloaded at*, 2012.

[30] Shinji Sunada, Nobuhiro Go, and Patrice Koehl. Calculation of nuclear magnetic resonance order parameters in proteins by normal mode analysis. *The Journal of chemical physics*, 104(12):4768–4775, 1996.

[31] Alan Cooper. Protein fluctuations and the thermodynamic uncertainty principle. *Progress in biophysics and molecular biology*, 44(3):181–214, 1984.

[32] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Current opinion in structural biology*, 9(2):164–169, 1999.

[33] Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Current opinion in structural biology*, 10(2):165–169, 2000.

[34] JY Lee, Burak Okumus, DS Kim, and Taekjip Ha. Extreme conformational diversity in human telomeric dna. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):18938–18943, 2005.

Chapter 6

Concluding Remarks

6.1 Formalizing Uncertainty

It is not uncommon to hear scientists and non-scientists alike discussing the merits of various scientific fields in comparison to each other, oftentimes condemning or placing one discipline in a tiered like scaffolding of importance. The story usually goes something like this; physics, being at the bottom, axiomatizes into chemistry, which then axiomatizes into biology, which then axiomatizes into social sciences and the like, almost like a food pyramid with the most "fundamental" at the bottom and the most "derived" at the top (of course it ironically leaves out art, music, poetry, literature and the like, which are just as much scientific disciplines in any functional sense as the natural sciences). Despite the obvious problems and complete lack of physical or epistemological evidence for such a simplified worldview, it is nevertheless repeatedly entertained as a kind of thought experiment; "yes, I know we could never know the positions and momenta of all the atoms in the universe, but if we *did* know them, we could then extrapolate an exact solution to any problem we could ever dream up!" or something along these lines.

The main problem with this type of thinking is not just that it demeans the mountains of effort that has gone into elucidating the various ingenious insights inherent in each field, but that it is also absolutely and undoubtedly wrong in any real sense. It has been shown and discussed at length that each mode of inquiry, be it physics, chemistry, or biology alike has fundamental, irrevocable, and indefinite uncertainties which it must embrace in order to fully capitalize on the information that it can utilize. It is always fun to point out to those which maintain physic's place at the bottom of the canonization pyramid that even if we *were* able to efficiently store all of the positions and momenta of the of atoms in someone's body for any considerable amount of time (which we definitely cannot, and never will be able to in the foreseeable future), it would be of very little use to biologists and medical practitioners alike, *and it would only be one sample!* Quantum mechanics is *not* the only field to stomach uncertainty as a formal part of its discipline, one only needs to read about the historical role of Godel in mathematics to see the amount of time and misdirection that can result if a discipline is unable to recognize, formally accept and canonize what it is systematically unable to determine. Many argue that Godel's simple theorem was directly and causatively integral in allowing for the work of Turing and Von Neumann, without which we would be bereft of digitally automated processes entirely. What we cannot know is at times just as important as what we can know, and specifics about the dynamics of molecules at the quantum-newtonian interface is no exception.

But what is it, exactly, that we cannot know about the dynamics of atoms? Alan Cooper answers this question masterfully in a concise and yet elegant exposition of modern thermodynamics and statistical mechanics written in 1984 in response to the newly emerging field of molecular dynamics simulation [18]. He states, "A distinct change in the way we view biological macromolecules, and proteins in particular, has taken place over the last few years. Despite much elegant and painstaking work in determining the structures of these molecules it has become increasingly clear that a description of the structure of a macromolecule, in terms of a set of atomic coordinates or, more manageably, as a static molecular model, is

insufficient- no matter how accurate- to describe or understand the mechanics of the functions for which these molecules are designed. Moreover, such pictures are inherently misleading in that they convey the impression that these complex structures are static, or that any motion that does occur, in response to ligand binding or a catalytic step, for example, can be described in purely mechanical terms much as one might describe the workings of a mechanical clock. This is at odds with much theory and numerous experimental observations, many of them quite old. We now, most of us, agree that protein molecules are "dynamic", though what we mean by that can depend on personal background and prejudices. But, broadly, we accept that the atoms and groups which make up a macromolecule are in a state of perpetual motion fueled by a thermal excitation, and that the static crystallographic image of conformation is only a first approximation, albeit an indispensable one, to describing the actual physical state of the molecule."

What I would hasten to add to this particularly poignant observation from over 30 years ago, is that the *actual, physical* state of the molecule as Cooper so elegantly puts it is also a thing rapidly losing its place at the center of the way we conceptualize about molecular function. While we can definitely estimate the *actual, physical,* state of some specifically given molecule (given we have enough initial conditions and environmental variables), we cannot solve any *actual, physical,* positions of any generalized atoms in any particularly useful, or directly canonizable way, and we definitely cannot solve the positions continuously or for more than a few atoms at a time. What we *can* do, however, is calculate the kinds of motions *available* to a generalized molecule, using statistical mechanics and thermodynamics to appropriately assign *probabilistic* populations of motion, and endeavor to design meaningful manipulations of said molecular motions.

6.2 Embodied Action in the Cell

What then, is the best recourse available to us? How can we design cellular agents of action if we cannot even be sure about how they move in any specific way? While it is a bit beyond the scope of this document to give a full treatment of the various inroads to altering entities within the cell, let us return to the "mechanical clock" Cooper discussed in the above quote. Ironically, he states that this view is being abandoned by his colleagues, and yet 30 years later it is still the predominant view for mechanistic action that is taught in undergraduate courses and described as tentative hypotheses posited by seasoned scientists as well. Perhaps even more ironically he states that the view of proteins as dynamic entities is rapidly replacing the static picture suggested by crystallographic data, and yet today we commonly talk of proteins as "rigid" in conformation as compared to nucleic acids. The ultimate message that one should take away from this emerging view is that perhaps the reason we fall for these seductive models is not simply because they are valuable first approximations (again a favorite line from Cooper), but that such a simple view is easier to canonize, to axiomatize, to build into a conceptually synergistic picture in which interacting with the cellular species is much more simple than in reality. It is much too seductive an idea to think that macromolecular conformational changes happen in a way that is directly observable; if we run our simulations long enough surely we can simply watch the solved trajectories and ascertain similarly "simple" ways to intervene .

Of course anyone who has actually set out to design, from first principles or otherwise, some sort of entity which can interact in a task oriented way in the cell knows that this is an unfortunately naive view; not only is knowledge of structure, function, and major motions preferred but also a mountain of expertise and a fair amount of nuanced ingenuity is indispensable. Pharmaceutical R&D departments and methodologists alike will tell you that a successful technique requires several inroads to the desired action, an incredible attention to environmental variables, and a solid grasp on a mountain of detail which is (often) well

beyond everyday comprehension. Is it any wonder, then, that we tend towards simple pictures of molecular motions, linking knowledge of structure and simple, vastly under-sampled dynamics alone to molecular embodiment of our intentions?

There is, of course, an alternative here, and it is the simple but difficult realization that we must *use* the increasing complexity which we find as a tool instead of a hindrance, even if it means letting go of our favorite old lock and key type models. In the very big and grand scheme of things, it is not the *actual, physical* positions or momenta of atoms that we are truly after, it is the *concerted* motions of these atoms which make up cellular interactions and encapsulate (unfortunately in a very hierarchically complicated way) the insight we are after.

6.3 Smooth Motions as Axes of Inquiry

So can we then propose something as an alternative? We have hinted many times now at what we *cannot* know, but have we covered what it is that we *can* know? Take a minute to consider, for example, the previously laid out implications of chapters 4 and 5. Despite the general distaste that many biophysicists and theoretical chemists alike express for simulations in which reaction coordinates are "chosen," we have shown with great clarity and certainty the power of such techniques as applied to a fuzzy set of molecular structures with a smooth continuum of states between them. Recall the example of a child swinging a toy above his head used to explain periodicity in the introduction to chapter 4. The main purpose of such a demonstration is to show that the ensemble techniques we use as a metric are simply measures of *correlated* motions across time, and that there are essentially an infinite number of nuanced variations (whether short lived or long lived) which could lead to identical results. Ironically, this is not a difficult concept for most to handle because it considers an ensemble average as the measure, and we are perfectly comfortable with formalized "indeterminacy"

when it comes to ensemble averaging. Why is it then difficult for us to think of the *individual* motions of the molecules in a similar way?

Returning to our discussion of chapter 5, let us summarize by saying that we have demonstrated one could ever so gently alter the fluctuations of a given macromolecule by first uncovering the major modes of vibrational access it has, and then using these states as a kind of axis along which to enhance sampling and arrive at concrete definitive descriptors. In our example we showed that we can significantly alter the way in which a small DNA molecule can fluctuate by simply presenting a rigid fullerene to the major groove. Despite the affinity of researchers to discuss the "tendency" of DNA to be in B form versus A form in solution, most simulations of short ds-DNA molecules actually show that nucleic acids fluctuate an average of 6 $\mathring{A}$ in RMSD (roughly the equivalent of an A to B transition) during normal simulations. By introducing a change in the environment, we answered the direct and specific question, how much energy does it require to hold the DNA in this A like state? How much energy does it require to hold it in this B like state? How does that amount change when I present a rigid, hydrophobic moiety to the major groove? This way, we can state the exact amount of energy it would take to alter each of the states in between A and B form, allowing us to deduce periodic motion and major vibrational alterations available to a *generalized* version of the molecule, not just a single iteration as is generated with simple brute force MD. Perhaps even more impressively, if the SWNT were removed *after* the fluctuations available to the dsDNA were altered, the DNA would return to its previous fluctuations somewhere between A and B as if nothing had happened at all! It strikes this author very likely that this situation, one in which the dynamics of the DNA can be significantly altered in a way that would leave absolutely no measurable structural transition, is how evolution found the most effective modes of influencing molecular entities to carry out important functions with high activation energy barriers without denaturing their previously attained structure. This is clearly the case for how transcription factors work while initiating DNA based functions; they must stack many "gentle" touches sequentially so as

to effect an aggregate *concerted* fluctuation. This concerted fluctuation can then exceed some activation energy, and induces a conformational change. The idea of a clock, or key, or any such picture finds no basis in this process, it is entirely probabilistically driven, and all of the motions are equally probable given the altering environment of the molecule. It is quite fascinating that simply by slow and gentle "nudges" towards a new fluctuation which can then exceed the activation energy, major cellular revisions can take place. If any of the nudges are not available, however, the DNA and the proteins involved are left without any measurable change to their overall structure, preserving them for another try in the future.

This brings us to the conclusion of our study, and perhaps the single most important implication as it was discussed in the the introduction. Conformational changes, no matter how dramatic and sudden (or *clocklike*) they appear to be effected, are likely strung along slowly when viewed from the nanosecond - microsecond timescale. Particularly in the case of nucleic acids, these large scale conformational changes are carried out by *correlated* sets of conformational fluctuations, encouraged (but definitely not caused) by new neighbors or cell signaling events in an ordered, concerted manner. Furthermore, when one considers the fluid nature of RNA and its promiscuous access to all timescales and molecular entities alike, it seems probable that this molecule has the most gentle control (or perhaps we should say finely tuned correlation of motions) of all the macromolecules, and it seems likely that we, as humans, are the recipients of an extremely long line of concerted, guided, motions which started sometime 3 or 4 billion years ago by this very old, yet very functional molecule.

Bibliography

[1] Fareed Aboul-ela, Jonathan Karn, and Gabriele Varani. Structure of hiv-1 tar rna in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic acids research*, 24(20):3974–3981, 1996.

[2] 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.

[3] Evgenia N Nikolova, Gavin D Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. Probing sequence-specific dna flexibility in a-tracts and pyrimidine-purine steps by nuclear magnetic resonance 13c relaxation and molecular dynamics simulations. *Biochemistry*, 51(43):8654–8664, 2012.

[4] Aaron T Frank, Andrew C Stelzer, Hashim M Al-Hashimi, and Ioan Andricioaei. Constructing rna dynamical ensembles by combining md and motionally decoupled nmr rdcs: new insights into rna dynamics and adaptive ligand recognition. *Nucleic acids research*, 37(11):3670–3679, 2009.

[5] Loïc Salmon, Gavin Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. A general method for constructing atomic-resolution rna ensembles using nmr residual dipolar couplings: the basis for interhelical motions revealed. *Journal of the American Chemical Society*, 135(14):5457–5466, 2013.

[6] Catherine Musselman, Qi Zhang, Hashim Al-Hashimi, and Ioan Andricioaei. Referencing strategy for the direct comparison of nuclear magnetic resonance and molecular dynamics motional parameters in rna. *The Journal of Physical Chemistry B*, 114(2):929–939, 2009.

[7] Eörs Szathmáry. The origin of the genetic code: amino acids as cofactors in an rna world. *Trends in genetics*, 15(6):223–229, 1999.

[8] Sven Siebert. Common sequence structure properties and stable regions in rna secondary structures. 2006.

[9] Alberto Pérez, F Javier Luque, and Modesto Orozco. Dynamics of b-dna on the microsecond time scale. *Journal of the American Chemical Society*, 129(47):14739–14745, 2007.

[10] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[11] Ralf Dahm. Friedrich miescher and the discovery of dna. *Developmental Biology*, 278(2):274–288, 2005.

[12] James D Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[13] Ignacio Tinoco Jr and Carlos Bustamante. How rna folds. *Journal of molecular biology*, 293(2):271–281, 1999.

[14] TE Cloutier and Jonathan Widom. Dna twisting flexibility and the formation of sharply looped protein–dna complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3645–3650, 2005.

[15] FHe Crick. Linking numbers and nucleosomes. *Proceedings of the National Academy of Sciences*, 73(8):2639–2643, 1976.

[16] M Nirenberg, P Leder, M Bernfield, R Brimacombe, J Trupin, F Rottman, and C O'neal. Rna codewords and protein synthesis, vii. on the general nature of the rna code. *Proceedings of the National Academy of Sciences of the United States of America*, 53(5):1161, 1965.

[17] John S Mattick. The hidden genetic program of complex organisms. *Scientific American*, 291(4):60, 2004.

[18] Alan Cooper. Protein fluctuations and the thermodynamic uncertainty principle. *Progress in biophysics and molecular biology*, 44(3):181–214, 1984.

[19] Cécile Morette DeWitt. Feynman's path integral. *Communications in Mathematical Physics*, 28(1):47–67, 1972.

[20] DA McQuarrie. Statistical mechanics, 1976. *Happer and Row, New York*.

[21] Jerry B Marion and Stephen T Thornton. *Classical dynamics of particles and systems*. Saunders College Pub., 1995.

[22] Shinji Sunada, Nobuhiro Go, and Patrice Koehl. Calculation of nuclear magnetic resonance order parameters in proteins by normal mode analysis. *The Journal of chemical physics*, 104(12):4768–4775, 1996.

[23] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Current opinion in structural biology*, 9(2):164–169, 1999.

[24] Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Current opinion in structural biology*, 10(2):165–169, 2000.

[25] J Andrew McCammon and Stephen C Harvey. *Dynamics of proteins and nucleic acids*. Cambridge University Press, 1988.

[26] Xiang-Jun Lu and Wilma K Olson. 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic acids research*, 31(17):5108–5121, 2003.

[27] Jeff Wereszczynski and Ioan Andricioaei. On structural transitions, thermodynamic equilibrium, and the phase diagram of dna and rna duplexes under torque and tension. *Proceedings of the National Academy of Sciences*, 103(44):16200–16205, 2006.

[28] Richard E Dickerson, Horace R Drew, Benjamin N Conner, Richard M Wing, Albert V Fratini, Mary L Kopka, et al. The anatomy of a-, b-, and z-dna. *Science*, 216(4545):475–485, 1982.

[29] Chun Yoon, Gilbert G Privé, David S Goodsell, and Richard E Dickerson. Structure of an alternating-b dna helix and its relationship to a-tract dna. *Proceedings of the National Academy of Sciences*, 85(17):6332–6336, 1988.

[30] Uli Schmitz, Gerald Zon, and Thomas L James. Deoxyribose conformation in [d (gtatatac)] 2: evaluation of sugar pucker by simulation of double-quantum-filtered cosy cross-peaks. *Biochemistry*, 29(9):2357–2368, 1990.

[31] Jason D Kahn, Elizabeth Yun, and Donald M Crothers. Detection of localized dna flexibility. 1994.

[32] Horace R Drew and Andrew A Travers. Dna bending and its relation to nucleosome positioning. *Journal of molecular biology*, 186(4):773–790, 1985.

[33] JY Lee, Burak Okumus, DS Kim, and Taekjip Ha. Extreme conformational diversity in human telomeric dna. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):18938–18943, 2005.

[34] Donald M Gray, Su-Hwi Hung, and Kenneth H Johnson. Absorption and circular dichroism spectroscopy of nucleic acid duplexes and triplexes. *Methods in enzymology*, 246:19–34, 1994.

[35] Thierry Dauxois, Michel Peyrard, and AR Bishop. Entropy-driven dna denaturation. *Phys. Rev. E*, 47(1):R44–R47, 1993.

[36] Ioulia Rouzina and Victor A Bloomfield. Force-induced melting of the dna double helix 1. thermodynamic analysis. *Biophysical journal*, 80(2):882–893, 2001.

[37] Helen G Hansma, Kenichi Kasuya, and Emin Oroudjev. Atomic force microscopy imaging and pulling of nucleic acids. *Current opinion in structural biology*, 14(3):380–385, 2004.

[38] Haran T. E. The unique structure of a-tracts and intrinsic dna bending. *Q. Rev. Biophys.*, 42:41, 2009.

[39] A. K. Mazur. Anharmonic torsional stiffness of DNA revealed under small external torques. *Phys. Rev. Lett.*, 105:018102, Jun 2010.

[40] Kathleen B Hall. Rna in motion. *Current opinion in chemical biology*, 12(6):612–618, 2008.

[41] Paul G Higgs. Rna secondary structure: physical and computational aspects. *Quarterly reviews of biophysics*, 33(03):199–253, 2000.

[42] Thomas Hermann and Dinshaw J Patel. Rna bulges as architectural and recognition motifs. *Structure*, 8(3):R47–R54, 2000.

[43] Aurelie Lescoute, Neocles B Leontis, Christian Massire, and Eric Westhof. Recurrent structural rna motifs, isostericity matrices and sequence alignments. *Nucleic Acids Research*, 33(8):2395–2409, 2005.

[44] Philippe Brion and Eric Westhof. Hierarchy and dynamics of rna folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997.

[45] Neocles B Leontis and Eric Westhof. Analysis of rna motifs. *Current opinion in structural biology*, 13(3):300–308, 2003.

[46] Jan Ferner, Alessandra Villa, Elke Duchardt, Elisabeth Widjajakusuma, Jens Wöhnert, Gerhard Stock, and Harald Schwalbe. Nmr and md studies of the temperature-dependent dynamics of rna ynmg-tetraloops. *Nucleic acids research*, 36(6):1928–1940, 2008.

[47] Martin Zacharias and Paul J Hagerman. Bulge-induced bends in rna: quantification by transient electric birefringence. *Journal of molecular biology*, 247(3):486–500, 1995.

[48] Tamás Kiss. Small nucleolar rnas: an abundant group of noncoding rnas with diverse cellular functions. *Cell*, 109(2):145–148, 2002.

[49] David P Bartel, Maria L Zapp, Michael R Green, and Jack W Szostak. Hiv-1 rev regulation involves recognition of non-watson-crick base pairs in viral rna. *Cell*, 67(3):529–536, 1991.

[50] Jung C Lee and Robin R Gutell. Diversity of base-pair conformations and their occurrence in rrna structure and rna structural motifs. *Journal of molecular biology*, 344(5):1225–1249, 2004.

[51] A Lescoute and E Westhof. The interaction networks of structured rnas. *Nucleic acids research*, 34(22):6587–6604, 2006.

[52] Michael F Bardaro, Zahra Shajani, Krystyna Patora-Komisarska, John A Robinson, and Gabriele Varani. How binding of small molecule and peptide ligands to hiv-1 tar alters the rna motional landscape. *Nucleic acids research*, 37(5):1529–1540, 2009.

[53] Gary D Stormo and Yongmei Ji. Do mrnas act as direct sensors of small molecules to control their expression? *Proceedings of the National Academy of Sciences*, 98(17):9465–9467, 2001.

[54] Nils G Walter. Structural dynamics of catalytic rna highlighted by fluorescence resonance energy transfer. *Methods*, 25(1):19–30, 2001.

[55] Jamie H Cate, Marat M Yusupov, Gulnara Zh Yusupova, Thomas N Earnest, and Harry F Noller. X-ray crystal structures of 70s ribosome functional complexes. *Science*, 285(5436):2095–2104, 1999.

[56] Christopher Hammond and Christopher Hammond. *Basics of crystallography and diffraction*, volume 214. Oxford, 2001.

[57] Joseph P Hornak. Basics of nmr, 1997.

[58] G Marius Clore, Attila Szabo, Ad Bax, Lewis E Kay, Paul C Driscoll, and Angela M Gronenborn. Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *Journal of the American Chemical Society*, 112(12):4989–4991, 1990.

[59] Eric R Henry and Attila Szabo. Influence of vibrational motion on solid state line shapes and nmr relaxation. *The Journal of chemical physics*, 82(11):4753–4761, 1985.

[60] Eike Brunner. Residual dipolar couplings in protein nmr. *Concepts in Magnetic Resonance*, 13(4):238–259, 2001.

[61] Alexander D MacKerell, Bernard Brooks, Charles L Brooks, Lennart Nilsson, Benoit Roux, Youngdo Won, and Martin Karplus. Charmm: the energy function and its parameterization. *Encyclopedia of computational chemistry*, 1998.

[62] Scott J Weiner, Peter A Kollman, David A Case, U Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular

mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984.

[63] Lennart Nilsson and Martin Karplus. Empirical energy functions for energy minimization and dynamics of nucleic acids. *Journal of computational chemistry*, 7(5):591–616, 1986.

[64] Alexander D MacKerell Jr, Joanna Wiorkiewicz-Kuczera, and Martin Karplus. An all-atom empirical energy function for the simulation of nucleic acids. *Journal of the American Chemical society*, 117(48):11946–11975, 1995.

[65] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[66] Mike P Allen and Dominic J Tildesley. Computer simulation of liquids. 1989.

[67] Dominique Levesque and Loup Verlet. Molecular dynamics and time reversibility. *Journal of Statistical Physics*, 72(3-4):519–537, 1993.

[68] Dennis C Rapaport. *The art of molecular dynamics simulation*. Cambridge university press, 2004.

[69] Elizabeth J Denning, U Priyakumar, Lennart Nilsson, and Alexander D Mackerell. Impact of 2-hydroxyl sampling on the conformational properties of rna: Update of the charmm all-atom additive force field for rna. *Journal of computational chemistry*, 32(9):1929–1943, 2011.

[70] Nicolas Foloppe and Alexander D MacKerell Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, 21(2):86–104, 2000.

[71] Alexander D Mackerell and Nilesh K Banavali. All-atom empirical force field for nucleic acids: Ii. application to molecular dynamics simulations of dna and rna in solution. *Journal of Computational Chemistry*, 21(2):105–120, 2000.

[72] Elzbieta Kierzek, Anna Pasternak, Karol Pasternak, Zofia Gdaniec, Ilyas Yildirim, Douglas H Turner, and Ryszard Kierzek. Contributions of stacking, preorganization, and hydrogen bonding to the thermodynamic stability of duplexes between rna and 2-o-methyl rna with locked nucleic acids. *Biochemistry*, 48(20):4377–4387, 2009.

[73] Darrin M York, Tom A Darden, and Lee G Pedersen. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: a comparison of the ewald and truncated list methods. *The Journal of chemical physics*, 99(10):8345–8348, 1993.

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

[75] Simon W de Leeuw, John W Perram, and Edgar R Smith. Simulation of electrostatic systems in periodic boundary conditions. i. lattice sums and dielectric constants. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 373(1752):27–56, 1980.

[76] Catherine Musselman, Hashim M Al-Hashimi, and Ioan Andricioaei. ired analysis of tar rna reveals motional coupling, long-range correlations, and a dynamical hinge. *Biophysical journal*, 93(2):411–422, 2007.

[77] Tim Zeiske, Kate A Stafford, Richard A Friesner, and Arthur G Palmer. Starting-structure dependence of nanosecond timescale intersubstate transitions and reproducibility of md-derived order parameters. *Proteins: Structure, Function, and Bioinformatics*, 81(3):499–509, 2013.

[78] David E Shaw, Martin M Deneroff, Ron O Dror, Jeffrey S Kuskin, Richard H Larson, John K Salmon, Cliff Young, Brannon Batson, Kevin J Bowers, Jack C Chao, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

[79] Takehiko Shimanouchi. Tables of molecular vibrational frequencies consolidated. volume i. Technical report, DTIC Document, 1972.

[80] Vincent Kräutler, Wilfred F van Gunsteren, and Philippe H Hünenberger. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry*, 22(5):501–508, 2001.

[81] CL Brooks III, MW Balk, and SA Adelman. Dynamics of liquid state chemical reactions: Vibrational energy relaxation of molecular iodine in liquid solution. *The Journal of Chemical Physics*, 79(2):784–803, 1983.

[82] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[83] John E Straub and D Thirumalai. Exploring the energy landscape in proteins. *Proceedings of the National Academy of Sciences*, 90(3):809–813, 1993.

[84] Joseph R Lakowicz and Gregorio Weber. Quenching of protein fluorescence by oxygen. detection of structural fluctuations in proteins on the nanosecond time scale. *Biochemistry*, 12(21):4171–4179, 1973.

[85] Stephen J Benkovic and Sharon Hammes-Schiffer. A perspective on enzyme catalysis. *Science*, 301(5637):1196–1202, 2003.

[86] Dominique Bourgeois, Beatrice Vallone, Friedrich Schotte, Alessandro Arcovito, Adriana E Miele, Giuliano Sciara, Michael Wulff, Philip Anfinrud, and Maurizio Brunori. Complex landscape of protein structural dynamics unveiled by nanosecond laue crystallography. *Proceedings of the National Academy of Sciences*, 100(15):8704–8709, 2003.

[87] Alan S Verkman. Solute and macromolecule diffusion in cellular aqueous compartments. *Trends in biochemical sciences*, 27(1):27–33, 2002.

[88] Bertil Halle. The physical basis of model-free analysis of nmr relaxation data from proteins and complex fluids. *The Journal of chemical physics*, 131(22):224507, 2009.

[89] DICK D Mosser, NICHOLAS G Theodorakis, and RICHARD I Morimoto. Coordinate changes in heat shock element-binding activity and hsp70 gene transcription rates in human cells. *Molecular and cellular biology*, 8(11):4736–4744, 1988.

[90] Richard Martin Ballew, Jobiah Sabelko, and Martin Gruebele. Observation of distinct nanosecond and microsecond protein folding events. *Nature Structural & Molecular Biology*, 3(11):923–926, 1996.

[91] De Witt Sumners. Untangling dna. *The Mathematical Intelligencer*, 12(3):71–80, 1990.

[92] Grégoire Altan-Bonnet, Albert Libchaber, and Oleg Krichevsky. Bubble dynamics in double-stranded dna. *Physical review letters*, 90(13):138101, 2003.

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

[94] Malka Kitayner, Haim Rozenberg, Remo Rohs, Oded Suad, Dov Rabinovich, Barry Honig, and Zippora Shakked. Diversity in dna recognition by p53 revealed by crystal structures with hoogsteen base pairs. *Nature structural & molecular biology*, 17(4):423–429, 2010.

[95] Nilesh K Banavali and Benoît Roux. Free energy landscape of a-dna to b-dna conversion in aqueous solution. *Journal of the American Chemical Society*, 127(18):6866–6876, 2005.

[96] Qi Zhang, Xiaoyan Sun, Eric D Watt, and Hashim M Al-Hashimi. Resolving the motional modes that code for rna adaptation. *Science*, 311(5761):653–656, 2006.

[97] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

[98] Wilma K Olson, Andrey A Gorin, Xiang-Jun Lu, Lynette M Hock, and Victor B Zhurkin. Dna sequence-dependent deformability deduced from protein–dna crystal complexes. *Proceedings of the National Academy of Sciences*, 95(19):11163–11168, 1998.

[99] Andrew C Stelzer, Aaron T Frank, Jeremy D Kratz, Michael D Swanson, Marta J Gonzalez-Hernandez, Janghyun Lee, Ioan Andricioaei, David M Markovitz, and Hashim M Al-Hashimi. Discovery of selective bioactive small molecules by targeting an rna dynamic ensemble. *Nature chemical biology*, 7(8):553–559, 2011.

[100] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[101] Bernard R Brooks, Charles L Brooks, Alexander D MacKerell, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bar-

tels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[102] David A Case, Thomas E Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.

[103] Alexander D MacKerell, Nilesh Banavali, and Nicolas Foloppe. Development and current status of the charmm force field for nucleic acids. *Biopolymers*, 56(4):257–265, 2000.

[104] Matthew A Young, G Ravishanker, and DL Beveridge. A 5-nanosecond molecular dynamics trajectory for b-dna: analysis of structure, motions, and solvation. *Biophysical journal*, 73(5):2313–2336, 1997.

[105] Natalie A Davis, Sangita S Majee, and Jason D Kahn. Tata box dna deformation with and without the tata box-binding protein. *Journal of molecular biology*, 291(2):249–265, 1999.

[106] Ludmila V Yakushevich. *Nonlinear physics of DNA*. John Wiley & Sons, 2006.

[107] David E Shaw, Ron O Dror, John K Salmon, JP Grossman, Kenneth M Mackenzie, Joseph A Bank, Cliff Young, Martin M Deneroff, Brannon Batson, Kevin J Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*, pages 1–11. IEEE, 2009.

Appendices

A Probing Sequence Specific DNA Flexibility in A-tracts and Pyrimidine Purine Steps by NMR ^{13}C Relaxation and MD Simulations

A.1 Abstract

Sequence-specific DNA flexibility plays a key role in a variety of cellular interactions that are critical for gene packaging, expression, and regulation. Yet, few studies have experimentally explored the sequence dependence of DNA dynamics that occur on biologically relevant timescales. Here, we use nuclear magnetic resonance (NMR) carbon spin relaxation combined with molecular dynamics (MD) simulations to examine the picosecond to nanosecond dynamics in a variety of dinucleotide steps as well as in varying length homopolymeric $A_n{\cdot}T_n$ repeats (A_n-tracts, n = 2, 4 and 6) that exhibit unusual structural and mechanical properties. We extend the NMR spin relaxation timescale sensitivity deeper into the nanosecond regime by using glycerol and a longer DNA duplex to slow down overall tumbling. Our studies reveal a structurally unique A-tract core (for n > 3) that is uniformly rigid, flanked by junction steps that show increasing sugar flexibility with A-tract length. High sugar mobility is observed at pyrimidine residues at the A-tract junctions, which is encoded at

the dinucleotide level (CA, TG and CG steps) and increases with A-tract length. The MD simulations reproduce many of these trends, particularly the overall rigidity of A-tract base and sugar sites, and suggest that the sugar-backbone dynamics could involve transitions in sugar pucker and phosphate backbone BI $\Leftrightarrow$ BII equilibria. Our results reinforce an emerging view that sequence-specific DNA flexibility can be imprinted in dynamics occurring deep within the nanosecond time regime that is difficult to characterize experimentally at the atomic level. Such large amplitude sequence-dependent backbone fluctuations might flag the genome for specific DNA recognition.

A.2 Introduction

The DNA double helix is not simply a uniform structure that carries the codon message for gene expression. Rather, different nucleotide sequences show distinct propensities to deform bend and twist on their own [1, 2] or upon binding to protein and drug targets [3, 4, 5]. Sequence-specific variations in DNA structure and flexibility form the basis of indirect DNA readout by regulatory proteins [6] and can also guide the positioning of nucleosomes along the genome [7]. These dynamic flags constitute a new layer of genetic information that remains poorly understood.

A crucial step towards decoding the functional roles of DNA sequence-specific mobility is to elucidate how the dynamic properties of duplex DNA vary with nucleotide sequence. Surveys of naked and protein-bound DNA crystal structures together with knowledge-based computational models have provided significant insight into the conformational flexibility of the DNA duplex at a dinucleotide level [3, 8] and for longer nucleotide stretches [9, 10]. These nearest-neighbor "rules" rank pyrimidine-purine (YR) steps, specifically CA, TG, TA and CG steps, as the most conformationally flexible dinucleotide sequences. Not surprisingly, these steps are frequently the loci of helical deformations in DNA assemblies

with transcription factors and their flexibility features could guide indirect readout of specific DNA sequences [11, 12, 13]. On the other end of the spectrum, purine-purine (RR) AA steps, and less so purine-pyrimidine (RY) AT steps, are the most conformationally rigid dinucleotide steps and exhibit structural parameters that vary the least with sequence context, making them more difficult to mold by proteins.

Numerous studies reveal that poly(dA)·poly(dT) stretches, so-called asymmetric An-tracts (n > 3), adopt a locally distinct and rigid B-DNA conformation that forms cooperatively and that cannot be purely described as a collection of individual AA steps as assumed by a nearest-neighbor model, reviewed by Haran [14]. This non-canonical conformation features a high propeller twist and negative inclination in A·T base pairs and a progressive narrowing of the minor groove in the 5' to 3' direction [15, 16]. The local conformational rigidity of A-tracts could explain their preferential exclusion from nucleosomes in vitro and avoidance in exon regions that are densely populated with nucleosomes in vivo [17, 18]. Thus, A-tracts could be stereochemically locked into "inflexible" frameworks that could make them less prone to interact with outside regulatory factors. The local A-tract structure tend to resist sharp bending [19] but A-tracts can induce macroscopic curvature when phased in tandem with the helical repeat [20], which is enhanced by placement of CA/TG steps at their 5' junction [21], and that is important for DNA looping in transcriptional regulation and chromatin packaging [14]. However, it remains unclear whether the local A-tract conformation or the helical bending is dynamic in nature as well as whether the global curvature of phased A-tracts originates at their junctions or is delocalized along the entire adenine stretch [14]. The ability of A-tracts to modulate DNA structure, and thus affect protein binding or enable long-range communication, has placed these unique elements at the forefront of research efforts to elucidate the relationship between structure, dynamics and function in DNA transactions.

A growing number of experimental and computational studies show that sequence-specific DNA deformability observed in crystal structures is encoded as intrinsic dynamic fluctu-

ations in naked DNA [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. For instance, the discrimination of unwanted uracil, the product of cytosine deamination, from thymine by the uracil DNA glycosylase (UNG) repair enzyme that removes the modified base from genomic DNA was shown to be dictated by differences in thermally induced opening of A · U versus A · T base pairs [35]. More recently, Hoogsteen base pairs observed in duplex DNA bound to transcription factors [36, 37] and antibiotic drugs [38] have been found to form spontaneously and sequence-specifically in naked duplex DNA [33]. Also, extruded nucleobases observed crystallographically within a junction between B-DNA and Z-DNA [39] have been shown to be sequence-specifically flexible in the context of B-DNA [32]. Moreover, experimental biophysical studies [31, 40, 41] and theoretical models [34] suggest that flexible CG steps, which are enriched in promoter regions and prone to C5-cytosine methylation as a mechanism to regulate gene expression, become more stiff and show a lower propensity to circularize or form nucleosomes. Collectively, these studies suggest that the intrinsic dynamic properties of DNA can provide a mechanism for genetic control.

Nuclear magnetic resonance (NMR) spectroscopy is a powerful technique for studying DNA sequence-specific flexibility at atomic resolution and over timescales spanning from picosecond to seconds and longer. To date, NMR carbon spin relaxation studies targeting specific biological sequences such have uncovered large-amplitude backbone motions in cytosine sugars of unmodified DNA [22, 26, 28, 42] and near sites of DNA damage [43] that may be important for specific DNA recognition. Surprisingly, no such NMR study has been used to explore sequence-specific flexibility of unusual A-tract sequences and systematically for dinucleotide steps and their dependence on immediate neighbors. Here, we use solution NMR spin relaxation techniques in conjunction with molecular dynamics (MD) simulations to probe the internal dynamics of varying length A-tracts. Our studies reveal a structurally unique A-tract with a uniformly rigid nucleotide core that exhibits a somewhat increased adenine over thymine sugar mobility. The A-tract is flanked by sequences that contain increasingly more flexible sugar moieties near the 5' and 3' A-tract junction steps with in-

creasing A-tract length. We observe unique sugar mobility in pyrimidine residues, which seems occur at nanosecond (ns) timescales based on measurements at variable temperature and viscosity, that are encoded at the dinucleotide level but that can be modulated by the A-tract length. MD simulations reproduce the majority but not all of these trends, and suggest that the sugar C1' mobility could be coupled to rapid backbone transitions in BI $\Leftrightarrow$ BII and/or sugar pucker rearrangements.

A.3 Materials and Methods

Materials and Sample Preparation

All unlabeled DNA oligonucleotides were purchased from Integrated DNA Technologies, Inc. (Coralville, IA). $^{13}C/^{15}N$-labeled labeled DNA dodecamers were synthesized in vitro by the method of Zimmer et al. [44] using a template hairpin DNA (IDT, Inc.), Klenow fragment DNA polymerase (NEB, Inc.), and uniformly $^{13}C/^{15}N$-labeled dNTPs (Isotec, Sigma-Aldrich). Single-stranded DNA products were purified by a 20% denaturing gel electrophoresis, isolated by passive elution from crushed gels and desalted on a C18 reverse-phase column (Sep-pak, Waters). Oligonucleotides were further lyophilized and complementary strands were resuspended separately in NMR buffer (15 mM sodium phosphate (pH 6.8), 25 mM sodium chloride, 0.1 mM EDTA) supplied with 10% D_2O. Sample annealing was monitored by quick ^{13}C,^{15}N HSQCs until single strand signal were not longer observed, with typically duplex concentrations of 0.5 1.0 mM for NMR studies. Unlabeled DNA constructs were prepared directly from oligonucleotides purchased from the manufacturer. Oligos were resuspended in NMR buffer at $\sim$200 μM concentration and their concentration measured by UV absorbance at 260 nm using extinction coefficients provided by the manufacturer. DNA duplexes were annealed by mixing an equal molar ratio of the complementary DNA strands, heating for 2 min at 95 °C and gradual cooling at room temperature. DNA preparations

were washed 3X in resuspension buffer by microcentrifugation using an Amicon Ultra-4 centrifugal filter (3 kDa cutoff), concentrated to a volume of ∼250 μl (∼ 2-4 mM) for NMR studies and supplied with 10% D_2O.

NMR Measurements and Analysis

All NMR experiments were performed on a Bruker Avance 600 MHz NMR spectrometer equipped with a 5mm triple-resonance cryogenic probe. Unlabeled DNA duplexes were assigned using conventional 2D 1H,1H NOESY (mixing time 175 ms) in 10% D_2O at 26 °C. Proton assignments were transferred to 2D ^{1}H,^{13}C and ^{1}H,^{15}N HSQC spectra, allowing convenient assignment of base C2H2, C6H6, C8H8, N1H1, N3H3 and sugar C1'H1' in unlabeled DNA constructs. Resonance intensities were obtained from ^{1}H,^{13}C and ^{1}H,^{15}N HSQC spectra and normalized for each type of bond vector to the intensity of a helical residue that was set to 0.1.

^{13}C relaxation rate constants R1 and R2 in $^{13}C/^{15}N$-labeled DNA dodecamers were measured using a 2D relaxation experiment [45] for base C2, C6, and C8, and sugar C1' spins using a 3.5 kHz spinlock field strength and a spinlock carrier centered at C6 (for C2/C6/C8) or C1' resonances. Spinlock powers were sufficiently high to suppress undesired chemical exchange contributions and ensure Hartmann-Hahn contributions of < 1 % for JCC ~ 10 Hz and < 0.1 % for JCC ~ 1 Hz. Relaxation data were collected with 8 scans (6 to 7 hrs) and delay series (20, 100, 250, 450 (X3) ms) for R_1 and (4, 16, 32, 48 (X3) ms) for $R_{1\rho}$ with triplicate measurements for error estimation. Relaxation profiles were processed with nmrPipe [46] and relaxation rate constants determined by fitting the resonance intensities to mono-exponential decays using Mathematica 6.0 (Wolfram Research, Inc.). R_2 relaxation rates were computed from R_1 and using the relationship $R_2 = (R_{1\rho} - R_1 \cos^2\theta)/\sin^2\theta$ [47]. Relative order parameters S^2_{Rel} were computed as $2R_2$ - R_1 values normalized to the largest value from a helical region for each carbon type (C2, C6, C8 or C1') and also each

residues type for C8 and C6, set to unity [45, 48, 49]. Hydrodynamic and S^2 predictions were conducted with HydroNMR [50] and an in-house software by employing a previously described protocol with a DNA dodecamer model constructed with 3DNA [51] or obtained from MD simulations, [45] assuming anisotropic diffusion and using only R_1 and R_2 values (without heteronuclear NOEs).

Molecular Dynamics Simulations

Atomic coordinates were built using the Nucleic Acid Builder (part of AmberTools [52]) of sequence (A_2, A_4, and A_6-DNA) in ideal helical B-form DNA. The structures were solvated with water and Na+ ions using Visual Molecular Dynamics[53] in a 64 x 64 x 64 Angstrom cube, with 25 Na+ ions and 3 Cl- ions to neutralize charge and bring molarity to experimental conditions. All structures were heated gradually with harmonic constraints placed on sugar-phosphate backbone atoms from 0 K to 300 K in 150,000 steps, with 1 fs time-steps in NAMD using a Langevin thermostat [54] with the CHARMM force field [55, 56]. Harmonic constraints were gradually released over 300 ps and the systems were equilibrated each for 10 ns. Independent trajectory ensembles were then generated from 10 independent Maxwel-Boltzmann distributed initial conditions for each sequence, producing 30 uncorrelated trajectories of 10 ns each.

S^2 values were determined using a generalized Lipari-Szabo model free approach [57, 58] in which the bond-bond autocorrelation function for the second order Legendre polynomial describing rotational decorrelation is parameterized by the sum of two-exponential forms [59] to obtain amplitude (S^2) and correlation time (τ) according to the following relationship:

$$C(t) = S^2 + (1 - S_f^2)e^{-t/\tau_f} + (S_f^2 - S^2)e^{-t/\tau_s} \tag{A.1}$$

where $S2 = S_f^2 S_s^2$ is the plateau of the function and subscripts f and s refer to fast and slow motions respectively, which are assumed to be uncorrelated. As a check of convergence, S^2 values were also calculated from the bond vector Cartesian coordinate equilibrium expression given by Szabo and Henry, [60] which gave good agreement with the extended exponential fit. Overall tumbling was removed by least squares fit alignment of heavy atoms in VMD. Time correlation functions were calculated using the CHARMM software package [59, 61]. Plateau values at 1ns (i.e., a tenth of the total trajectory time [62]) were determined by averaging the tail autocorrelation function values and the results were then averaged across ensembles. Sugar pucker statistics and sugar-backbone dihedral angles were calculated from the MD trajectories using 3DNA [51].

A.4 Results

A-tract Specific Dynamics from NMR Spectra and Resonance Intensities

We used solution NMR to study the dynamic properties of three uniformly $^{13}C/^{15}N$-labeled DNA dodecamers containing two (A_2-DNA), four (A_4-DNA), and six (A_6-DNA) adjacent adenines capped by GC-rich helices (Figure 1A). A_6-DNA appears in a context (5' CA6T) commonly encountered in kinetoplast DNA, which was originally found to exhibit microscopic bending when regularly phased with the helical repeat (10.5 base pairs per turn) [63].

We first examined the NMR spectral variations as a function of A-tract length. 1H,^{13}C HSQC and 2D NOESY spectra of A_6-DNA and A_4-DNA displayed chemical shifts (CS) and nuclear Overhauser effect (NOE) connectivities characteristic of asymmetric A-tracts, [64, 65] which have been shown to deviate from a canonical B-DNA conformation [15]. For example, we observed strong interstrand NOE cross-peaks between the nucleobase H2 proton of adenine and the sugar H1' or imino H3 protons of the 3'-neighboring thymine on the complementary

strand (data not shown), which has previously been correlated with minor groove compression and a large propeller twist [64, 65]. In addition, purine H8/H1' and pyrimidine H3/H1' protons typically displayed upfield and downfield shifted CSs, respectively, characteristic of A-tract sequences (Figure 1B). The highly unusual upfield shifted proton CS of the cytosine and adenine sugar moiety at the 5' CA/TG junction also represent unique spectral signature of the distinct A-tract conformation [64, 65].

The above NMR spectroscopic signature of the A-tract diminish slightly from A_6-DNA to A_4-DNA, and are no longer observed in A_2-DNA does not adopt the unusual conformation of longer adenine runs or induce any appreciable global curvature when periodically phased relative to a random sequence (Figure 1B). Specifically, curtailing the A-tract from six to two AT base pairs caused a downfield shift for adenine H8/H1' and an upfield shift for thymine H3/H1' protons, in the direction of the CS space generally occupied by heterogeneous sequences (Figure 1B). Certain base and sugar protons at the common 5' CA/TG junction (C15 H1' and A16 H8/H1') also experienced significant downfield shift from A_6-DNA to A_4-DNA (up to 0.04 ppm) and once again an even larger shift (up to 0.2 ppm) from A_4-DNA to A_2-DNA. Such sizeable perturbations in proton CS are not expected to arise due to remote changes in sequence (> 2 base pairs away) and point to conformational changes within the A-tract that vary with A-tract length.

We further investigated the dynamic behavior of DNA dodecamers by comparing spin-normalized resonance intensities for nucleobase C2H2, C6H6, C8H8 and deoxyribose C1'H1' spin pairs (Figure 2C). The resonance intensity provides a qualitative assessment regarding the relative mobility for a given site over timescales spanning picoseconds to milliseconds and/or variations in bond vector orientation relative to the magnetic field [45, 66, 67]. Generally, high peak intensity or line narrowing is associated with increased net dynamics (local or collective) at a given site on the pico-to-nanosecond (ps-ns) timescales, whereas weak peak intensities could reflect relative rigidity on the ps-ns timescale, micro-to-millisecond (?s-ms)

conformational exchange and/or an orientation for the bond vector that is more parallel with respect to the long duplex axis. As expected, we observed increased intensities for residues near the terminal end likely arising from end-fraying at ps-ns timescales. Although, in general, the intensities observed within the duplex are quite uniform, some unique dynamic signatures are apparent. We observed reduced intensities at the 5' CA/TG A-tract junction (A16 C8H8 and C2H2), which points to ?s-ms chemical exchange that we previously showed corresponds to transient excursions towards non-canonical Hoogsteen base pairs.33 In addition, we observe elevated intensities for cytosine and thymine sugar C1'H1' sites in CA, TG and CG steps that indicate elevated ps-ns sugar flexibility similar to those previously been observed for cytosines in CG steps embedded inside a different DNA sequence[28]. Interestingly, these normalized intensities increase by lowering the temperature from 40 oC to 10 oC (Figure 2C). This suggests the presence of sugar-backbone fluctuations occurring at ns timescales that are masked by overall rotational diffusion; lowering the temperature decouples the two motions by slowing down overall diffusion, allowing better resolution of the local dynamics [66, 67].

Picosecond-to-nanosecond Dynamics from Carbon Spin Relaxation Measurements.

We used carbon ^{13}C spin relaxation measurements [68, 69] to more quantitatively characterize ps-ns dynamics in the three DNA constructs. Specifically, we measured longitudinal (R1) and transverse (R2) ^{13}C spin relaxation data for C2, C6, C8 and C1'. The measured R1 and R2 values were then used to compute a relative order parameter, S^2_{rel} [49, 57], which provides an estimate for relative motional amplitudes across different sites (Figure 2B). The value of ranges between zero for a highly flexible site to one for a perfectly rigid site. The values were normalized independently for each carbon type relative to the most rigid site [49].

The S^2_{rel} values reinforced many of the trends obtained from analysis of the resonance intensities (Figure 2). The nucleobases of non-terminal residues were uniformly rigid across different residues and DNA constructs (S^2_{rel} range of 0.94 to 1.0.) (Figure 2). The sugar moieties exhibited larger variations in the S^2_{rel} values that cannot be accounted for by typical variations in C1'H1' vector orientation (Figure 2). Within the A-tract, thymine C1' sites exhibited the lowest flexibility with fairly uniform S^2_{rel} values approaching unity. Somewhat higher sugar flexibility was observed at the complementary adenine residues in the longer A-tracts of A_6-DNA and A_4-DNA (average S^2_{rel} 0.92) (A17 C1', Figure 2). The highest flexibility was seen for the second adenine from the 5' junction in A_6-DNA that gradually diminished with shortening of the A-tract (from 0.86 to 0.98). Similar A-tract dependent sugar-backbone dynamics were observed for residues at the A-tract junctions, including the common G10 and T9 at the 5' junction and the variable adenine (A17, A19 and A21) and thymine (T18, T20, and T22) at the 3' junction (Figure 2). Overall, the pattern of enhanced sugar-backbone dynamics at the junctions with longer A-tracts correlated well with the conformational changes observed by NMR chemical shift and NOE data, which indicates that the increased local mobility arises in part from a shift towards a distinct conformation. Once again, we observed elevated sugar-backbone pyrimidine dynamics in YR dinucleotide steps, specifically for CA/TG (T9 and C15 in all DNAs; C2 and T22 in A_2-DNA) and CG (C3 in A_4-DNA; C5 and C19 in A_2-DNA) steps with greater mobility observed in cytosine (S^2_{rel} 0.65 0.78) as compared to thymine (S^2_{rel} 0.82 0.86) sugars. These motions were less dependent on A-tract length when the YR step was placed at the 5' A-tract junction. Thus, it follows that the ps-ns motions at YR sites that are known to be flexible is encoded at the sequence-specific dinucleotide level rather than relative position to or length of A-tract.

Dynamics of A-tract and Dinucleotide Sequences from MD Simulations

Next, we conducted an analysis of the set of ten 10 ns molecular dynamics (MD) simulations for each of the sequences to gain further insights into the dynamics observed using NMR relaxation. We first compared results from the MD simulations with the NMR data by computing generalized order parameters (S^2) from the autocorrelation function averaged over ten simulation runs (Figure 3A). These order parameters were converted into S^2_{rel} values using an approach analogous to the NMR relaxation analysis (Figure 3B). Below, we focus on trends rather than quantitative comparison of S^2_{rel} values given the relatively short MD simulations and that the determination of S^2_{rel} rather than absolute S^2 values complicates quantitative comparisons.

In agreement with NMR data, uniform and high S^2_{rel} values were observed for non-terminal nucleobase sites, independent of A-tract length (Figure 3B). Likewise, overall lower S^2_{rel} values and higher disorder was observed for terminal nucleobase and particularly sugar C1' sites that is consistent with end-fraying effects (Figure 3B). More importantly, the simulations captured the decreased and uniform mobility for the thymine and most adenine sugars within the A-tract core and the tendency for increased sugar mobility at the junction residues of longer A-tracts as compared to the central residues. Similar to NMR observations, cytosine and thymine sugars in YR steps displayed lower S^2_{rel} values relative to pyrimidines in central A-tract positions, although their flexibility was underestimated by MD as compared to NMR.

At the same time, many of the specific trends observed by NMR were not very well reproduced by the MD simulations. Notably, the increasing sugar C1' flexibility with A-tract length at the shared 5' G10 and T9 residues was not observed in the MD data. Instead, MD C1' S^2_{rel} values for these sites exhibited little to no variation with A-tract length, with the thymine being more rigid and the guanine being more flexible than measured by NMR. More

generally, internal C1' spins exhibited greater dynamic variability in MD simulations as the nucleotide sequence became more heterogeneous from A_6-DNA to A_2-DNA, especially due to increase in guanine sugar dynamics in A_2-DNA that were not observed by experiment. These discrepancies may be due to insufficient sampling within the simulation timeframe, uncertainties in the structures used to carry out the simulations, or may represent deficiencies in the force field. The MD simulations also show behavior that is not observed by NMR, including the observation of elevated motions in C6 sites for A_6-DNA and A_4-DNA but not for A_2-DNA (Figure 3A). These motions are not expected based on previous NMR/MD studies [26, 28] model-free S^2_{rel} calculations performed here (data not shown) [45]. Therefore, these differences likely do not represent true differences in dynamics between C8H8 and C6H6 bond vectors but, rather, signify an issue with nucleobase force field parameterization that may stem from more homogeneous sequences (i.e., A-tracts) not being used in initial parameter optimization.

Not withstanding some of the discrepancies between NMR and MD, good agreement was obtained for the general trends of intrinsic DNA mobility within A-tracts, some junction sites and flexible pyrimidine resides in YR steps. Therefore, we examined the MD trajectories more closely to gain insight into the molecular motions that underlie the observed variations in sugar and base S^2_{rel} . Analysis of sugar pucker distributions and time-dependent fluctuations revealed that purines, especially in A-tracts, adopted primarily South (S, C2'-endo/C3'-exo) sugar pucker angles with rare and short-lived transitions towards North (N, C3'-endo/C2'-exo) conformers, while pyrimidines exhibited greater diversity in sugar pucker angles with more frequent and long-lived transitions to non-canonical North and East (E, O4'endo) conformers (Figure S1). For example, T9 at the 5' A-tract junction and in a TG step occupied the C3'-endo state at least 20% of the time, which gradually increased to about 50% with longer A-tracts. Core A-tract thymines also exhibited elevated C3'-endo populations relative to their adenine partners. Interestingly, the broadest sugar pucker distribution with a significant fraction of E states (30%) was adopted by thymines in AT steps

and was independent of A-tract length. Thus, the greater population of non-canonical C3'-endo puckers that entail large-amplitude sugar motions (150%) observed for thymine and cytosine residues by MD could partially account for the reduced sugar C1' order parameters observed at YR steps by NMR. The higher proportion of C3'-endo sugar puckers in A-tract thymines and especially in cytosines is also reflected in their more downfield shifted C1' chemical shifts than for purines.33 However, the appreciable C3'-endo populations in central A-tract thymines versus adenines could not explain the lower pyrimidine sugar mobility there. Sugar repuckering events were always accompanied by much lower amplitude ($< 50\%$) changes in the glycosidic torsion angle χ towards a high *anti* base orientation, that can also explain the absence of increased mobility in DNA bases for non-terminal sites of increased sugar mobility.

We further examined the equilibrium between BI $\Leftrightarrow$ BII backbone phosphate conformers that could potentially give rise to high-amplitude sugar-backbone motions. The major BI and minor BII backbone phosphate states are determined by the difference in ϵ and ζ dihedral angles (ϵ - ζ < 0 for BI and ϵ - ζ > 0 for BII). The BII conformer occurred most frequently in terminal nucleotides (15 65%), followed by CG, CA, and TG steps (4 20 %) and, finally, adenines within A-tracts (3 - 5%) (Figure S3). The BII conformer was nearly absent in the backbone of internal A-tract thymines. This trend in BII populations resembles closely the trend in NMR C1' S^2_{rel} values and could be used to explain the gradation in sugar mobility across different dinucleotide steps and A-tract motifs in DNA duplexes. Together, analysis of the MD DNA simulations suggested that deoxyribose order parameters obtained by NMR relaxation could be influenced by both backbone BI $\Leftrightarrow$ BII and sugar pucker transitions, which is consistent with a previously established correlation between ensemble BII, S populations and C1' order parameters obtained solely by NMR [23]. However, we did not observe direct coupling between the BI $\Leftrightarrow$ BII and sugar S/N re-puckering fluctuations, suggesting that these two motions could be semi-independent of each other.

Probing Nanosecond Motions by Slowing Down Overall Tumbling

Apart from uncovering an overall helical rigidity in A-tracts of four to six consecutive adenines, we found large-amplitude fluctuations of pyrimidine sugar C1'-H1' bond vectors in two types of YR dinucleotide steps CA, TG, and CG steps. Based on the temperature dependence of resonance profiles, it appeared that these motions occurred on relatively slower timescales, possibly within the nanosecond window. To probe whether the elevated flexibility at YR steps represent ns motions, we devised a strategy to selectively slow down global molecular diffusion and reduce coupling between internal and overall dynamics that would resolve such motions. Our goal was to achieve these conditions without the use of multiple isotopically labeled samples that are required by a domain elongation approach [66]. We employed a combination of minimal elongation of unlabeled DNA samples and glycerol addition, which increases the solvent (water) viscosity and retards the overall rotational diffusion in a predictable manner. Specifically, we collected resonance intensities for unlabeled DNA constructs of the same size (12-mer) or elongated by one C-G base pair on each end (14-mer) in the absence and presence of 20% (v/v) glycerol.

First, as a benchmark for this method, we used a 27-nt HIV-1 TAR construct containing a mutant UUCG tetraloop (mTAR), whose ns dynamics have been extensively characterized by ^{15}N [66] and ^{13}C [45] spin relaxation using a helical elongation technique. Upon addition of 25% (v/v) glycerol that increases the rotational correlation time of mTAR from 6 ns to 11 ns, we observed line narrowing for several nucleotides that were among those previously shown to exhibit ns internal dynamics (data not shown) [66].

To probe for ns motions in DNA sugars, implied by the reduced S^2_{rel} parameters, resonance intensities were first measured for C1'H1' of unlabeled A_6-DNA, A_4-DNA, a4nd A_2-DNA in the presence 20% glycerol that is expected to increase the duplex rotational correlation time (τ_m) by 1.6 times to 7.2 ns at 26 degrees C (Figure 4). The intensity profiles showed

a selective increase in peak intensity at cytosines that were part of CA and CG steps as well as terminal residues (Figure 4A). This suggested that the increased backbone dynamics could involve slower, ns motions that are absent or suppressed in other sequence contexts. The effect was even more pronounced when the same experiment was repeated with a 14-mer A_6-DNA, where we obtained up to 1.5 times larger fractional increase in intensities as compared to 12-mer A_6-DNA (Figure 4B). There, the longer DNA with respectively slower diffusion (τ_m about 9 ns in 20% glycerol) allowed us to probe even deeper into the ns window. Moreover, a noticeable increase in C1'-H1' intensity was observed for the AT step at the 3' A6-tract junction that hinted towards ns dynamics at that site as well. The nearly perfect spectral overlay with and without the retarding agent excluded the possibility that changes in the duplex dynamics are a result of specific interactions with glycerol or major structural changes in DNA (Figure 4C).

A.5 Discussion

In this study, we examine the conformation and dynamics of DNA sequences that contain variable length A-tracts using experimental solution NMR carbon relaxation in conjunction with computational simulations. Our data indicate variations in DNA flexibility that are dependent on A-tract length as well as local dinucleotide environment and support the presence of sequence-specific DNA dynamics. Moreover, such differences in the internal ps-ns dynamics are found to reside primarily at the DNA sugar backbone, which is easily accessible to DNA-targeting agents and can be utilized by proteins and small molecules for indirect readout of specific DNA sequences, nucleotide modifications and damaged sites.

First, the chemical shift analysis confirmed the unusual structure adopted by longer A-tracts and provided evidence for A-tract length-dependent conformational changes near the 5' and 3' junctions. These changes correlated with the increase in internal backbone dynamics for

residues near the A-tract junctions as the A-tract was elongated. This implies that heterogeneity in DNA dynamics of different sequences may correlate with sequence-specific DNA structure. Analysis of the resonance intensities and ^{13}C spin relaxation profiles indicated that the sugar moiety, but not the base, of residues found at A-tract junctions progressively gained flexibility with longer A-tracts, while base and sugar sites of core A-tract residues, especially thymines, remained rigid. Order parameters obtained from MD simulations of the DNA dodecamers yielded excellent agreement with trends of internal mobility for core and certain junction A-tract residues. Some discrepancies values could be rationalized by inadequate sampling in the MD runs or inaccurate duplex structures. Thus, it seems that as the A-tract becomes longer and stiffer (up to a certain point) by stacking AA steps that collectively favor a distinct B-DNA conformation, residues at the A-tract ends become increasingly flexible and perhaps subject to helical deformations to retain a favorable base-stacking arrangement with the A-tract. We cannot rule out large-scale helical bending motions as the source for the enhanced backbone dynamics at A-tract junctions, even though such motions could not be previously detected using a DNA domain-elongation approach [67].

The structural rigidity of A-tracts is not a new concept. As discussed before, AA steps comprising longer A-tracts are ranked as the most rigid dinucleotide sequences. Moreover, increased base pair stability [70] and helical stiffness [19] has been observed for poly(dA)-poly(dT) sequences of at least three consecutive AT base pairs. However, here we report the first quantitative NMR relaxation study of ps to ns dynamics confirming that A-tract residues do not exhibit unusual large-amplitude base or sugar motions, except for residues near the junctions. The formation of these inflexible DNA blocks of AA steps can be traced to their strong conformational preference to adopt a large propeller twist and limited slide mobility that could have a severe stereochemical locking effect and, in principle, introduce a mechanical strain in longer A-tracts. Yet, its effect could be potentially offset by stabilizing interactions improved $\pi - \pi$ base stacking, bifurcated hydrogen bonds, and formation of an ordered hydration spine in the narrow minor groove, coupled with helical bending that

are proposed to be specific features of A-tract sequences based on a number of biochemical and biophysical studies [14], but that still remain controversial. The distinct structure and higher rigidity of A-tract motifs, granted by these interactions, could stabilize helical bends or a narrow minor groove, which frequently serves as an accessory indirect recognition site in transcription factor binding to its cognate DNA [12].

The NMR data also revealed extensive sugar, but not base, dynamics at C1' spins of cytosine and thymine nucleotides located at CA, TG, and CG sequences that appear to be a general feature of pyrimidine-purine dinucleotide steps. The intensity data at different temperatures and glycerol levels strongly suggest that these dynamics, particularly in cytosine nucleotides, occur at ns timescales and are suppressed in purine nucleotides. The ^{13}C relaxation data further showed that these dynamics are somewhat modulated by the A-tract size when the step was positioned at the 5' A-tract junction and by variable nearest-neighbor nucleotides (i.e. TCG vs. GCG). Moreover, the trend of reduced order parameters at these sites, but not the extent of these amplitudes, was captured by MD simulations; this likely results from undersampling (i.e., from broken ergodicity) and/or structural discrepancies between NMR and MD ensembles. These differences can be addressed in the future by running longer MD simulations and possibly by using known NMR structures as initial DNA coordinates.

Previously, several solution NMR investigations have reported increased backbone disorder in cytosine and, less so, in thymine sugars of YR (CG, CA, TG, and TA) and YY (CT or TC) context in B-DNA [22, 26, 28, 43]. While the anomalously high mobility of cytosine sugars was linked to cytosine-specific backbone motions and the sequence dependence was under-stressed, the higher mobility of thymine sugars at TA and TG steps relative to more rigid TT and AT steps has escaped attention likely due to the paucity of sequence-specific probes. Also, there is previous evidence for ns motions at CG dinucleotides [41], which we extend here to CA dinucleotide. There, the observation of increased flexibility at the HhaI methyltransferase target dodecamer comprised of two CG steps by both solid and solution

state NMR, which are sensitive to different timescales, could be reconciled with a specific motional model that involves slower-than-diffusion cytosine sugar fluctuations [41].

Indeed, the events that underlie these molecular transitions are difficult to probe solely by NMR. The development of motional models for DNA flexibility can benefit tremendously from state-of-the-art computational simulations, as has been demonstrated for canonical [28, 33], non-canonical [71] and damaged [72] DNA. In one particular NMR study informed by MD, Duchardt et al. proposed a motional model for the rapid picosecond mobility observed at cytosine sugar moieties that involves sugar re-puckering (S/N) transitions [28]. This model, supported by the higher preference of cytosines for the N conformer determined by experiment and ab initio calculations [73, 74], could well be physically plausible. Here, analysis of the sugar pucker distributions and re-puckering transition rates in MD simulations with DNA sequence also uncovered increased populations and longer lifetimes for non-canonical N (C3'-endo) conformers in cytosines and thymines located in CA, TG, and CG steps as compared to other sequences. The increased population of N puckers in A-tract thymine and especially cytosine sugars is further supported by the more downfield shifted C1' chemical shifts. At the same time, we found that the backbone for pyrimidines in CG, and less so CA/TG, steps was particularly enriched with the minor BII conformer. These findings are in agreement with prior solution NMR studies based on 31P chemical shifts, $^{3}J_{H3'-P}$ couplings, and inter-proton distances as well as with surveys of DNA crystal structures and MD simulations showing that the BI $\Leftrightarrow$ BII balance is sequence-specific, with the rare BII conformation having higher occupancy in CA, TG, and CG dinucleotide steps [75, 76, 77]. These two sugar-backbone motions, which do not appear to be directly coupled to each other from the MD trajectories, could provide a plausible explanations for the markedly lower C1' order parameters at YR steps.

The increased populations of C3'-endo states within core A-tract thymines over adenines in MD failed to explain the somewhat higher thymine C1' order parameters obtained by exper-

iment. However, we observed a direct link between the lower mobility of A-tract thymines and negligible fractions of the minor BII conformer, that were at least 10-fold higher in the opposite adenines or in TG steps. Therefore, we hypothesize that excursions to the minor BII conformer could in fact contribute to the C1' order parameters for internal nucleotides. In conjunction with this hypothesis is a study that links stabilization of the BI over BII backbone conformer of cytosines upon C5-methylation observed by MD [78] with dampened sugar and phosphate backbone dynamics observed by NMR [31, 40, 41]. These findings give credence to the emerging idea that MD simulations are capable of providing physically relevant models for the intrinsic dynamics of nucleic acids and their sequence dependence. In a biological context, unusual sugar-backbone dynamics can ultimately facilitate recognition of specific DNA sequences by their protein or small molecule binders. NMR/MD studies by the groups of Schleucher [28], Drobny and Varani [26, 40, 41] have made significant progress in understanding the sugar-backbone dynamics of AT-rich EcoRI endonuclease and CG-rich HhaI methyltransferase target site as well as the impact of methylation at CG steps on backbone flexibility, which may play a role in methylation-dependent protein recognition. Flexible CA/TG steps are also targeted by many biological factors, such the ubiquitous and gene-regulating CAP [12, 79] and p53 [80] proteins that are known to induce large deformations or trap non-canonical base-pairing conformations. Similar recognition strategies that take advantage of sequence-specific duplex flexibility are also utilized by DNA-binding drugs [5]. DNA is emerging as a prominent drug target and effective tools for analysis of DNA-drug recognition can facilitate the development of therapeutics. Thus, the prospect of engineering gene regulation by protein- or drug-DNA interactions places a tremendous importance on how well we understand and can manipulate sequence-dependent DNA dynamics.

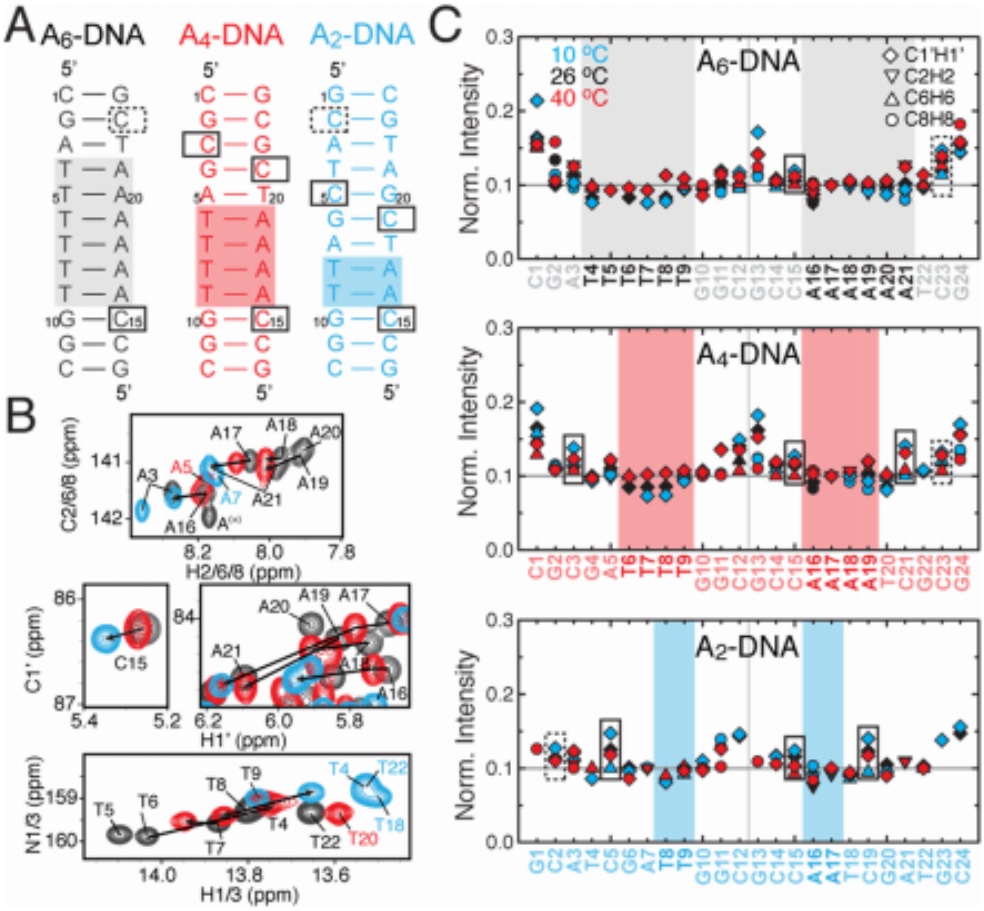

Figure A.1: (A) DNA constructs for A_6-DNA (black), A_4-DNA (red) and A_2-DNA (blue). The A-tract position is highlighted and flexible cytosines in CA/TG and CG steps are boxed, corresponding to plots in (C). (B) NMR overlays of 1H,^{13}C-HSQC and 1H,^{15}N-HSQC spectra, color-coded for the three DNA sequences in (A). (C) NMR resonance intensity profiles for base (C2H2, C6H6, C8H8) and deoxyribose (C1'H1') DNA sites obtained from 2D 1H,^{13}C-HSQC spectra at three temperatures (see inset). Boxed residues correspond to cytosine sugar sites that show enhanced intensities and also an unusual increase in intensity with lower temperatures (near terminal sites are dashed).

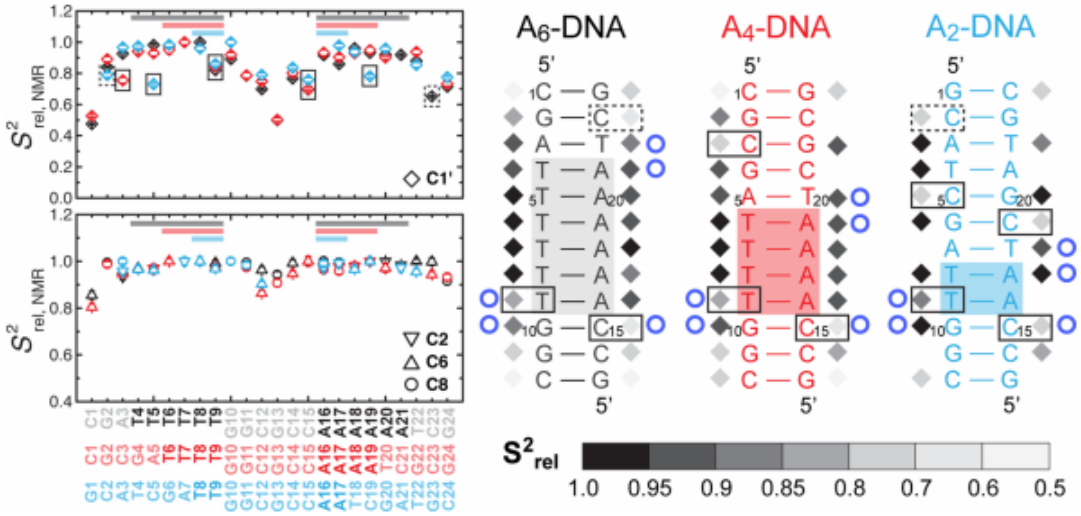

Figure A.2: . Relative order parameters S^2_{rel} (right) obtained from ^{13}C NMR spin relaxation data for base (C2H2, C6H6, C8H8) and deoxyribose (C1'H1') sites in A_6-DNA, A_4-DNA, and A_2-DNA (26 deg C) and DNA constructs (left) showing the variation in sugar backbone S^2_{rel} (diamond). Pyrimidine residues with reduced sugar C1' S^2_{rel} values are boxed in plots and DNA sequences (near terminal sites are dashed), while A-tract junction residues that are modulated by A-tract length are marked with a blue circle.

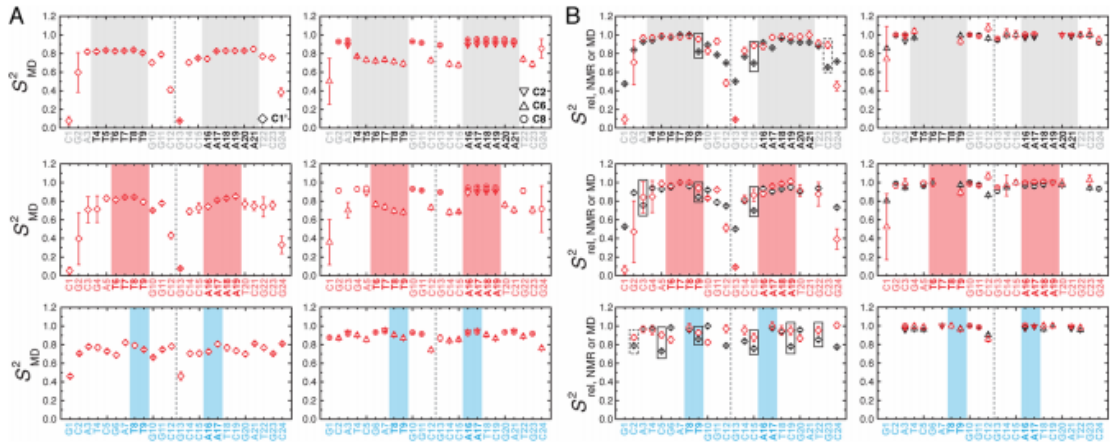

Figure A.3: (A) S^2 order parameters obtained by MD simulations for base (C2H2, C6H6, C8H8) and deoxyribose (C1'H1') sites in A_6-DNA, A_4-DNA, and A_2-DNA. (B) Comparison between relative order parameter S^2_{rel} obtained by NMR ^{13}C spins relaxation and MD simulations. Pyrimidine residues with reduced sugar C1' S^2_{rel} values are boxed in plots (near terminal sites are dashed).

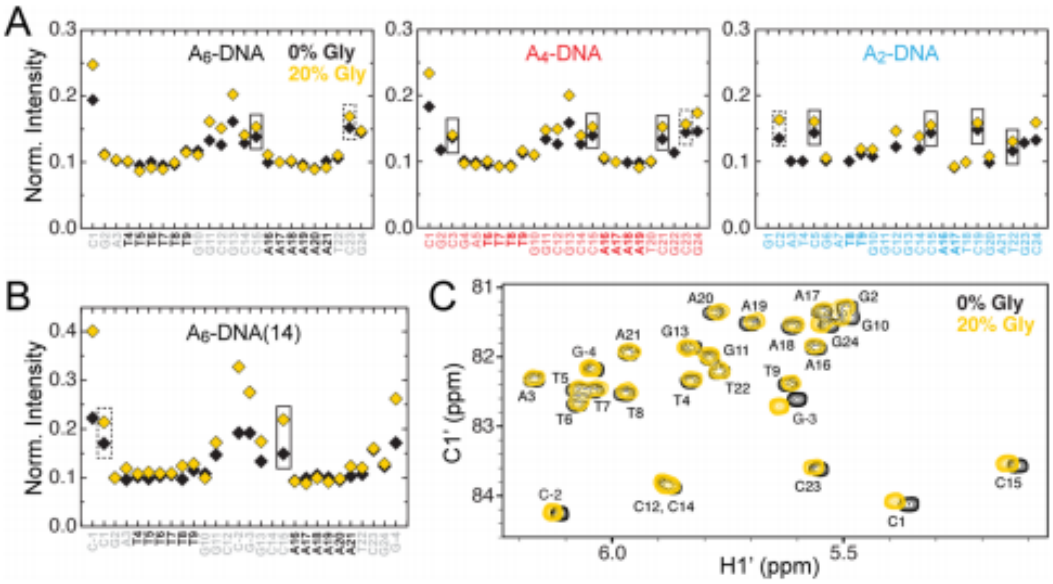

Figure A.4: (A) NMR resonance intensity profiles for base (C2H2, C6H6, C8H8) and deoxyribose (C1'H1') sites in A_6-DNA, A_4-DNA, and A_2-DNA obtained in the absence (black) and presence (green) of 20 % glycerol. Pyrimidine residues with reduced sugar C1' S^2_{rel} values that show increase in intensity with glycerol addition are boxed in plots (near terminal sites are dashed). (C) Corresponding intensity profiles for A_6-DNA [14]. (B) Overlay of 2D 1H,^{13}C-HSQC spectra of A_6-DNA [14] in the absence and presence of 20 % glycerol

Bibliography

[1] Yanagi K. Analysis of local helix geometry in three b-dna decamers and eight dodecamers. *J. Mol. Biol.*, 217:201, 1991.

[2] Dickerson R. E. Dna bending: The prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.*, 26:1906, 1998.

[3] Olson W. K. Dna sequence-dependent deformability deduced from protein-dna crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 95:11163, 1998.

[4] Svozil D. Dna conformations and their sequence preferences. *Nucleic Acids Res.*, 36:3690, 2008.

[5] Arauzo-Bravo M. J. Indirect readout in drug-dna recognition: Role of sequence-dependent dna conformation. *Nucleic Acids Res.*, 36:376, 2008.

[6] Rohs R. The role of dna shape in protein-dna recognition. *Nature*, 461:1248, 2009.

[7] Segal E. A genomic code for nucleosome positioning. *Nature*, 442:772, 2006.

[8] El Hassan M. A. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in dna. *J. Mol. Biol.*, 259:95, 1996.

[9] Packer M. J. Sequence-dependent dna structure: Tetranucleotide conformational maps. *J. Mol. Biol.*, 295:85, 2000.

[10] Gardiner E. J. Sequence-dependent dna structure: A database of octamer structural parameters. *J. Mol. Biol.*, 332:1025, 2003.

[11] Suzuki M. Stereochemical basis of dna bending by transcription factors. *Nucleic Acids Res.*, 23:2083, 1995.

[12] Chen S. Indirect readout of dna sequence at the primary-kink site in the cap-dna complex: Dna binding specificity based on energetics of dna kinking. *J. Mol. Biol.*, 314:63, 2001.

[13] Kim Y. Crystal structure of a yeast tbp/tata-box complex. *Nature*, 365:512, 1993.

[14] Haran T. E. The unique structure of a-tracts and intrinsic dna bending. *Q. Rev. Biophys.*, 42:41, 2009.

[15] MacDonald D. Solution structure of an a-tract dna bend. *J. Mol. Biol.*, 306:1081, 2001.

[16] Stefl R. Dna a-tract bending in three dimensions: Solving the da4t4 vs. dt4a4 conundrum. *Proc. Natl. Acad. Sci. U.S.A.*, 101:1177, 2004.

[17] Cohanim A. B. The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. *Nucleic Acids Res.*, 37:6466, 2009.

[18] Segal E. Poly(da:dt) tracts: Major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, 19:65, 2009.

[19] Hogan M. Dependence of dna helix flexibility on base composition. *Nature*, 304:752, 1983.

[20] Hagerman P. J. Sequence dependence of the curvature of dna: A test of the phasing hypothesis. *Biochemistry*, 24:7033, 1985.

[21] Nagaich A. K. Ca/tg sequence at the 5 end of oligo(a)-tracts strongly modulates dna curvature. *J. Biol. Chem.*, 269:7824, 1994.

[22] Paquet F. Selectively 13c-enriched dna: Evidence from 13c1 relaxation rate measurements of an internal dynamics sequence effect in the lac operator. *J. Biomol. NMR*, 8:252, 1996.

[23] Isaacs R. J. Nmr evidence for mechanical coupling of phosphate b(i)-b(ii) transitions with deoxyribose conformational exchange in dna. *J. Mol. Biol.*, 311:149, 2001.

[24] Okonogi T. M. Sequence-dependent dynamics of duplex dna: The applicability of a dinucleotide model. *Biophys. J.*, 83:3446, 2002.

[25] Kojima C. Dna duplex dynamics: Nmr relaxation studies of a decamer with uniformly 13c-labeled purine nucleotides. *J. Magn. Reson.*, 135:310, 1998.

[26] Shajani Z. Nmr studies of dynamics in rna and dna by 13c relaxation. *Biopolymers*, 86:348, 2007.

[27] Shajani Z. 13c relaxation studies of the dna target sequence for hhai methyltransferase reveal unique motional properties. *Biochemistry*, 47:7617, 2008.

[28] Duchardt E. Cytosine ribose flexibility in dna: A combined nmr 13c spin relaxation and molecular dynamics simulation study. *Nucleic Acids Res.*, 36:4211, 2008.

[29] Perez A. Dynamics of b-dna on the microsecond time scale. *J. Am. Chem. Soc.*, 129:14739, 2007.

[30] Mura C. Molecular dynamics of a b dna element: Base flipping via cross-strand intercalative stacking in a microsecond-scale simulation. *Nucleic Acids Res.*, 36:4941, 2008.

[31] Tian Y. 31p nmr investigation of backbone dynamics in dna binding sites. *J. Phys. Chem. B*, 113:2596, 2009.

[32] Bothe J. R. Sequence-specific b-dna flexibility modulates z-dna formation. *J. Am. Chem. Soc.*, 133:2016, 2011.

[33] Nikolova E. N. Transient hoogsteen base pairs in canonical duplex dna. *Nature*, 470:498, 2011.

[34] Perez A. Impact of methylation on the physical properties of dna. *Biophys. J.*, 102:2140, 2012.

[35] Stivers J. T. Extrahelical damaged base recognition by dna glycosylase enzymes. *Chemistry*, 14:786, 2008.

[36] Patikoglou G. A. Tata element recognition by the tata box-binding protein has been conserved throughout evolution. *Genes Dev.*, 13:3217, 1999.

[37] Aishima J. A hoogsteen base pair embedded in undistorted b-dna. *Nucleic Acids Res.*, 30:5244, 2002.

[38] Ughetto G. A comparison of the structure of echinomycin and triostin a complexed to a dna fragment. *Nucleic Acids Res.*, 13:2305, 1985.

[39] Ha S. C. Crystal structure of a junction between b-dna and z-dna reveals two extruded bases. *Nature*, 437:1183, 2005.

[40] Meints G. A. Dynamic impact of methylation at the m.hhai target site: A solid-state deuterium nmr study. *Biochemistry*, 40:12436, 2001.

[41] Echodu D. Furanose dynamics in the hhai methyltransferase target dna studied by solution and solid-state nmr relaxation. *J. Phys. Chem. B*, 112:13934, 2008.

[42] Borer P. N. 13c-nmr relaxation in three dna oligonucleotide duplexes: Model-free analysis of internal and overall motion. *Biochemistry*, 33:2441, 1994.

[43] Spielmann H. P. Dynamics in psoralen-damaged dna by 1h-detected natural abundance 13c nmr spectroscopy. *Biochemistry*, 37:5426, 1998.

[44] Zimmer D. P. Nmr of enzymatically synthesized uniformly 13c15n-labeled dna oligonucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, 92:3091, 1995.

[45] Hansen A. L. Dynamics of large elongated rna by nmr carbon relaxation. *J. Am. Chem. Soc.*, 129:16072, 2007.

[46] Delaglio F. Nmrpipe: A multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR*, 6:277, 1995.

[47] Palmer A. G. Characterization of the dynamics of biomacromolecules using rotating-frame spin relaxation nmr spectroscopy. *Chem. Rev.*, 106:1700, 2006.

[48] Tjandra N. An approach to direct determination of protein dynamics from n-15 nmr relaxation at multiple fields, independent of variable n-15 chemical shift anisotropy and chemical exchange contributions. *J. Am. Chem. Soc.*, 121:8577, 1999.

[49] Dethoff E. A. Characterizing complex dynamics in the transactivation response element apical loop and motional correlations with the bulge by nmr, molecular dynamics, and mutagenesis. *Biophys. J.*, 95:3906, 2008.

[50] Garcia de la Torre J. Hydronmr: Prediction of nmr relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *J. Magn. Reson.*, 147:138, 2000.

[51] Lu X. J. 3dna: A versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, 3:1213, 2008.

[52] DA ea Case, TA Darden, TE Cheatham Iii, CL Simmerling, J Wang, RE Duke, R Luo, RC Walker, W Zhang, KM Merz, et al. Amber 11. *University of California, San Francisco*, 142, 2010.

[53] Humphrey W. Vmd: Visual molecular dynamics. *J. Mol. Graphics*, 14:27, 1996.

[54] Phillips J. C. Scalable molecular dynamics with nam. *J. Comput. Chem.*, 26:1781, 2005.

[55] MacKerell A. D. Development and current status of the charmm force field for nucleic acids. *Biopolymers*, 56:257, 2000.

[56] Barsky D. New insights into the structure of abasic dna from molecular dynamics simulations. *Nucleic Acids Res.*, 28:2613, 2000.

[57] Szabo A. Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules. 1. theory and range of validity. *J. Am. Chem. Soc.*, 104:4546, 1982.

[58] Musselman C. Referencing strategy for the direct comparison of nuclear magnetic resonance and molecular dynamics motional parameters in rna. *J. Phys. Chem. B*, 114:929, 2010.

[59] Clore G. M. Deviation from the simple 2-parameter model-free approach to the interpretation of n-15 nuclear magnetic-relaxation of proteins. *J. Am. Chem. Soc.*, 112:4989, 1990.

[60] Henry E. R. Influence of vibrational motion on solid-state line-shapes and nmr relaxation. *J. Chem. Phys.*, 82:4753, 1985.

[61] Brooks B. R. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187, 1983.

[62] Mike P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 1989.

[63] Marini J. C. Bent helical structure in kinetoplast dna. *Proc. Natl. Acad. Sci. U.S.A.*, 79:7664, 1982.

[64] Kintanar A. Two-dimensional nmr investigation of a bent dna fragment: Assignment of the proton resonances and preliminary structure analysis. *Nucleic Acids Res.*, 15:5845, 1987.

[65] Katahira M. One- and two-dimensional nmr studies on the conformation of dna containing the oligo(da)oligo(dt) tract. *Nucleic Acids Res.*, 16:8619, 1988.

[66] Zhang Q. Resolving the motional modes that code for rna adaptation. *Science*, 311:653, 2006.

[67] Nikolova E. N. Preparation, resonance assignment, and preliminary dynamics characterization of residue specific 13c/15n-labeled elongated dna for the study of sequence-directed dynamics by nmr. *J. Biomol. NMR*, 45:9, 2009.

[68] Duchardt E. Residue specific ribose and nucleobase dynamics of the cuucgg rna tetraloop motif by mnmr 13c relaxation. *J. Biomol. NMR*, 32:295, 2005.

[69] Shajani Z. 13c nmr relaxation studies of rna base and ribose nuclei reveal a complex pattern of motions in the rna binding site for human u1a protein. *J. Mol. Biol.*, 349:699, 2005.

[70] Leroy J. L. Evidence from base-pair kinetics for two types of adenine tract structures in solution: Their relation to dna curvature. *Biochemistry*, 27:8894, 1988.

[71] Isaacs R. J. Insight into gt mismatch recognition using molecular dynamics with time-averaged restraints derived from nmr spectroscopy. *J. Am. Chem. Soc.*, 126:583, 2004.

[72] Chen J. Dna oligonucleotides with a, t, g or c opposite an abasic site: Structure and dynamics. *Nucleic Acids Res.*, 36:253, 2008.

[73] LaPlante S. R. 13c-nmr of the deoxyribose sugars in four dna oligonucleotide duplexes: Assignment and structural features. *Biochemistry*, 33:2430, 1994.

[74] Foloppe N. Intrinsic conformational properties of deoxyribonucleosides: Implicated role for cytosine in the equilibrium among the a, b, and z forms of dna. *Biophys. J.*, 76:3206, 1999.

[75] Lefebvre A. Solution structure of the cpg containing d(cttcgaag)2 oligonucleotide: Nmr data and energy calculations are compatible with a bi/bii equilibrium at cpg. *Biochemistry*, 35:12560, 1996.

[76] Madhumalar A. Sequence preference for bi/bii conformations in dna: Md and crystal structure data analysis. *J. Biomol. Struct. Dyn.*, 23:13, 2005.

[77] Heddi B. Quantification of dna bi/bii backbone states in solution. implications for dna overall structure and recognition. *J. Am. Chem. Soc.*, 128:9170, 2006.

[78] Rauch C. C5-methylation of cytosine in b-dna thermodynamically and kinetically stabilizes bi. *J. Am. Chem. Soc.*, 125:14990, 2003.

[79] Parkinson G. Structure of the cap-dna complex at 2.5 angstroms resolution: A complete picture of the protein-dna interface. *J. Mol. Biol.*, 260:395, 1996.

[80] Kitayner M. Diversity in dna recognition by p53 revealed by crystal structures with hoogsteen base pairs. *Nat. Struct. Mol. Biol.*, 17:423, 2010.