Coarse-Grained Molecular Dynamics Simulations of DNA Representing Bases as Ellipsoids

by

Herbert Alexander Morriss-Andrews

B.Sc.H., Queen's University, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Physics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2009

Abstract

We conducted a study of various physical and geometric properties of single and double stranded DNA by developing a coarse-grained computational model. To represent the steric shapes of the bases more accurately, we modelled their interactions using the ellipsoidal RE^2 potential [18]. The bond, angle and dihedral potentials describing the DNA backbone are parametrized using statistics we gathered from atomistic simulations, and the RE^2 interaction parameters were likewise obtained from all-atom data.

We find reasonable agreement of the dependence of the persistence length on ionic concentration with polymer theory [44]. We also observed the collapse of single stranded DNA by checking the radius of gyration as we lowered the temperature. Our simulations show a very strong dependence of helical twist and stacking interactions on the strength of the RE^2 potential, although increasing the twist frustrates all other potentials. The dependence of twist and stacking on ionic concentration and temperature was also analyzed.

Our DNA also exhibits the correct right-handed chirality, with the chiral symmetry broken by the dihedral potentials. The model also gives clear major and minor grooves. Our methodical approach to the parametrization and analysis provides insight into the effect of the parameters in a coarse-grained model on the properties observed in the simulated DNA.

Table of Contents

List of Tables

List of Figures

List of Programs

Acknowledgements

I would like to thank my supervisors Steve Plotkin and Joerg Rottler for their helpful advice and direction. Dr. Plotkin has given me great insight into the field of biophysics, and Dr. Rottler has given me similar insight into computational physics and has, in addition, provided me with the computational resources needed to carry out these simulations. I thank Erik Abrahamsson, who developed the molecular visualization tool BioVEC, which was essential in providing me with a method of visualizing my simulations. I also thank Christopher Yearwood, who laid some of the ground work for this project. This work was supported in part by NSERC.

Chapter 1

Introduction

1.1 Properties of DNA

The goal of computer simulation of DNA is to study its properties in an "experiment" over which we have absolute control of the environment. We will refer to a number of these properties in what follows. A double stranded

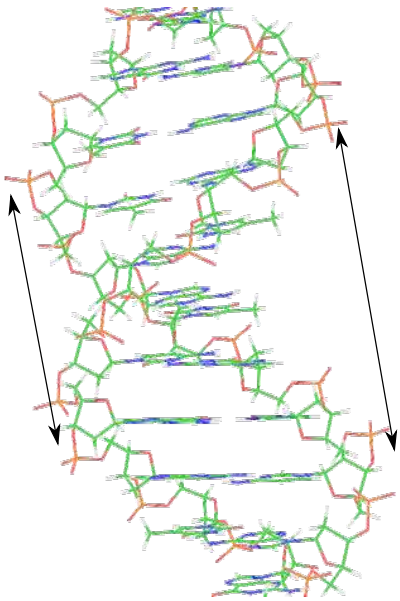

Figure 1.1: Depicts an all-atom visualization of dsDNA. The long and short arrow segments mark the major and minor grooves respectively. Rendered in PyMOL [12].

DNA molecule forms a double helix, with two grooves, shown in Figure 1.1. The groove is the distance from a point along the backbone of one strand to the other strand half a helical rotation above or below it. DNA is observed to form two sizes of grooves: the major and minor. The distance "up" to the

next strand will be different than the distance "down" to the next strand. The helix has a structural periodicity called *pitch*: this is the number of base pairs for the helix to make one cycle. DNA naturally has a pitch of 10 base pairs (or 10bp) [3]. This means that each base pair contributes *twist* of 36°, so that after 10 base pairs the total rotation of the helix is 360°. The *persistence length*, or the distances over which correlations in the tangent vector of the strand decay, is a measure of the stiffness of the DNA. This is discussed in greater detail in Section 4.2. *Stacking* refers to the degree to which adjacent bases on the same strand are stacked to each other (see Section 4.5). DNA *dehybridizes* when the two strands come apart. A DNA *hairpin* forms when a single strand of DNA hydrogen bonds to itself because the bases on one half are complementary to the bases on the other half, forming a hairpin-like structure.

These properties characterize the physical state and dynamics of a DNA molecule, and are amenable to both computational and experimental analysis. Thus they have been used repeatedly throughout the development of computer simulations of biomolecules as a means to confirm the accuracy of the proposed models, and we will be investigating many of these properties in this study.

1.2 Literature Review

1.2.1 Molecular Dynamics

The molecular dynamics simulation technique, or MD, was developed by Alder and Wainwright in the late 1950s as a method for simulating systems of atoms as an alternative to the Monte Carlo algorithm [2]. They were able to study a nonequilibrium system of up to about 100 particles and analyze collisions between the particles. A distinct advantage of the molecular dynamics method is that it incorporates a real notion of time, and therefore realistically represents the dynamics of the system.

The basic idea of molecular dynamics is to assume that the particles being simulated obey Newton's laws of motion for some potential energy function. Time is then discretized into timesteps of finite size Δt and the equations of motion are integrated over each timestep using a finite difference method. Details on molecular dynamics are well-discussed in textbooks on the subject, e.g.: [22, 48]. This thesis will focus mainly on the development of molecular dynamics simulations of DNA and coarse-graining.

Simulating large systems of atoms for long timescales is an extremely computationally intensive problem [45]. There have been many attempts

to improve the speed of the algorithms, most of which necessarily come at the cost of decreased resolution of the model. One such improvement is the SHAKE algorithm [49], an iterative method used to freeze out degrees of freedom which are not deemed essential in the analysis. This is most often applied to protons in a molecule, which have a bond length of only about 1Å and thus oscillate at a very high frequency tightly about the potential minimum. High frequency motions require that the timestep be smaller so that the integration routine remains stable; applying the SHAKE algorithm to these motions allows the timestep to be increased.

1.2.2 Atomistic Simulations of DNA

All-atom simulations are the most highly resolved classical approximations of molecules that are in use today. Common examples of all-atom potentials are CHARMM (Chemistry at HARvard Molecular Mechanics) [9], AMBER (Assisted Model Building with Energy Refinement) [10] and GROMACS (GROningen MAchine for Chemical Simulations) [27]. Each atom is represented as a single sphere, typically with the Lennard Jones potential [31] discussed in Section 1.2.4.

An early example of atomistic simulation of DNA appeared in the 1980s [36]. This model was able to analyze pitch, flexibility, kinking along the strand and oscillations of the DNA about the equilibrium structure. At that time, computational limitations restricted the model to simulations of DNA of a length of 24 base pairs (abbreviated 24bp) and times of up to 90ps. This is nowhere near long enough spatially or temporally to study phenomena such as persistence length. The model showed a number of effects, e.g.: the bases tended to kink into the minor groove, the DNA absorbed thermal motions by bending rather than base pair opening, the hydrogen bond stability was due more to the helix structure itself than the nature of the hydrogen bond and the pitch of the DNA was found to be 11-11.15bp, instead of the experimental value of 9.6-10.6bp.

A paper in the late 1990s [19] compared how the CHARMM and AMBER potentials represented the structure of DNA, with respect to the twist: the helical rotation about one base pair along the DNA. The authors found that the average helical twist in CHARMM was 31.85° per base pair, and 30.45° in AMBER. The correct value of the twist from the crystal structure is 36° so that the helical pitch is 10bp [3]. They simulated a strand of length 10 base pairs for 10 ns.

Another group simulated a 17bp sequence for 20 ns [35]. They considered properties such as torsional rigidity and pitch. They found a pitch of 11-11.5,

admitting that AMBER underestimates the twist of DNA, agreeing with [19]. They also saw a persistence length of 50 nm in high salt concentration, although the length of a 17bp segment of DNA is only about 56 nm. Their sequence is barely long enough to get a persistence length accurately. Other AMBER studies [39] found a helical twist of 22-34°.

A review article on molecular dynamics simulations of nucleic acids [29] discussed the successes and issues with simulating DNA using all-atom force fields. Although the article is eight years old, much of it still applies today. Improvement to force fields creeps forward slowly at best, and the main problem of computational complexity has changed hardly at all. While computers have become faster, they have not increased their computational power by the many orders of magnitude necessary to access significantly longer time or length scales. The paper lists the following as successes of atomistic modeling of DNA: ion interactions in DNA grooves, DNA bending, the effect of ions on DNA and DNA structures, such as hairpins or base pair mismatches. The biggest concern was conformational sampling, i.e. the ability to simulate for long enough durations for good statistical significance of thermodynamic results. At the time the paper was written, the state-of-the-art was to simulate strands of length about 100 base pairs for up to about 25 ns. This is only a long enough time to sample around a local minimum, and not the entire configuration space. The other main concern, which is still very relevant, is that the parametrization of an all-atom force field has many degrees of freedom. Thus it is a very complex problem to find the optimal energy functional to model the system. Different possible energy functionals e.g., CHARMM, AMBER and GROMACS, have different strengths and weaknesses. We are a long way from determining a "perfect" energy functional that is always guaranteed to yield accurate simulations. Therefore, many researchers develop coarse-grained models to allow for better conformational sampling and which has fewer parameters to be tweaked when studying specific behaviours in a particular system.

1.2.3 United Atom DNA Models

A common form of coarse-graining, and the kind that we have used in this work, is the use of united atom force fields. A contrast of a coarse-grained model (the one used in this thesis) and an all-atom model is shown in Figure 1.2. Generally, researchers developing such a model design a set of potential energy functions between the coarse-grained particles to analyze certain properties of the system they are studying. One bead-spring model of DNA [13] developed in 2000 was able to model melting and renaturation

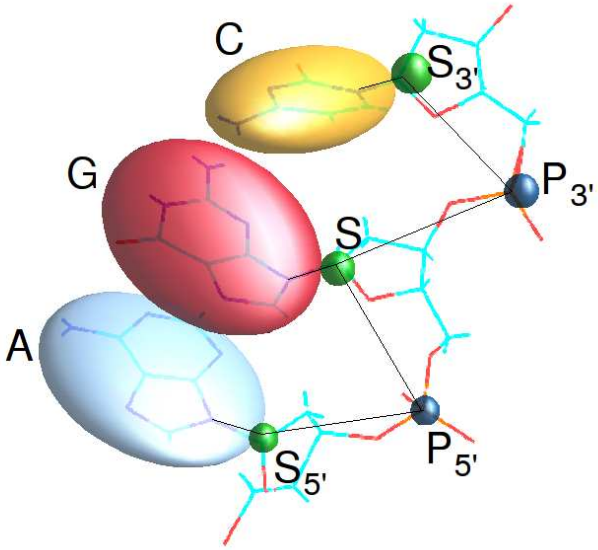

Figure 1.2: Our coarse-graining scheme, overlaid onto an all-atom visualization. The bases are identified by their first letter and the sugars and phosphates are identified by "S" and "P", with a number indicating whether they are on the $3'$ or $5'$ side of the base (in this case the guanine).

of DNA that was consistent with experimental results. These coarse-grained particles consisted of five types: a backbone particle and four base particles. Another similar model in 2007 [32], in addition to modeling rehybridization and melting, analyzed the ionic dependence of the persistence length of the DNA and bubble dynamics, where a segment of DNA dehybridizes like a bubble within the helix and may propagate over time. This technique improved the resolution of the previous model [13] by adding an extra backbone type: the sugar and the phosphate were treated each as a single particle. The backbone then alternated between these two types of particles. A variation of this sugar-phosphate-base spherical particle model was more recently used [50] to obtain realistic data for the melting temperature, persistence length in ionic solution, heat capacity profiles and DNA renaturation.

A fundamentally different model was developed in 2004 [53] that was designed to model the twisting of a ladder of dsDNA spontaneously into a helix. It used the alternating sugar-phosphate backbone as in [32, 50], but had a much higher resolution for the bases. A single base pair was

represented by a group of 14 beads attached by springs. These bases differ from other models in that each base pair is a single cluster of bead-springs so that the bases cannot separate. The researchers focused on the stacking interaction, which was tweaked so that the potential preferred stacked bases to be separated by the experimental value of 3.3Å. They claimed that the only way to satisfy the phosphate-phosphate separation of 6Å and have the bases stacked 3.3Å apart was to spontaneously form a helix. Their model behaved as it was designed to do; however, since their model was chiral symmetric, they sometimes found kinks in the DNA where the chirality reversed direction along the strand.

1.2.4 Pair Potentials: From Spheres to Ellipsoids

Usually in united atom simulations, the coarse-grained particles have spherically symmetric potentials. The most common pair interaction between particles in molecular simulations is the 6-12 Lennard-Jones potential proposed in 1924 as the classical potential which best fit the quantum interactions between neutral atoms [31]. A discussion of the precise form of the Lennard-Jones potential is found in Section 2.2.

At short separations there is a sharp impenetrable barrier representing the hard core repulsion between atoms. At intermediate distances there is an attractive well representing quantum electronic attraction between the atoms, and at long distances the potential fades off to zero. The potential acting between all pairs of atoms comes at a computational cost of $O(N^2)$ per timestep. The cost can be eased somewhat by applying a cutoff $r_{\max}$ and saving the atoms that lie within the cutoff in a pair list [11].

While single atoms are spherically symmetric, this is not always a good approximation for groups of atoms. DNA bases, for example, are too flat along their stacking direction. An extension of a spherical potential is an ellipsoidal potential. Gay and Berne created a straightforward extension of the Lennard-Jones potential to anisotropic particles [24], although what is commonly referred to as the Gay-Berne potential is an extension of their function to possibly dissimilar ellipsoidal particles [7].

In 2003, Everaers and Ejtehadi [18] proposed a modification of the Gay-Berne potential to soft ellipsoids where the parameters have a well defined microscopic interpretation. This potential, along with the Gay-Berne potential, was incorporated into the LAMMPS simulation package [47], and is the potential that we use to represent the DNA bases. It is known as the RE^2 potential.

A previous model of DNA using ellipsoids for the bases [40] exists. This

model, however, used Monte Carlo moves instead of molecular dynamics and, like the studies of Tepper and Voth [53], does not distinguish between different bases and treats base pairs as a single unit. Each pair of bases is represented by a single ellipsoid that is bonded to both backbones of the double helix. This model prevents analysis of properties such as dehybridization or simulation of single stranded DNA. The potentials, again similar to [53] were set so that base stacking of 3.3Å was the preferred separation. Their model has three free parameters, each designed to adjust a certain property of the DNA: stacking, torsional rigidity, and the ellipsoid energy well depth. The authors found a persistence length for double stranded DNA of 150bp, which translates into a length of about 50nm. They were able to tweak their potential so that the helical period became 10bp.

Representing single bases as ellipsoids gives a better representation of the steric shape of these molecular groups and the anisotropic potential between them. It comes, however, at the cost of simulation time because the form of the potential is much more complicated than a Lennard-Jones potential, and involves a number of changes of reference frame and matrix algebra. As a result, simpler models [13] are able to run approximately 45 times faster than our simulations on a single CPU. However, a coarse grained ellipsoid simulation is still much faster than an all-atom simulation. Even for potentials between one base and another, an atomistic simulation has ~ 15 atoms in one base interacting with ~ 15 in the other, giving a total of ~ 225 interactions. The all-atom simulation is much more computationally expensive.

1.3 Motivation

The need for coarse-grained models of DNA will be present as long as atomistic simulations of large systems over tens of nanoseconds or more is infeasible. As one can see from reading the literature on coarse-grained models of DNA, such as the papers mentioned in Section 1.2, our understanding of the coarse-grained classical interactions of DNA is lacking. No model exists which can describe all the features of DNA that are needed and is efficient enough computationally to handle strands of DNA hundreds of base pairs long. By studying the effects of coarse-grained potentials on various observables of the DNA, we can gain a better understanding of how coarse-graining works, which will lead to improvements in coarse-grained modeling.

Our work is mainly focused on analysis of the coarse-graining procedure itself. We use systematic methods to obtain the functional forms and

parametrization of our potentials. After observing the DNA with the *a priori* parameters, we scale the energetic parameters to analyze what effect each of our potentials has on DNA properties, such as persistence length, twist and base stacking. This kind of analysis emphasizes the need to understand the coarse-graining process more precisely in order to improve methods in DNA modeling.

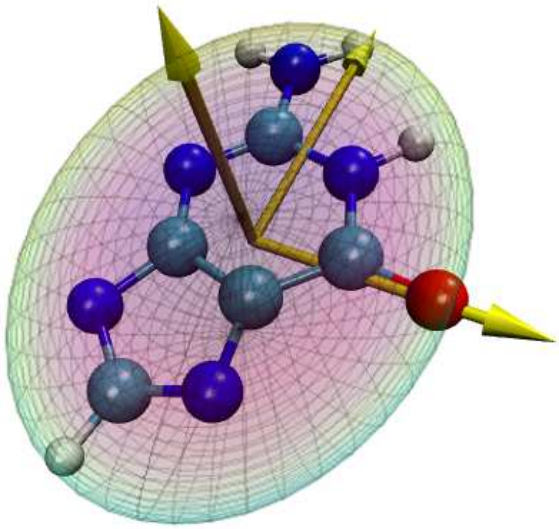

Figure 1.3: An ellipsoid overlaid onto a molecular representation of a base. The principal axes of the base are shown as yellow arrows.

We use ellipsoids to represent the four bases in DNA. As mentioned previously, spheres are crude representations of the steric shape of the bases (see Figure 1.3), whereas an ellipsoid can be flattened to better fit the volume that the bases occupy. Since the ellipsoidal RE^2 potential is already implemented in the LAMMPS simulation software, it is a natural choice for the pair potential between bases.

The steric shape of the bases becomes especially important in any experiment where DNA is forced into a small area. An experiment of particular interest is the translocation of DNA through a nanopore, which is discussed in a review article [43] describing experimental, analytical and computational work on the subject. A bacterial protein called α-haemolysin self-assembles into a structure that creates a pore in the membrane of a cell, although the

experiment can also be done using punctures in artificial membranes. This pore is just large enough for a single strand of DNA to pass through one base at a time. The experiment has a Coulter counter which measures ionic current passing through the pore due to a potential difference put across the membrane. When the negatively charged DNA is also forced through the pore, it causes a blockage in the current as measured by the Coulter counter. In principle, because each of the four bases has a different shape, the drop in measured current can be used to determine which base is blocking the pore. The signal-to-noise ratio has not been perfected enough to read off the sequence of a strand as it passes though. However, once it is possible, it will be a very fast, cheap and nondestructive way to sequence DNA. The methods we currently have to sequence DNA are slower and very expensive; it cost roughly $300 million to sequence the human genome [46]. Atomistic simulations of DNA nanopore translocation have been done for 20 base pair sequences [1]. It would be very useful to have a more computationally efficient model to describe the experiment, and perhaps it could lead to new insights which might improve the experimental procedure to give an effective way to sequence DNA.

1.4 Present Approach and Overview

Our approach is to produce a systematically coarse-grained model representing the bases as ellipsoids. Chapter 2 describes the computational model in detail and how the potentials were obtained from atomistic simulations. Our model is the first to use ellipsoids to represent individual bases with a finite strength hydrogen bond. Chapter 3 provides a careful overview of the additions made to the software simulation package LAMMPS [47] and the tests to which we subjected our modifications. In Chapter 4, explicit definitions are provided for the properties of DNA we will be discussing in Chapter 5, our results and their analysis. Chapter 6 gives conclusions and final thoughts on this research, as well as suggestions for future work. The Appendix contains all the relevant tabulated data referenced in the body of this work.

Chapter 2

Computational Model: Parametrization and Potentials

2.1 Parametrization of the Model

The bond, angle and dihedral potentials, as well as the friction coefficients for the Langevin thermostat, were parametrized by fitting them to the results of atomistic simulations that we conducted using the CHARMM potentials.

2.1.1 Atomistic Simulations of Short Strands

The bond, angle and dihedral potentials were simulated using the smallest piece of DNA that contained all possible types of these potentials. First we took a strand consisting of three base pairs (Figure 2.1A) and we replaced the outer two base pairs with hydrogens, resulting in Figure 2.1B. We made the bond lengths of the added protons equal to the length of the bonds that were removed. We did this to remove base-base interactions from the potentials we wished to parametrize. We addressed the base-base interactions separately using the RE^2 potential.

Then we solvated the molecule in a box of water that extended 8Å beyond the boundary of the molecule in all six directions. We added sufficient K^+ and Cl^- ions so that the net charge of the system was neutral and the average salt concentration was 200mM. This procedure generally added about five ions to the system. This gives Figure 2.1C. The enforcement of charged neutrality biased the addition of salt to the positive ions, because DNA carries a net negative charge from the phosphate backbone.

We prepared four of these simulations, one for each of the possible bases in the system: A, C, G and T.

The method we used to obtain the potentials from the all-atom simulations was Boltzmann inversion. The effective potentials are found from the statistics using:

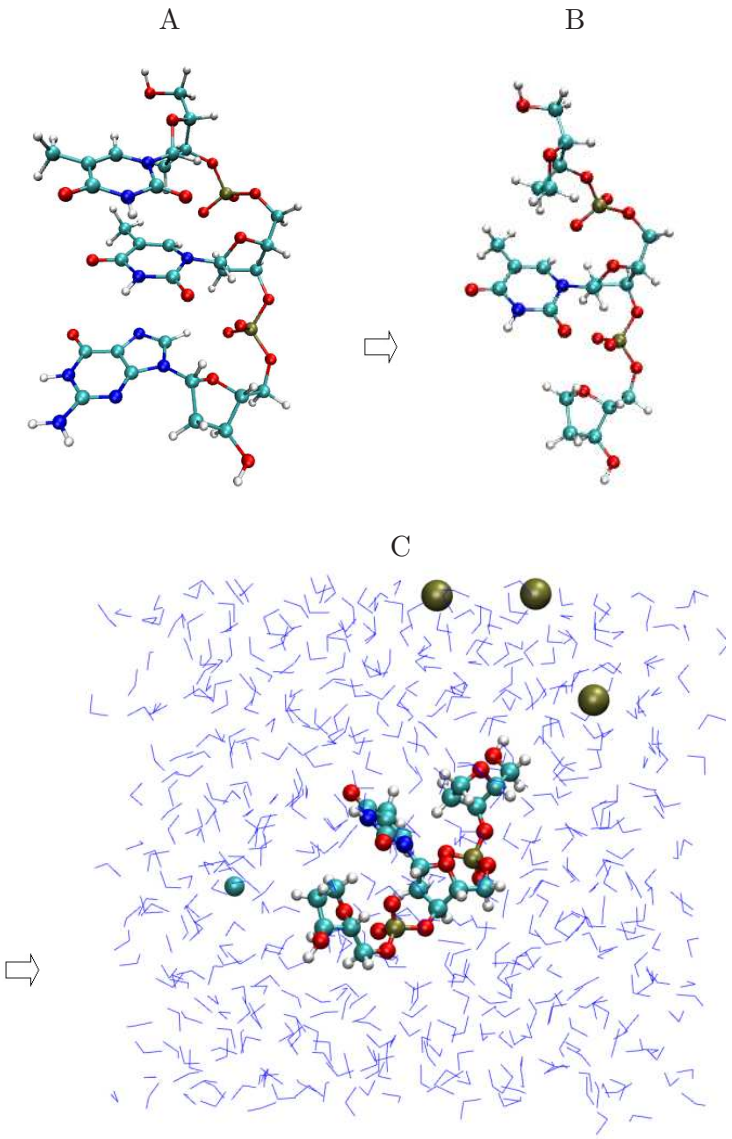

Figure 2.1: Recipe for the determination of bond, angle, and dihedral potentials from all-atom simulations. Rendered in VMD [28].

$$V_{\text{eff}}(\mathbf{x}) = -k_{\text{B}}T \ln p(\mathbf{x}) \tag{2.1}$$

Here $\mathbf{x}$ is the order parameter of the potential and $p(\mathbf{x})$ is the probability

density of observing the system with the value of the order parameter being $\mathbf{x}$. The order parameter is the bond length for bond potentials, the bond angle for angle potentials and the dihedral angle for dihedral potentials (refer to the discussion of these potentials in Section 2.2).

No functional form is assumed *a priori* for the potentials obtained from all-atom data. Rather, once the statistics $p(\mathbf{x})$ were obtained, we analyzed plots of $p(\mathbf{x})$ and $-k_{\mathrm{B}}T \ln p(\mathbf{x})$ choose a good functional form. The criteria we used were that it be straightforward to implement in LAMMPS, and that it not be computationally prohibitive.

We conducted the atomistic simulations for ~ 250ns for each of the four bases. This long simulation time was necessary to allow sufficient sampling of the phase space of the system for the statistics gathered to converge. The parameters that we fit to the bond potentials converged to within 1% of their asymptotic values within 0.02ns. Single well angle potentials converged in 0.5ns, double well angle potentials converged in 2.5ns, and the dihedral potentials converged in 5ns. Dihedral potentials converge slower than the others even though their potentials are shallower because of the larger phase space involved (dihedral potentials involve the greatest number of atoms: 4).

Figure 1.2 depicts both how we label the interaction sites from Figure 2.1 and where the interaction sites are located. The base in the simulation is labelled by its first letter (or sometimes by "X" to refer to a generic base) and the sugar that it is bonded to is labelled by "S". The sugars and phosphates on either side are likewise labelled as "S" and "P", but with an additional identifier indicating which side of the base that it is located on. We do not assume, for example, that the angle potentials $\mathrm{GSP}_{3'}$ and $\mathrm{GSP}_{5'}$ are identical.

2.2 Coarse-Grained Potentials

This section explicitly describes the potential functions used in our course-grained model. The full potential consists of the sum of eight different potential types: bond, angle, dihedral, Yukawa, Lennard-Jones, RE^2 (asphere-asphere), RE^2 (sphere-asphere) and hydrogen bonds.

$$\begin{aligned} V &= V_{\mathrm{bond}} + V_{\mathrm{angle}} + V_{\mathrm{dihedral}} + V_{qq} \\ &\quad + V_{\mathrm{LJ}} + V_{\mathrm{RE}^2} + V_{\text{sph-RE}^2} + V_{\mathrm{HB}}, \end{aligned} \tag{2.2}$$

where V_{type} is understood to mean the sum of all instances of that type, for example V_{bond} consists of the sum of the energies of every bond in the molecule.

2.2.1 Bond Potentials

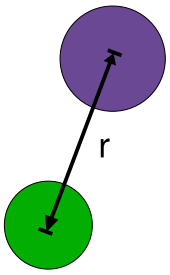

Figure 2.2: Depiction of the bond separation used in the bond potential between two atoms

The bond potential $V_{\text{bond}}(r)$ depends on the separation r between the centers of two bonded atoms, as shown in Figure 2.2.

In Figure 2.3, we see three representative plots of bond potentials being fit to the all-atom statistics. Note that the spring constant of the bond between the sugar and the cytosine is much stiffer than the other bonds; this in part motivated our decision to treat the base-sugar bonds as rigid, as discussed in Section 3.3.

It is not entirely clear that the bond potential for the $P_{3'}S$ bond is best fit to a harmonic potential. However, it is not clearly a double well like we see for some of the angle potentials (see Figure 2.6B). Since the bond potential does not have a distinct secondary well, we decided to fit the data to a harmonic potential about the minimum for simplicity. Adding a second well would change much in our coarse-grained simulations, because the bond is oscillating about its minimum most of the time, as can be seen from the histogram at the bottom of Figure 2.3C.

Table 2.1 gives the complete set of parameters obtained from Equation 2.1. In instances where the potential is base-dependent, the uncertainty is the 95% confidence interval in the fit to the potentials. In cases where the potential does not depend on the base, the uncertainty is obtained from the standard error of the parameters obtained from all for base simulations.

The weakness of the bonds between the sugar and phosphates is caused by the fact that the two participating sites are farther away from each other. This extra separation causes the effective bond between them to be screened, as coarse-grained interactions in statistical physics are always screened on the order of the correlation length of the system. In fact the equilibrium

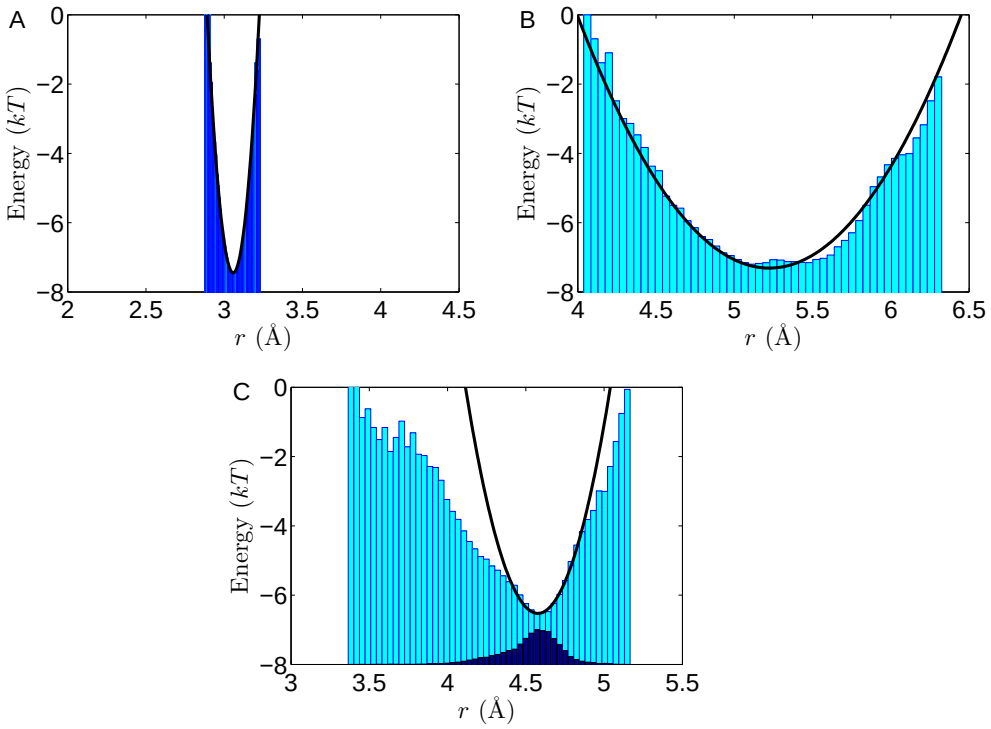

Figure 2.3: Three effective bond potentials in Cytosine: A) r_{CS}, B) $r_{\mathrm{P}_{5'}\mathrm{S}}$ C) the bond between the $3'$ phosphate residue $P_{3'}$ and the sugar, $r_{\mathrm{P}_{3'}\mathrm{S}}$, which shows an anharmonic shoulder that may be accounted for by adding another Gaussian (see discussion of angle potentials in Section 2.2 where this is done), however the amount of time the bond spends at these distances is small: to see this, the scaled probability distribution is shown. The harmonic approximation is thus reasonably good for this bond.

bond separation decreases with the strength of the spring, as seen in Figure 2.4.

The results of this correlation are compared to the phonon dispersion (effective interaction strength) of a 1-D chain of coupled oscillators. This refers to the dependence of strengths of coarse-grained couplings of vibrational modes as a function of the distance between two sites on the chain (see [4]).

The form we obtained for the potential is:

$$V_{\mathrm{bond}}(r) = \frac{1}{2}k_r(r - r_o)^2, \tag{2.3}$$

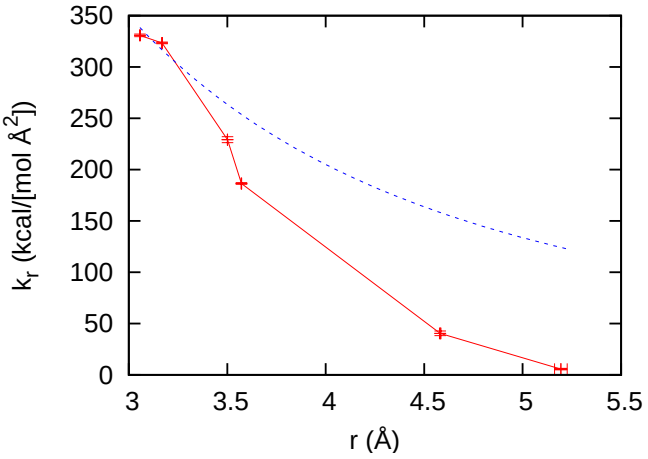

Figure 2.4: Depicts the stiffness coefficient k_r for bond potentials as a function of the mean distance between the participating residues, along with the phonon dispersion relation for a 1-D chain of harmonic oscillators [4], $k_{eff}(\lambda) = k_o \sin^2(\pi a/\lambda)$, with wavelength λ here taken to be the distance r. The correlation coefficient is $r = 0.995$ and chance probability $p \sim 4 \times 10^{-5}$.

where the parameter r_o is the equilibrium separation of the two atoms, measured in Å and k_r is the stiffness of the harmonic spring bonding them, measured in kcal/mol·Å^2. The harmonic approximation is the usual one made for classical models of atomic systems. Since $\lim_{r\to\infty} V_{\text{bond}}(r) = \infty$ the breaking of bonds is impossible, so this model does not involve chemistry at all.

2.2.2 Angle Potentials

The angle potential $V_{\text{angle}}(\theta)$ depends on the bond angle θ between the centers of three consecutively bonded atoms, as shown in Figure 2.5. The angle is calculated from $\theta = \arccos(\hat{\mathbf{r}}_{ij} \cdot \hat{\mathbf{r}}_{kj})$, where the circumflex marks unit vectors.

We found that the bond angle statistics fell into two types: those that were well approximated by a harmonic potential, such as Figure 2.6A, and those that had a clear double well, such as Figure 2.6B. Note that unlike the dihedral angles, the bond angles are calculated purely as the arccos of a dot product, and therefore V_{angle} is not a periodic function of θ. We chose the functional form of Equation 2.5 since we found that the probability density $p(\theta)$ fit perfectly to the sum of two Gaussian distributions, but fit poorly to

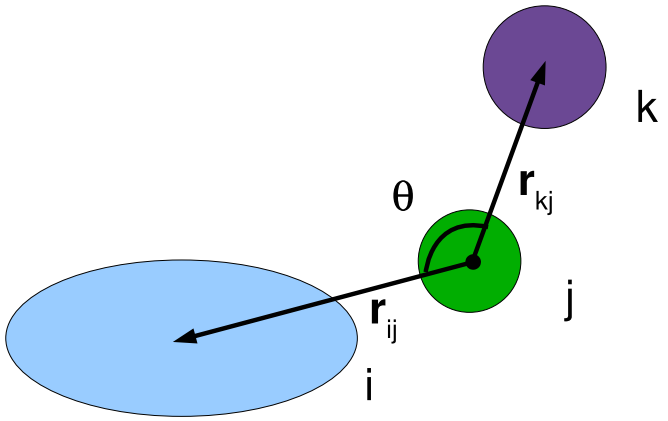

Figure 2.5: Depiction of the bond angle used in the angle potential. Note that an ellipsoid particle may be used in an angle potential, and its center is used for computing $\mathbf{r}_{ij}$.

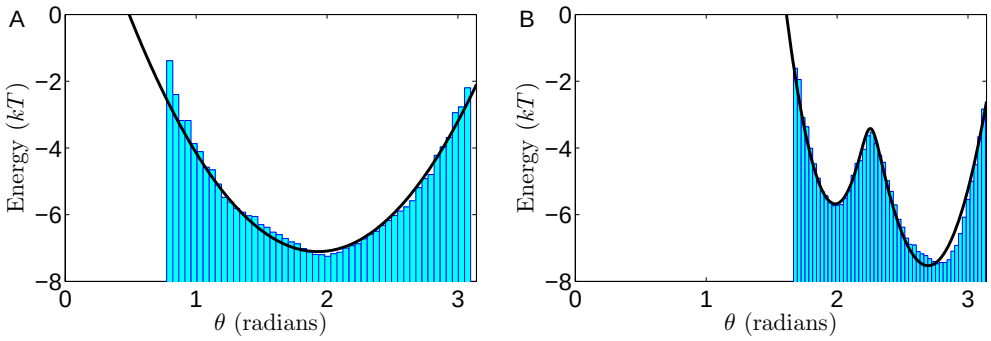

Figure 2.6: Two representative examples of angle potentials A) Harmonic potential between $SP_{3'}S_{3'}$, and B) double well potential between $TSP_{3'}$.

other functions such as polynomials, even up to 8^{th} order. The largest barrier between two wells was about $4k_{\text{B}}T$, for $\theta_{TSP_{3'}}$ (Figure 2.6B). Typically barrier heights between wells of the double well potentials were about $1k_{\text{B}}T$. The parameters for the angle potential are tabulated in Table 2.2.

The explicit forms we find for the angle potentials are as follows:

$$V_{\text{angle}}(\theta) = \frac{1}{2}k_\theta(\theta - \theta_o)^2, \tag{2.4}$$

where the parameter θ_o is the equilibrium separation of the two atoms, measured in radians and k_θ is the stiffness of the harmonic spring bonding them, measured in kcal/mol·rad^2. The other kind of potential is the double well potential, shown in Figure 2.6B, whose form is:

$$V_{\text{angle}}(\theta) = -k_B T \ln\left(e^{-k_1(\theta-\theta_1)^2/2k_BT} + Ae^{-k_2(\theta-\theta_2)^2/2k_BT}\right). \quad (2.5)$$

This potential has five parameters: k_1 and k_2 (kcal/mol·rad^2) giving the stiffness of each well, θ_1 and θ_2 (radians) related to the two minima of the double well, and A which is a number between 0 and 1 that indicates how strong the secondary well is compared to the primary well. The form of the potential 2.5 by Boltzmann statistics means that the probability distribution of the angle is the sum of two Gaussian distributions.

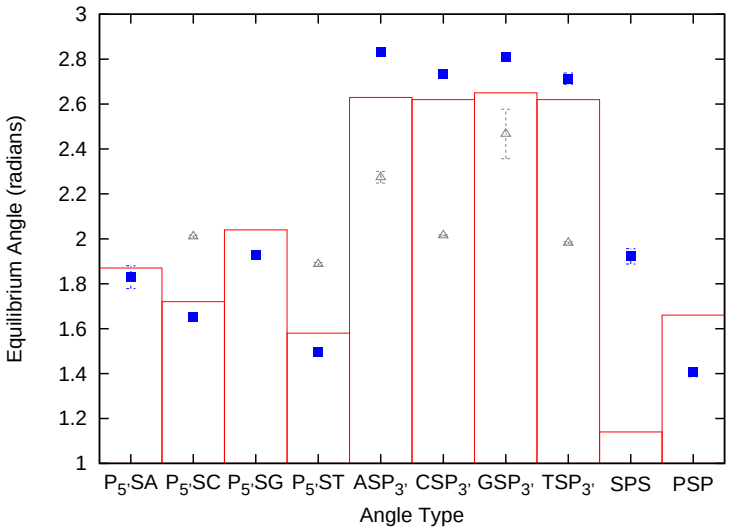

Figure 2.7: Angle equilibrium values. The angles that are in the PDB crystal structure are shown as a histogram. The data points are the minima of the coarse-grained potentials we use. In the cases where the potential is a double well, the locations of both the primary (blue boxes) and secondary minima (grey triangles) are shown.

In Figure 2.7 we see how the minima of the angle potentials that we obtained from all-atom statistics compare with the angles seen in the PDB

structure DNA [3]. For most types of angle potentials, we find that the well lies fairly close to the PDB structure, and in cases with double welled angle potentials, the PDB angles lie between the two parametrized wells and are closer to the deeper primary one. Therefore our parametrization, which is taken from a piece of DNA too small to form anything resembling a double helix, in fact biases the angle potentials to a double helix. There are two exceptions: the angles along the backbone differ substantially from the PDB structure. This is not a problem, however, as there are many factors which will influence what angles the coarse-grained particles will take in relation to each other. The most significant factor is when adjacent bases along either strand try to stack with each other and pull the backbone into the necessary angles. This observation was made by Tepper *et al* in [53], who set the equilibrium of the backbone angle potentials to π, which lies above the entire graph in Figure 2.7.

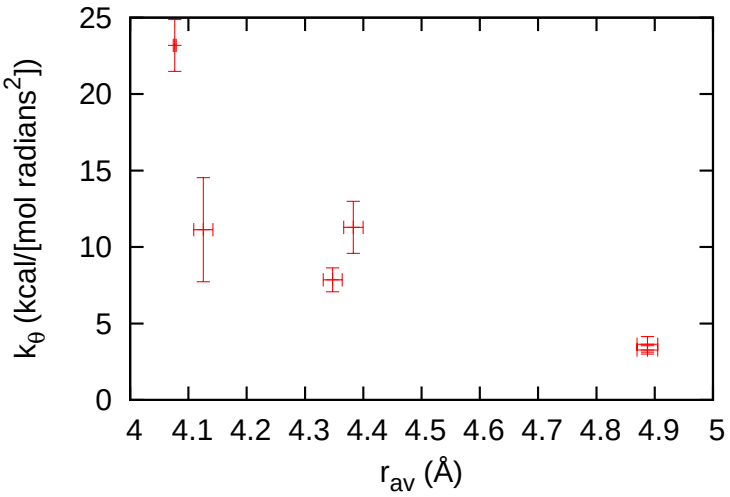

Figure 2.8: Stiffness coefficient k_θ for angle potentials is plotted as a function of the mean length of the two bonds participating in the angle.

Similarly in Figure 2.4, the strength of the angle potential is related to how close on average the participating atoms are to one another. The equivalent angle plot is shown in Figure 2.8.

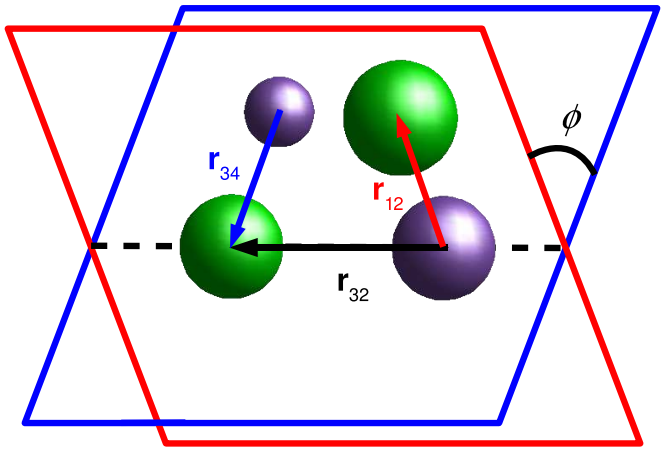

Figure 2.9: Depiction of the dihedral angle used in the dihedral potential. Note that an ellipsoid particle may be used in a dihedral potential, and its center is used for computing the bond vectors.

2.2.3 Dihedral Potentials

The dihedral angle ϕ is determined from four consecutively bonded atoms. It is the angle between the plane of the first three atoms and the plane of the last three atoms. The angle may take on any value from $-\pi$ to π (see Figure 2.9):

$$\phi = \text{sign}((\mathbf{r}_{12} \times \mathbf{r}_{32}) \cdot \mathbf{r}_{34}) \arccos\left(\frac{(\mathbf{r}_{12} \times \mathbf{r}_{32}) \cdot (\mathbf{r}_{32} \times \mathbf{r}_{34})}{|(\mathbf{r}_{12} \times \mathbf{r}_{32})||(\mathbf{r}_{32} \times \mathbf{r}_{34})|}\right), \tag{2.6}$$

where sign$(\cdots)$ is ± 1 depending on the sign of the argument.

The dihedral parameters are fit to periodic functions of ϕ:.

$$V_{\text{dihed}}(\phi) = \sum_{n=1}^{3} K_n(1 - \cos(n\phi - \phi_n)), \tag{2.7}$$

with six parameters $\{K_n, \phi_n\}_{n=1,2,3}$. Four representative fits are shown in Figure 2.10. The form of potentials such as $P_{5'}SP_{3'}S_{3'}$, which did not involve the base, had very close parametrizations, and the true parameters were taken as the average. One can see in Figure 2.10D that the fit is not perfect for this reason, but the qualitative shape of the potential is more or less unchanged by the averaging. Barriers between local minima on the dihedral

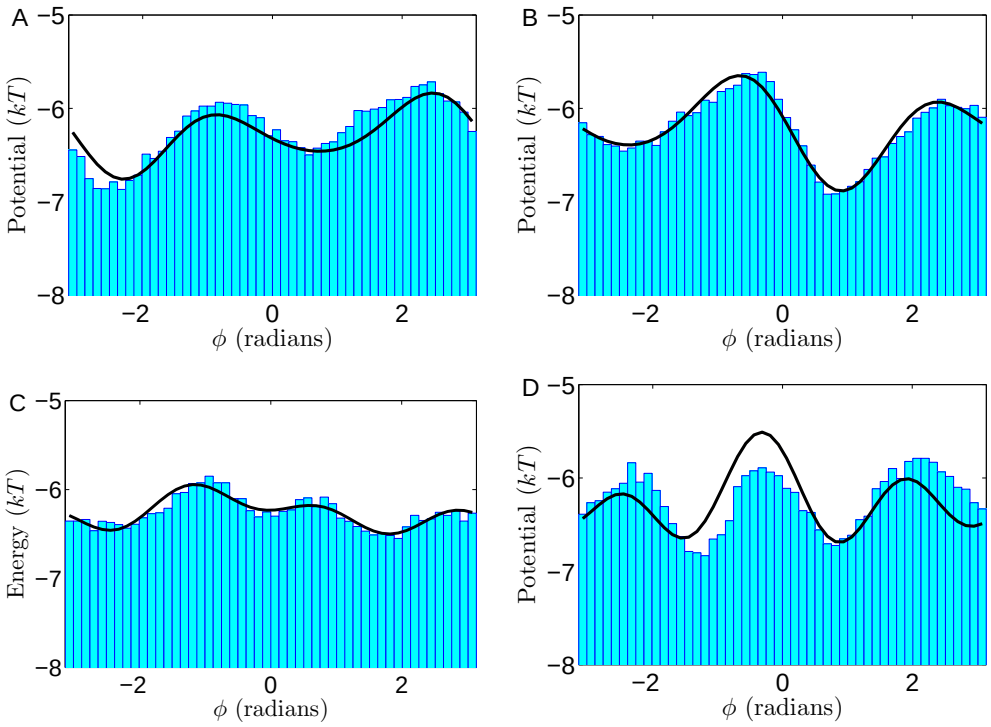

Figure 2.10: Figures are for dihedral potentials for the angles A) $S_{5'}P_{5'}SP_{3'}$, B) $S_{5'}P_{5'}SA$, C) $GSP_{3'}S_{3'}$ and D) $P_{5'}SP_{3'}S_{3'}$. The potential fits for $S_{5'}P_{5'}SP_{3'}$ and $P_{5'}SP_{3'}S_{3'}$ are the result of the parameters obtained by averaging over all bases and given in Table 2.3. We do not take these potentials to be base dependent, although they could be in principle.

potential were all less than about $1.5k_{\mathrm{B}}T$, indicating rotations with respect to the dihedral are much less energetically inhibited than the bond or angle potentials.

There is an improper angle between the normals to the planes defined by $P_{3'}SXP_{5'}$ where X represents the base. This angle is taken to be zero when the $P_{3'}$ residue coincides with $P_{5'}$, and π when $P_{3'}SXP_{5'}$ all lie in the same plane. The calculation of the improper angle is performed as if it were a normal dihedral angle as per Figure 2.9, only the bond structure of atoms 1, 2, 3, 4 is different. The parametrizations of the improper dihedral potentials are summarized in Table 2.4.

2.2.4 Yukawa Potential

The Yukawa potential represents the electrostatic interaction between two charged atoms. The Coulomb energy between the charges is considered to be exponentially screened with their separation because of the ionic solution between them. This screened potential is well approximated by Debye-Hückel theory

$$V_{qq}(r) = \frac{q_i q_j e^{-r/\ell_{\mathrm{D}}}}{4\pi\epsilon_o \varepsilon r}, \tag{2.8}$$

where $\ell_{\mathrm{D}} = (\epsilon_o \varepsilon k_{\mathrm{B}} T / 2e^2 C)^{1/2}$ is the Debye length, ϵ_o is the permittivity of free space, ε the dielectric constant (78 for water at STP) and C is the ionic concentration. For a typical condition of $200mM$ ionic concentration of KCl, the Debye length is $\ell_{\mathrm{D}} \approx 6.8$ Å. In our case, phosphates are the charged residues, and they each carry a charge of $-e$. This purely repulsive potential gives various energies at different separations of the phosphate sites shown in Table 2.5.

2.2.5 Lennard-Jones Potential

The van der Waals attraction and repulsion between any pair of spherical particles is modelled by a Lennard-Jones potential. It is attractive for intermediate atomic separations, repulsive for short separations and flat for large separations.

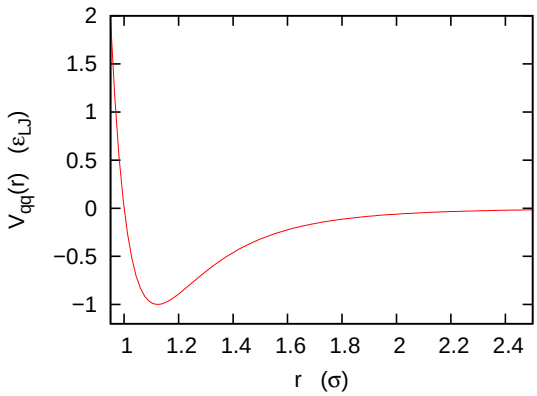

Figure 2.11: The functional form of the Lennard-Jones potential.

The functional form is:

$$V_{LJ}(r) = 4\epsilon_{\text{LJ}} \left[(\sigma/r)^{12} - (\sigma/r)^{6} \right], \tag{2.9}$$

where ϵ_{LJ} is the energy well depth of the potential, r is the separation between atoms, and σ is the separation at which the potential is zero and is a measure of the sum of the radii of the constituent atoms. A plot of the potential is shown in Figure 2.11. Because the potential goes to zero very quickly for large separations, we apply a cutoff r_{cut}, where for $r > r_{\text{cut}}$ we do not compute the Lennard-Jones potential.

We use the Lennard-Jones potential only between the spherical sugar residues, and the potential is calculated for all sugars more than one base away from a given sugar. From Equation 2.9 we use a weak potential of $\epsilon_{\text{LJ}} = 0.25$ kcal/mol with a separation distance of 2Å. This was chosen qualitatively to prevent steric overlap of the backbone residues, without inducing a long range attraction. We did not observe a substantial attraction between sugar residues from our all-atom simulations.

2.2.6 Base Orientation Using Quaternion Vectors

We keep track of the orientations of the ellipsoidal objects using a 4-vector $\mathbf{q} = (q_1, q_2, q_3, q_4)$ in the mathematical space of quaternions, a well known extension of the imaginary numbers to a four dimensional space. We restrict $\mathbf{q}$ to the unit 3-sphere in the space ($\sum_{i=1}^{4} q_i^2 = 1$). A rotation about a unit vector (a, b, c) in the lab frame by an angle θ is represented as a quaternion as $\mathbf{q} = (\cos(\theta/2), a\sin(\theta/2), b\sin(\theta/2), c\sin(\theta/2))$.

Quaternions have the distinct advantage of being more computationally stable than the Euler angle representation and they turn composite rotations into purely algebraic operations. For more information, see [16, 17].

2.2.7 RE^2 Potential (Asphere-Asphere)

We use this potential to model the van der Waals attraction and repulsion between ellipsoidal particles, and it is by far the most complex and computationally expensive potential that we use. It involves significant amounts of matrix algebra and coordinate rotations between the frame or reference of the two participating ellipsoids and the lab frame.

The RE^2 potential, developed by Everaers and Ejtehadi [18], is a generalization of the Lennard-Jones interaction to ellipsoidal particles. The extra rotational degrees of freedom accounted for by the RE^2 potential come with

a greater computational burden: the RE^2 potential takes five times the number of input parameters per interaction pair (including the cutoff distance) as does the Lennard-Jones potential.

The first six of these parameters come from the radii of the two ellipsoids, expressed as the following shape tensor:

$$\mathbf{S_i} = \mathrm{diag}(\sigma_x^{(i)}, \sigma_y^{(i)}, \sigma_z^{(i)}) \tag{2.10}$$

The anisotropic well depths are expressed as the relative potential well depth tensor. Its entries are dimensionless and are inversely proportional to the well depths in their respective directions.

$$\mathbf{E}_i = \mathrm{diag}(\epsilon_x^{(i)}, \epsilon_y^{(i)}, \epsilon_z^{(i)}) \tag{2.11}$$

Another input parameter with a dimension of distance is the atomic interaction radius, σ_c, which characterizes the distance scale for interactions between the atomic constituents to be coarse-grained. The energy scale is given by the Hamaker constant, A_{ij}. Finally, there is the necessary interaction cutoff distance, R_{cut}, as the computation time to compute all $\mathrm{O(N^2)}$ RE^2 interactions is too prohibitive, and is unnecessary because of how quickly the potential decays with separation.

The RE^2 potential between two ellipsoids, labeled as $i = 1, 2$, can be conveniently written in terms of attractive and repulsive components. In the expressions below, the tensor $\mathbf{A}_i$ is the rotation matrix from the lab frame to the rotated principal axis frame of particle i.

$$\begin{aligned} V_A^{RE^2}(\mathbf{A}_1, \mathbf{A}_2, \mathbf{r}_{12}) &= -\frac{A_{12}}{36}\Big(1 + 3\eta_{12}\chi_{12}\frac{\sigma_c}{h_{12}}\Big) \\ &\quad \times \prod_{i=1}^{2}\prod_{e=x,y,z}\left(\frac{\sigma_e^{(i)}}{\sigma_e^{(i)} + h_{12}/2}\right) \end{aligned} \tag{2.12}$$

$$\begin{aligned} V_R^{RE^2}(\mathbf{A}_1, \mathbf{A}_2, \mathbf{r}_{12}) &= \frac{A_{12}}{2025}\Big(\frac{\sigma_c}{h_{12}}\Big)^6\Big(1 + \frac{45}{56}\eta_{12}\chi_{12}\frac{\sigma_c}{h_{12}}\Big) \\ &\quad \times \prod_{i=1}^{2}\prod_{e=x,y,z}\left(\frac{\sigma_e^{(i)}}{\sigma_e^{(i)} + h_{12}/60^{\frac{1}{3}}}\right) \end{aligned} \tag{2.13}$$

The vector $\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2$ is the vector pointing from the center of ellipsoid 2 to the center of ellipsoid 1.

The final potential between the two ellipsoids, depending on the positions and quaternion vectors of the two ellipsoids, is:

$$V_{\mathrm{RE}^2}(\mathbf{r}_1, \mathbf{q}_1, \mathbf{r}_2, \mathbf{q}_2) = V_A^{RE^2}(\mathbf{A}_1, \mathbf{A}_2, \mathbf{r}_{12}) + V_R^{RE^2}(\mathbf{A}_1, \mathbf{A}_2, \mathbf{r}_{12}) \tag{2.14}$$

The quantities η_{12}, χ_{12} and h_{12} in equations (2.12-2.12) are recursively defined by

$$\eta_{12}(\mathbf{A}_1, \mathbf{A}_2) = \frac{\det[\mathbf{S}_1]/\sigma_1^2 + \det[\mathbf{S}_2]/\sigma_2^2}{[\det[\mathbf{H}_{12}]/(\sigma_1+\sigma_2)]^{1/2}} \tag{2.15}$$

$$\mathbf{H}_{12}(\mathbf{A}_1, \mathbf{A}_2, \hat{\mathbf{r}}) = \frac{1}{\sigma_1}\mathbf{A}_1^T\mathbf{S}_1^2\mathbf{A}_1 + \frac{1}{\sigma_2}\mathbf{A}_2^T\mathbf{S}_2^2\mathbf{A}_2 \tag{2.16}$$

$$\sigma_i(\mathbf{A}_i, \hat{\mathbf{r}}_{12}) = \left(\hat{\mathbf{r}}_{12}^T \;\; \mathbf{A}_i^T\mathbf{S}_i^{-2}\mathbf{A}_i\hat{\mathbf{r}}_{12}\right)^{-1/2} \tag{2.17}$$

$$\chi_{12} = 2\hat{\mathbf{r}}_{12}^T\mathbf{B}^{-1}\hat{\mathbf{r}}_{12} \tag{2.18}$$

$$\hat{\mathbf{r}}_{12} = \mathbf{r}_{12}/|\mathbf{r}_{12}| \tag{2.19}$$

$$\mathbf{B} = \mathbf{A}_1^T\mathbf{E_1}\mathbf{A}_1 + \mathbf{A}_2^T\mathbf{E_2}\mathbf{A}_2 \tag{2.20}$$

$$h_{12} = |\hat{\mathbf{r}}_{12}| - \sigma_{12} \tag{2.21}$$

$$\sigma_{12} = \left[\frac{1}{2}\hat{\mathbf{r}}_{12}^T\mathbf{G}^{-1}\hat{\mathbf{r}}_{12}\right]^{-1/2} \tag{2.22}$$

$$\mathbf{G} = \mathbf{A}_1^T\mathbf{S}_1^2\mathbf{A}_1 + \mathbf{A}_2^T\mathbf{S}_2^2\mathbf{A}_2 \tag{2.23}$$

While it is very difficult to show the full functional form of the RE^2 potential, two cross-sections are shown in Figure 2.12. The dependence of the potential on the separation of the ellipsoids (Figure 2.12 A and C) is very similar to the Lennard-Jones potential in Figure 2.11. The orientation dependence of the potential (Figure 2.12 B and D) prefers certain alignments of two stacked bases over others. Here we see that the angular dependence prefers to align the bases with a cost of a little over 1 kcal/mol to have them placed orthogonal to each other. Note that the potential is symmetric about 0, $\pi/2$ and π because of various rotational and reflective symmetries of an ellipsoid.

There are 16 parameters in the RE^2 potential as discussed in Section 2.2.7: the radii for each of the two bases: σ_x, σ_y and σ_z; the "reciprocal well-depths" ϵ_x, ϵ_y and ϵ_z for each of the two bases; a parameter σ_c characterizing the interaction range between atoms in the all-atom potential; and the Hamaker constant A_{12} which scales the energy of the interaction. These parameters are found for the 10 possible base-base interactions (A-A, A-C, etc...) by fitting the RE^2 functional form to the all-atom MM3 force field [37] with Buckingham exponential-6 potentials, following the parametrization

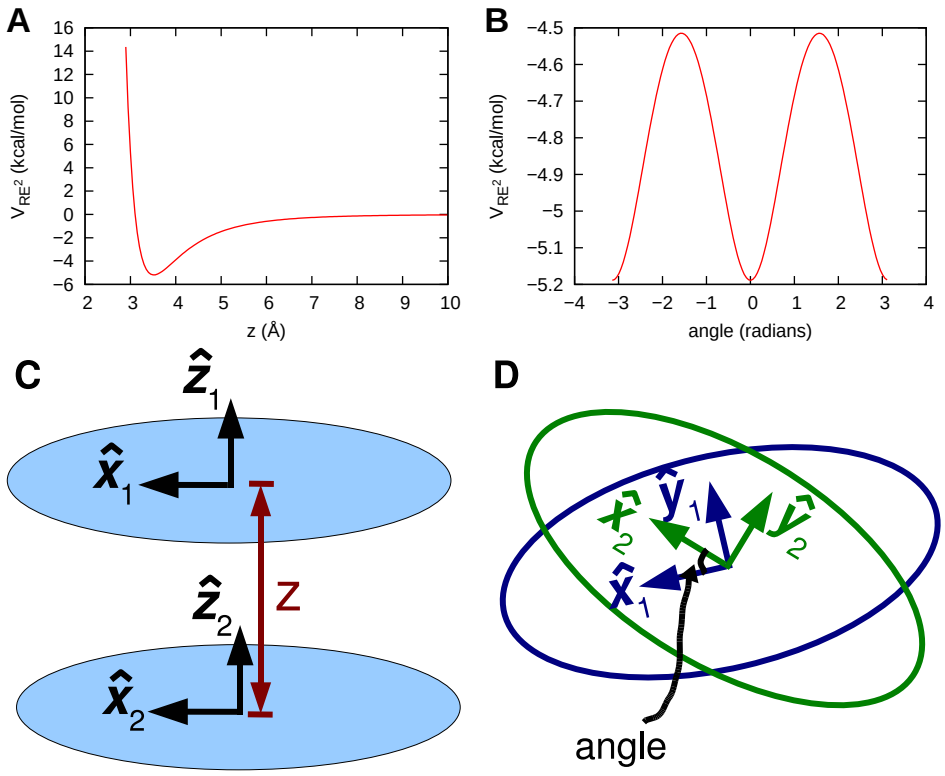

Figure 2.12: These depict two cross-sections of the functional form of the RE^2 potential. The parameters used for these graphs are those of two guanine bases. Figure A shows the dependence of the potential on the z separation of two perfectly stacked ellipsoids. Figure C defines z pictorially. The $\hat{\mathbf{y}}$ axis is coming out of the page. Figure B shows part of the dependence of the RE^2 potential on orientation. The two bases are placed at the minimizing value of z from Figure A and allowed to rotate around the $\hat{\mathbf{z}}$ axis. The angle is the amount of rotation about that axis.

procedure described in [6]. Base-phosphate sphere-asphere interactions were similarly obtained. Parameters for base-base and base-phosphate potentials are shown in Tables 2.6-2.7.

The base-sugar interactions were chosen to have a very small well depth and serve to prevent steric overlap, similar to the Lennard-Jones sugar-sugar interaction. These interaction parameters are tabulated in Table 2.8, and

more physical energy and distance parameters are given in Table 2.9.

2.2.8 RE^2 Potential (Sphere-Asphere)

In LAMMPS, an RE^2-like interaction between an ellipsoid and a sphere may be specified by reducing the radial parameters of the sphere to zero. The radius of the sphere is given by σ_c. In this case, the RE^2 interaction between the objects is calculated in the limit that $\mathbf{S}_2 \to \mathbf{0}$ and $A_{12} \to \infty$ at a rate of $A_{12} \sim 1/\det(\mathbf{S}_2)$. For this interaction, we will call the relevant energy parameter $\tilde{A}_{12} \equiv A_{12}/\rho\sigma_c^3$, where ρ is the number density of the sphere. That is, $\frac{4}{3}\pi \det(\mathbf{S}_2)\rho = 1$. The potential may then be straightforwardly evaluated by substituting $\mathbf{A}_2 = \mathbf{I}$ and $\mathbf{S}_2 = \mathbf{0}$ into the RE^2 potential:

$$\begin{aligned}
\tilde{V}_A(\mathbf{A}_1, \mathbf{r}_{12}) &= -\frac{\tilde{A}_{12}}{36}\frac{4\sigma_c^3}{3\pi}\frac{8}{h_{12}}\Big(1 + 3\chi_{12}\frac{\sigma_c}{h_{12}}\Big) \\
&\quad \times \prod_{e=x,y,z}\left(\frac{\sigma_e^{(1)}}{\sigma_e^{(1)} + h_{12}/2}\right) \qquad (2.24) \\
\tilde{V}_R(\mathbf{A}_1, \mathbf{r}_{12}) &= \frac{\tilde{A}_{12}}{2025}\frac{4\sigma_c^3}{3\pi}\frac{60}{h_{12}}\Big(\frac{\sigma_c}{h_{12}}\Big)^6\Big(1 + \frac{45}{56}\chi_{12}\frac{\sigma_c}{h_{12}}\Big) \\
&\quad \times \prod_{e=x,y,z}\left(\frac{\sigma_e^{(1)}}{\sigma_e^{(1)} + h_{12}/60^{\frac{1}{3}}}\right) \qquad (2.25)
\end{aligned}$$

The final potential between the sphere and the ellipsoid, depending on the positions of the two particles and the quaternion vector of the ellipsoid, is:

$$V_{\text{sph-RE}^2}(\mathbf{r}_1, \mathbf{q}_1, \mathbf{r}_2) = \tilde{V}_A(\mathbf{A}_1, \mathbf{r}_{12}) + \tilde{V}_R(\mathbf{A}_1, \mathbf{r}_{12}) \qquad (2.26)$$

A more complete discussion of the RE^2 potential including its advantages over the alternative biaxial Gay-Berne potential [7] is discussed in the literature [6].

2.2.9 Hydrogen Bond Potential

In addition to the RE^2 interaction between bases, we introduce a hydrogen bond potential that acts between any two bases of complementary type according to Watson-Crick base pairing (A-T and G-C). It depends on both the orientation and separation of the two bases and is highly directional.

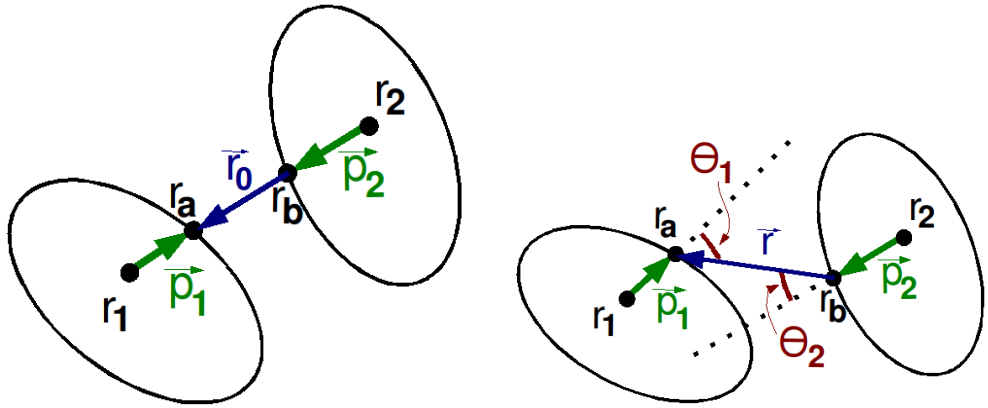

Figure 2.13: Explanation of variables used in the hydrogen bond potential in Equation 2.27. The bases participating in the interaction are labelled as base 1 and 2. The left image shows the two bases aligned in the minimum energy configuration, and the right figure shows the bases in a higher energy configuration where $\mathbf{r} \neq \mathbf{r}_0$ and $\theta_1 \neq 0 \neq \theta_2$.

The form of the potential that we use is:

$$\begin{aligned} V_{\mathrm{HB}}(\{\mathbf{r}_i, \mathbf{q}_i\}_{i=1,2}) &= \epsilon_{\mathrm{HB}} \left[5 \left(\tfrac{r_o}{r}\right)^{12} - 6 \left(\tfrac{r_o}{r}\right)^{10}\right] \\ &\quad \times \Big(\cos^4(3\theta_1)\cos^4(3\phi_1) + \\ &\quad \cos^4(3\theta_2)\cos^4(3\phi_2) \Big) \end{aligned} \tag{2.27}$$

In addition to the energy parameter ϵ_{HB} and equilibrium separation parameter r_o evident from equation Equation 2.27, there are two vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ used to define how the angles θ and ϕ are calculated in the hydrogen bond potential between the bases labelled 1 and 2. These vectors, in the frame of reference of their respective bases, point to the location on the edge of the ellipsoid which indicates the site on the base participating in the hydrogen bond. They are chosen such that in the minimal energy configuration of the hydrogen bond, they are antiparallel as shown on the left of Figure 2.13. Also, $|\mathbf{p}_1| + |\mathbf{p}_2| + |\mathbf{r}_0|$ is chosen to be the distance separating the center of masses of the two bases in their equilibrium structure.

The separation r is the simplest to calculate. With $\mathbf{r}_1$ and $\mathbf{r}_2$ being the center of masses of bases 1 and 2, we define $\mathbf{r}_a = \mathbf{r}_1 + \mathbf{p}_1$ and $\mathbf{r}_b = \mathbf{r}_2 + \mathbf{p}_2$. Then the vector $\mathbf{r}$ is defined as $\mathbf{r} = \mathbf{r}_a - \mathbf{r}_b$ and $r = |\mathbf{r}|$.

The hydrogen bond potential for a Watson-Crick pair is the sum of two

potentials $V_{\mathrm{HB}}^{(1)}+V_{\mathrm{HB}}^{(2)}$, representing the hydrogen bonds from base $1 \rightarrow 2$ and base $2 \rightarrow 1$ respectively. $V_{HB}^{(1)}$ is a three-body potential between the center of mass of base 1, $\mathbf{r}_1$, the origin of the hydrogen bond on the exterior of the ellipsoid, $\mathbf{r}_{\mathrm{A}}$, and the donor or acceptor on the other base, $\mathbf{r}_{\mathrm{B}}$. Hydrogen bonding in the other direction is similarly defined. The following discussion of θ and ϕ will refer only to the $1 \rightarrow 2$ direction, owing to the symmetric definitions of the reverse direction.

At the right of Figure 2.13, we can see that θ_1 is defined as:

$$\theta_1 = \arccos\left(-\frac{\mathbf{r}\cdot\mathbf{p}_1}{|\mathbf{r}||\mathbf{p}_1|}\right) \tag{2.28}$$

Introducing ϕ_1-dependent terms breaks the symmetry of rotations of the first base about $\mathbf{p}_1$ whereas rotations of the second base about $\mathbf{p}_2$ are broken by the ϕ_2 term. Defining $\mathbf{n}_1$ and $\mathbf{n}_2$ as the normals of bases 1 and 2 respectively, we proceed to define ϕ_1. In Fig. 2.13, these normals are all pointing out of the page for simplicity. We project the normal $\mathbf{n}_2$ onto the plane perpendicular to $\mathbf{p}_1$, and ϕ_1 will be the angle between this projected vector and $\mathbf{n}_1$, or precisely:

$$\cos(\phi_1) = \mathbf{n}_1 \cdot \frac{\mathbf{n}_2 - \mathbf{n}_2\cdot\mathbf{p}_1}{|\mathbf{n}_2 - \mathbf{n}_2\cdot\mathbf{p}_1|} \tag{2.29}$$

By projecting $\mathbf{n}_2$ onto the plane perpendicular to $\mathbf{p}_1$, the ϕ_1 contribution to the hydrogen bond potential is purely torsional, with the torque on base 1 act entirely on the $\mathbf{p}_1$ axis. This breaks the symmetry of rotations of base 1 about this axis in the simplest manner. Note that we do not project $\mathbf{n}_1$ onto $\mathbf{p}_1$ because we define $\mathbf{p}_1$ to be purely in the $x-y$ plane of base 1 for simplicity (and similarly for base 2).

It is undesirable to have strong hydrogen bonds form between one base and several other bases at once. To get around this, we make the potential (2.27) very orientation dependent, thus making secondary hydrogen bond energies substantially weaker than the strongest hydrogen bond in which the base is participating. In fact with our hydrogen bond potential, we find from our simulations that on average, $95.00 \pm 1.63\%$ of the hydrogen bond energy is concentrated in the strongest hydrogen bond at any given time. This strong orientation dependence in our model is attained by raising the cosine terms in θ and ϕ to the 4^{th} power.

Since hydrogen bonds only form when the bases' hydrogen acceptor/-donor sites are facing each other, we apply an angle cutoff of $\pi/6$ to both the θ and ϕ angles. We do not compute a hydrogen bond if either θ_1 or ϕ_1

is larger than the cutoff. The 3 inside the cosine terms in Equation 2.27 is a result of the fact that our angle cutoff is precisely $\pi/6$. Therefore, when either angle reaches $\pi/6$ its contribution to the hydrogen bond potential vanishes. This also ensures a well-depth ϵ_{HB}, and avoids multiple minima in the cosine terms. If we applied the cutoff condition when the potential was not at zero, the effective well depth would be less than ϵ_{HB} by exactly $|V_{\text{HB}}(\text{angle cutoff})|$.

In implicit water, the well depth of the hydrogen bond represents the free energy difference between bonded and unbonded base pairs in solvent, which has been determined experimentally to be 1.2 ± 0.4 kcal/mol for each hydrogen bond [21]. All three hydrogen bonds are cooperative, which makes the G-C hydrogen bonds more than 3/2 the strength of the A-T pairs. The stability of a base pair is also reduced by the presence of neighbours in the crystal structure, which induces strain on the base pair [38], and reduces the value from that of an isolated base pair [41], to stabilities of 10.2 kcal/mol and 17.2 kcal/mol for A-T and G-C pairs respectively *in vacuo*. The ratio between these two hydrogen bond well depths is 1.686, which when applied to the reasonable estimate of 2.4 kcal/mol for the hydrogen bond strength of A-T pairs in water, gives an estimate of 4.0 kcal/mol for the hydrogen bond strength of G-C pairs. We therefore take 2.4 and 4.0 kcal/mol as our well depths ϵ_{HB} for A-T and G-C hydrogen pairings respectively.

The bond separation r_o, as discussed in Section 2.2.9, is taken so that the energetic minimum reproduces the PDB structure.

2.3 Langevin Thermostat

A Langevin thermostat applies two types of forces to the simulated particles, neither of which is derived from a potential function. It acts as the implicit solvent of the DNA by applying random forces each timestep which simulate the thermal kicks the atoms receive from the surrounding water, acting as a thermal bath. The second feature of water that the Langevin force emulates is the dissipative viscous force which opposes the motions of the particles in the system by a term proportional to the speed of the particles. Likewise, for ellipsoidal particles, the Langevin thermostat applies random and dissipative torques to simulate the asymmetric thermal kicks by the surrounding thermal bath and to oppose rotations in a viscous medium. The functional form of the forces and torques is diagonal with respect to the linear and angular velocities in the principal axis frame of the ellipsoid:

$$F_i^{Lang} = -\zeta_i^{(t)} v_i + \xi_i^{(t)} \tag{2.30}$$

$$\tau_i^{Lang} = -\zeta_i^{(r)}\omega_i + \xi_i^{(r)}. \tag{2.31}$$

Here the subscript i denotes the Cartesian components along each of the ellipsoid's principal axes, v_i and ω_i are the translational and rotational velocities, $\zeta_i^{(t)}$ and $\zeta_i^{(r)}$ are the eigenvalues of the translational and rotational friction tensors, and ξ_i^t and ξ_i^r are components of the translational and rotational thermal noise for whose amplitudes are related to the friction coefficients by the fluctuation dissipation theorem.

For greater computational speed, the ξ thermal noise coefficients are evaluated using a uniform random number distribution instead of the usual Gaussian distribution [14]:

$$\xi_i^t = \sqrt{\zeta_i^{(t)} 24 k_B T/\Delta t} \cdot \text{rand} \tag{2.32}$$

$$\xi_i^r = \sqrt{\zeta_i^{(r)} 24 k_B T/\Delta t} \cdot \text{rand}, \tag{2.33}$$

where Δt is the timestep, and rand is a random number uniformly generated in the interval $-0.5 < \text{rand} < 0.5$. The parameters in the Langevin force are therefore the three translational and three rotational friction coefficients, ζ_i^t and ζ_i^r.

To parametrize the friction coefficients used in the Langevin equations (Equations 2.30-2.31) we performed more atomistic simulations to ensure that the expressions for the friction tensor faithfully reproduced the diffusive motion of the bases. The simulation box set up was similar to that found in Figure 2.1, except that here we simulated the base only, with no backbone. The all-atom simulations were performed with a single base solvated in a $(20\ \text{Å})^3$ box of water in a 0.2 mol/L KCl solution with a net charge of 0. The simulations were performed for 10 ns in the NPT ensemble. Four such simulations were performed, one for each type of base. Temperature was maintained at 310K by Langevin dynamics and pressure was maintained at 1 atm by a Langevin piston.

The values of the friction coefficients which correctly reproduce the base's translational and rotational diffusion statistics can be obtained from the Stokes-Einstein diffusion relations in the frame of reference of the base:

$$D_{r_i}^{(t)} = kT/\zeta_i^{(r)} = \frac{\langle(\Delta r_i(t))^2\rangle}{2t} \tag{2.34}$$

$$D_{\theta_i}^{(r)} = kT/\zeta_i^{(t)} = \frac{\langle(\Delta \theta_i(t))^2\rangle}{2t}, \tag{2.35}$$

where $D^{(t,r)}$ is the translational or rotational diffusion coefficient, the subscript i refers to the i^{th} principal axis of the base, and $\langle(\Delta r_i(t))^2\rangle$ and

$\langle(\Delta\theta_i(t))^2\rangle$ are respectively the average mean square displacement and rotation of the base due to diffusion in time t. We computed the mean square displacements from the all-atom simulations and obtained the correct friction coefficients from the inverse slope of the mean square displacements with time.

It should be noted that the Stokes-Einstein equations hold for any shape, and we have not assumed that the bases are ellipsoids. It is these friction coefficients that we use to parametrize the Langevin thermostat. It is reasonable to ask whether there exists an ellipsoid which reproduces the correct friction coefficients. The friction coefficients of ellipsoidal bodies can be obtained from the low Reynolds number solution to the Navier-Stokes hydrodynamics equations [34] and can be expressed in terms of the three radii r_i of the ellipsoid in the principal axis frame:

$$\zeta_i^{(t)} = \frac{16\pi\eta}{S+G_i} \tag{2.36}$$

$$\zeta_i^{(r)} = \frac{16\pi\eta(r_j^2+r_k^2)}{3(G_j+G_k)}, \tag{2.37}$$

with the following elliptic integrals

$$S = \int_0^\infty d\lambda \left[(r_1^2+\lambda)(r_2^2+\lambda)(r_3^2+\lambda)\right]^{-1/2} \tag{2.38}$$

and

$$G_i = r_i^2 \int_0^\infty d\lambda (r_i^2+\lambda)^{-1} \left[(r_1^2+\lambda)(r_2^2+\lambda)(r_3^2+\lambda)\right]^{-1/2} \tag{2.39}$$

From Equations 2.36-2.37, we used a functional minimization procedure to find the best fit set of radii which minimized the following functional of $\mathbf{r}$:

$$F[\mathbf{r}] = \sqrt{\sum_{i=1}^{3}\left[\left(\frac{\zeta_i^{(t)}(\mathbf{r})-\zeta_{i,o}^{(t)}}{\zeta_{i,o}^{(t)}}\right)^2 + \left(\frac{\zeta_i^{(r)}(\mathbf{r})-\zeta_{i,o}^{(r)}}{\zeta_{i,o}^{(r)}}\right)^2\right]}, \tag{2.40}$$

where the subscript o is used to mean the value of the friction coefficient obtained from the all-atom simulations. These best-fit radii are referred to as the hydrodynamic radii. The obtained hydrodynamic radii and the simulated friction coefficients are shown in Table 2.10.

Figure 2.14 shows a plot of how the hydrodynamic radii correlate with the equipotential radii, which is half the separation distance where the

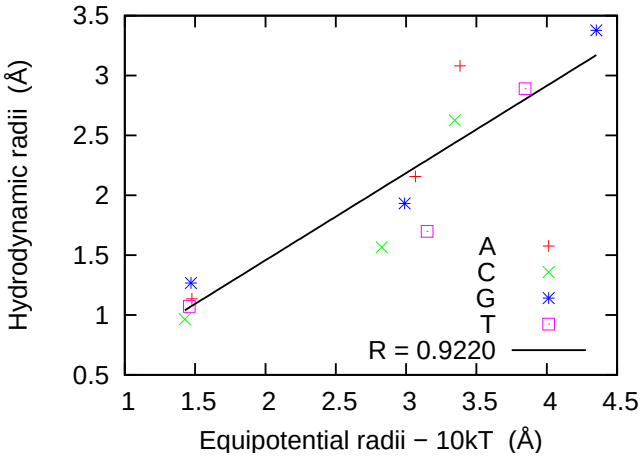

Figure 2.14: Energetic and hydrodynamic radii for effective ellipsoids. In this figure, hydrodynamically-derived radii (see Table 2.10) are plotted vs. the equipotential radii at $10k_BT$ (see Table 2.5).

RE^2 energy attains a given value (e.g. one of them would be half of z at $V_{RE^2} = 10k_BT$ in Figure 2.12). The two radii are strongly correlated, however the equipotential radii tend to be smaller than energetically-derived radii. For the thinnest axis, the assumption of continuum hydrodynamics is expected to be least accurate, and relative modifications due to hydration effects are expected to be largest [15, 20, 23, 26, 51].

It is interesting that the minimization of Equation 2.40 still gives hydrodynamic friction coefficients that differ from the target values by roughly 20%. This indicates that an ellipsoid is not the best approximation to the diffusion statistics of the base. This is not an issue because we use the correct friction coefficients directly rather than assuming an ellipsoid with certain hydrodynamic radii.

2.4 Tables

Table 2.1: Bond potential parameters. Note that the first four bonds were not used in the simulations, as the base-sugar system was treated as a rigid body not two bodies attached by a spring.

Bond	k_r (kcal/mol $\cdot$ Å^2)	r_o (Å)
r_{CS}	330.8 ± 1.2	3.057 ± 0.00074
r_{TS}	323.4 ± 0.84	3.169 ± 0.00018
r_{AS}	229.2 ± 2.9	3.501 ± 0.00065
r_{GS}	186.5 ± 0.74	3.572 ± 0.0014
$r_{\mathrm{SP}_{3'}}$	40.45 ± 2.3	4.581 ± 0.0037
$r_{\mathrm{P}_{5'}\mathrm{S}}$	5.596 ± 0.75	5.194 ± 0.032

Table 2.2: Bond angle potential parameters.

	Harmonic	
	k_θ (kcal/mol)	θ_0 (rad)
θ_{PSP}	19.51 ± 0.69	1.406 ± 0.02
$\theta_{\mathrm{P_{5'}SG}}$	11.29 ± 1.7	1.926 ± 0.0089
$\theta_{\mathrm{P_{5'}SA}}$	7.852 ± 0.78	1.829 ± 0.051
θ_{SPS}	3.453 ± 0.39	1.922 ± 0.034

	Bi-harmonic				
	A	k_1(kcal/mol)	k_2(kcal/mol)	θ_1	θ_2
$\theta_{\mathrm{ASP_{3'}}}$	0.3229 ± 0.087	24.19 ± 5.8	34.57 ± 21	2.83 ± 0.00016	2.274 ± 0.026
$\theta_{\mathrm{CSP_{3'}}}$	0.2169 ± 0.0065	20.69 ± 3.2	29.75 ± 2.4	2.734 ± 0.02	2.015 ± 0.0036
$\theta_{\mathrm{TSP_{3'}}}$	0.168 ± 0.015	28.12 ± 3.2	52.8 ± 3.1	2.713 ± 0.025	1.982 ± 0.0055
$\theta_{\mathrm{GSP_{3'}}}$	0.3702 ± 0.096	32.12 ± 9.2	16.34 ± 9.2	2.81 ± 0.017	2.467 ± 0.11
$\theta_{\mathrm{P_{5'}SC}}$	0.7811 ± 0.018	32.57 ± 3	29.55 ± 0.97	1.652 ± 0.0084	2.01 ± 0.0069
$\theta_{\mathrm{P_{5'}ST}}$	0.5165 ± 0.0041	65.58 ± 2.4	22.18 ± 3.4	1.498 ± 0.0076	1.887 ± 0.0077

Table 2.3: Dihedral potentials

	K_1(kcal/mol)	K_2(kcal/mol)	K_3(kcal/mol)	ϕ_1	ϕ_2	ϕ_3
$\phi_{\mathrm{ASP_{3'}S_{3'}}}$	-0.09 ± 0.01	0.03 ± 0.01	-0.05 ± 0.01	-1.41 ± 0.09	0.8 ± 0.2	-4.0 ± 0.2
$\phi_{\mathrm{CSP_{3'}S_{3'}}}$	-0.10 ± 0.01	0.10 ± 0.01	-0.06 ± 0.01	-1.26 ± 0.07	-0.02 ± 0.07	-8.3 ± 0.1
$\phi_{\mathrm{GSP_{3'}S_{3'}}}$	-0.10 ± 0.01	-0.04 ± 0.01	-0.08 ± 0.01	-0.79 ± 0.08	-1.5 ± 0.2	-3.9 ± 0.1
$\phi_{\mathrm{TSP_{3'}S_{3'}}}$	-0.11 ± 0.01	-0.17 ± 0.01	0.07 ± 0.01	-0.70 ± 0.08	1.85 ± 0.04	-6.7 ± 0.3
$\phi_{\mathrm{S_{5'}P_{5'}SA}}$	-0.13 ± 0.01	-0.26 ± 0.01	-0.05 ± 0.01	-1.62 ± 0.06	-1.39 ± 0.03	-0.7 ± 0.2
$\phi_{\mathrm{S_{5'}P_{5'}SC}}$	-0.07 ± 0.02	-0.22 ± 0.02	-0.05 ± 0.02	-0.9 ± 0.1	-0.72 ± 0.04	0.4 ± 0.2
$\phi_{\mathrm{S_{5'}P_{5'}SG}}$	-0.12 ± 0.02	-0.23 ± 0.02	-0.02 ± 0.02	-14.71 ± 0.08	-1.24 ± 0.04	-0.2 ± 0.3
$\phi_{\mathrm{S_{5'}P_{5'}ST}}$	-0.18 ± 0.03	-0.22 ± 0.03	-0.14 ± 0.03	0.2 ± 0.1	-0.48 ± 0.09	-0.1 ± 0.1
$\phi_{\mathrm{P_{5'}SP_{3'}S_{3'}}}$	0.10 ± 0.02	0.15 ± 0.02	0.21 ± 0.02	0.0 ± 0.1	2.44 ± 0.02	2.17 ± 0.05
$\phi_{\mathrm{S_{5'}P_{5'}SP_{3'}}}$	-0.08 ± 0.01	-0.20 ± 0.01	0.044 ± 0.01	8.0 ± 0.1	-1.50 ± 0.05	-0.9 ± 0.2

Table 2.4: Improper potentials

	Class 2					
	K_1(kcal/mol)	K_2(kcal/mol)	K_3(kcal/mol)	ϕ_1	ϕ_2	ϕ_3
$\phi_{\mathrm{P}_{5'}\mathrm{SAP}_{3'}}$	-1.4 ± 0.2	-0.3 ± 0.2	-0.3 ± 0.1	-0.77 ± 0.04	-0.9 ± 0.2	-0.6 ± 0.2
$\phi_{\mathrm{P}_{5'}\mathrm{SGP}_{3'}}$	-2.0 ± 0.3	-0.7 ± 0.2	-0.4 ± 0.1	-0.51 ± 0.02	-1.04 ± 0.09	-1.0 ± 0.1
$\phi_{\mathrm{P}_{5'}\mathrm{SCP}_{3'}}$	2.36 ± 0.06	0	0	1.97 ± 0.02	0	0
$\phi_{\mathrm{P}_{5'}\mathrm{STP}_{3'}}$	2.66 ± 0.09	0	0	1.62 ± 0.03	0	0

Table 2.5: Equipotential radii (Å) and minimum energies (kT at 310K). The phosphate interaction is purely a screened Coulomb interaction and thus has its minimum at infinity.

		Adenine	Cytosine	Guanine	Thymine	Sugar	Phosphate
$1kT$	r_x	3.5002	3.4702	4.5140	3.9947	0.9418	2.7039
	r_y	3.1963	2.9713	3.1189	3.2961	0.9418	2.7039
	r_z	1.5479	1.5145	1.5391	1.5386	0.9418	2.7039
$10kT$	r_x	3.3836	3.3457	4.3526	3.8454	0.8315	0.3348
	r_y	3.0654	2.8266	2.9891	3.1497	0.8315	0.3348
	r_z	1.4768	1.4263	1.4707	1.4586	0.8315	0.3348
$V_{\min}$	ϵ_x	1.1790	1.1075	0.6649	0.6902	0.1540	0
	ϵ_y	0.9433	0.8102	1.0689	0.7212	0.1540	0
	ϵ_z	2.7053	2.0562	3.1969	2.3001	0.1540	0

Table 2.6: RE[2] Parameters for base-base interactions. Units: σ (Å), ϵ (none) and A_{12} (kcal/mol). The row indicates which base the radii $\sigma_{x,y,z}$ and inverse relative well depths $\epsilon_{x,y,z}$ belong to. The column indicates which base the row-base is interacting with. Note that all the properties of each base depend on what base it is interacting with.

		A	C	G	T
A	σ_x	2.4652	2.2860	2.4370	2.6909
	σ_y	2.1472	1.4051	1.5951	1.9398
	σ_z	0.5578	0.1262	0.5821	0.6384
	ϵ_x	3.6229	4.3800	3.6178	3.7656
	ϵ_y	5.0023l	2.6585	2.8420	2.1395
	ϵ_z	1.4057l	2.4634	1.5088	1.6393
	A_{12}	3455	17115	4188	6195
	σ_c	3.5969	3.6915	3.8501	3.7700
C	σ_x	2.4997	2.3127	2.1181	2.1897
	σ_y	2.0252	1.7950	1.6871	1.9022
	σ_z	0.7082	0.3903	0.4637	0.3069
	ϵ_x	2.9957	2.5848	1.7108	1.1875
	ϵ_y	7.8481	3.8092	3.6023	6.3381
	ϵ_z	1.0379	1.3424	1.163	1.2818
	A_{12}	17115	7826	3720	4839
	σ_c	3.6915	3.9060	3.961	3.8553
G	σ_x	3.1278	3.1633	3.3321	3.3842
	σ_y	2.5891	2.6258	1.9640	1.0974
	σ_z	0.4201	0.5201	0.4342	0.0290
	ϵ_x	3.5849	3.3655	4.3347	5.4026
	ϵ_y	4.4283	2.4695	2.3717	0.8356
	ϵ_z	0.6581	0.8089	0.7752	1.1117
	A_{12}	4188	3720	4274	51144
	σ_c	3.8501	3.961	4.0250	3.8664
T	σ_x	2.7194	3.1334	3.2489	2.9091
	σ_y	1.6628	2.0441	2.2016	2.2128
	σ_z	0.2994	0.5824	0.7588	0.5062
	ϵ_x	4.8654	4.5408	5.5485	3.6127
	ϵ_y	5.5447	3.3117	4.9745	3.4088
	ϵ_z	0.8189	0.9092	0.9895	0.9666
	A_{12}	6195	4839	51144	2437
	σ_c	3.7700	3.8553	3.8664	3.8096

Table 2.7: RE^2 Parameters for base-phosphate sphere-asphere interactions. The $\sigma_{x,y,z}$ and $\epsilon_{x,y,z}$ parameters are properties of the base. Spherical particles under the RE^2 potential are considered to have $\sigma_{x,y,z} = 0$ and $\epsilon_{x,y,z} = 1$. Units: σ (Å), ϵ (none) and A_{12} (kcal/mol).

	A	C	G	T
σ_x	1.9844	1.7323	73.294	1.9983
σ_y	1.1571	1.2331	1.1070	1.5560
σ_z	0.1533	0.2070	0.2470	0.0966
ϵ_x	13.443	12.917	8.7902	15.0740
ϵ_y	9.0257	8.0570	4.1412	12.8730
ϵ_z	2.6714	3.5227	1.7311	2.9361
$\tilde{A}_{12}$	729.67	534.84	244.45	934.7900
σ_c	4.2489	4.1348	4.1409	4.2940

Table 2.8: RE^2 Parameters for base-sugar sphere-asphere interactions. The $\sigma_{x,y,z}$ and $\epsilon_{x,y,z}$ parameters are properties of the base. Spherical particles under the RE^2 potential are considered to have $\sigma_{x,y,z} = 0$ and $\epsilon_{x,y,z} = 1$. Units: σ (Å), ϵ (none) and A_{12} (kcal/mol).

	A	C	G	T
σ_x	2.4652	2.3127	3.3321	2.9091
σ_y	2.1472	1.7950	1.9640	2.2128
σ_z	0.5578	0.3903	0.4342	0.5062
ϵ_x	3.6229	2.5848	4.3347	3.6127
ϵ_y	5.0023	3.8092	2.3717	3.4088
ϵ_z	1.4057	1.3424	0.7752	0.9666
$\tilde{A}_{12}$	0.25	0.25	0.25	0.25
σ_c	2	2	2	2

Table 2.9: Equipotential separation distances (Å) and minimum energies (kT at 310K) between Bases and S or P residues. Here distance rather than radius is used between unlike partners.

		A-P	C-P	G-P	T-P	A-S	C-S	G-S	T-S
$1kT$	d_x	6.1314	5.6869	7.7365	6.2227	4.5239	4.2680	5.7290	5.1273
	d_y	4.9390	4.9517	4.6075	5.5883	4.0709	3.5264	3.8037	4.1450
	d_z	3.4256	3.4359	3.3415	3.4049	1.8533	1.5695	1.6753	1.7636
$10kT$	d_x	5.7067	5.2886	7.3162	5.8186	4.3228	4.0728	5.5260	4.9296
	d_y	4.5398	4.5829	4.2459	5.1922	3.8663	3.3322	3.6158	3.9445
	d_z	3.1302	3.1330	3.0542	3.1168	1.6528	1.3725	1.4730	1.5676
$V_{\min}$	ϵ_x	0.2561	0.2644	0.2361	0.2419	0.0054	0.0039	0.0037	0.0047
	ϵ_y	0.3152	0.3417	0.3872	0.2600	0.0043	0.0030	0.0056	0.0049
	ϵ_z	0.7223	0.5997	0.7177	0.6692	0.0102	0.0060	0.0108	0.0109

Table 2.10: Translational and rotational friction coefficients of the bases found from atomistic simulations and the radii of ellipsoids which best fit the friction coefficients. These are referred to as the hydrodynamic radii of the bases.

	ζ^{trans} $(10^{-11} kg/s)$				ζ^{rot} $(10^{-30} kg\ m^2/s)$			
	A	C	G	T	A	C	G	T
ζ_x	0.3961	0.2975	0.4730	0.3344	0.1467	0.0779	0.1616	0.1038
ζ_y	0.3972	0.3771	0.4974	0.3994	0.2183	0.1371	0.2910	0.1840
ζ_z	0.4835	0.5163	0.5524	0.4988	0.4046	0.1988	0.4091	0.2607

	$\sigma^{(H)}$ (Å)			
	A	C	G	T
σ_x	3.0809	2.6255	3.3773	2.889
σ_y	2.1573	1.5675	1.9315	1.698
σ_z	1.1348	0.9648	1.2677	1.071

Chapter 3

LAMMPS Modifications and Tests

3.1 About LAMMPS

The software package used for all coarse-grained simulations is the C++ program "Large-scale Atomic/Molecular Massively Parallel Simulator", referred to by its abbreviation LAMMPS [47]. It simulates molecules and atoms by integrating Newton's laws of motion using a finite timestep, a process known as "Molecular Dynamics" simulation, or MD. LAMMPS uses the Velocity-Verlet method (see [22, 48]) to integrate the equations of motion. LAMMPS is a well-maintained program which is easy to modify, being highly modular and open source. This is necessary for this project, as it involved the addition of new potentials and other modifications which were not previously a part of any available simulation package. LAMMPS is additionally highly optimized for parallel computation, however our work does not take advantage of this because optimizing the code to be parallelized is a more technical problem that would have detracted significantly from the scientific questions we wished to answer. Parallelizing our code would increase the efficiency of the MD simulations, but it must be left for future work.

3.2 Potentials Added to LAMMPS

Some of the potential function that we observed from the statistics of all-atom simulations, as discussed in Section 2.2, did not fit to any of the many potentials provided in the LAMMPS package. Thus it was required that we add new potentials to LAMMPS. The forces, or the derivatives of the potential, must be given analytically, since having the computer differentiate the potential numerically each timestep would add too greatly to the computational cost of the simulation. The first potential was the double well angle potential Equation 2.5. At each timestep, the simulation stores

the atomic coordinates of three atoms: $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$. The angle is calculated according to:

$$\theta = \arccos\left(\frac{\mathbf{r}_{12} \cdot \mathbf{r}_{32}}{|\mathbf{r}_{12}|\ |\mathbf{r}_{32}|}\right), \tag{3.1}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$. By taking the spatial derivatives of Equation 2.5, we obtain the forces on atoms $1, 2$ and 3. These forces were then coded directly into LAMMPS.

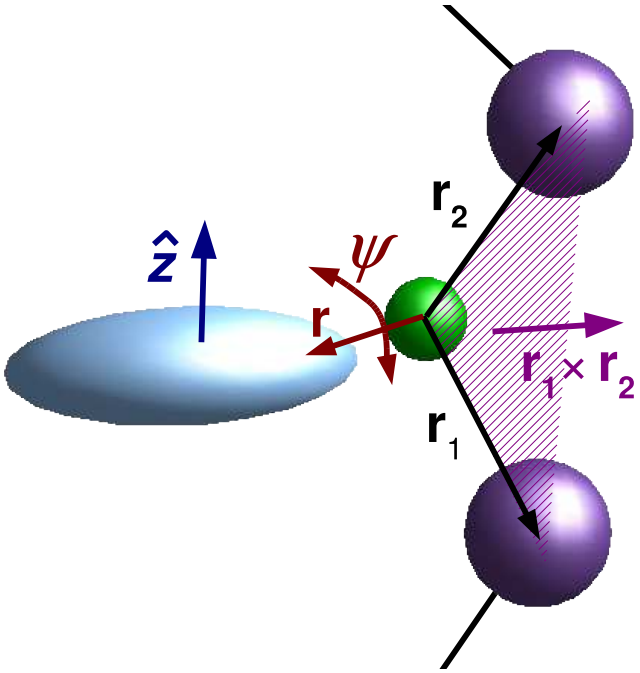

Figure 3.1: Depiction of the base orientation, ψ, with respect to the P-S-P plane of the backbone. The orientation potential was initially added to LAMMPS, but ultimately abandoned.

There were several other modified potentials that we added to LAMMPS, which we considered but ultimately decided not to use: the double well bond potential and the base orientation potential. We found that the second well in the bond potentials was not strong, and the bond was rarely in the second well, as seen in Figure 2.3. The base orientation potential induced a torque on the base given the relative orientation of the base's z principal axis and the normal of the P-S-P plane formed by the sugar bonded to the base and

the two phosphates next to the sugar. The base orientation, ψ, is shown in Figure 3.1. The angle ψ is calculated as:

$$\psi = \text{sign}\left([(\hat{\mathbf{r}}_1 \times \hat{\mathbf{r}}_2) \times \hat{\mathbf{z}}] \cdot \mathbf{r}\right) \arccos\left((\hat{\mathbf{r}}_1 \times \hat{\mathbf{r}}_2) \times \hat{\mathbf{z}}\right), \tag{3.2}$$

where sign$(\cdots)$ is ± 1 depending on the sign of the argument and the circumflex means the vector has been normalized. We fit the all-atom statistics to potentials like the dihedrals (see Equation 2.7). Namely, an orientation potential would be as follows:

$$V_{\text{orient}}(\psi) = \sum_{n=1}^{3} K_n(1 - \cos(n\psi - \psi_n)), \tag{3.3}$$

with six parameters $\{K_n, \psi_n\}_{n=1,2,3}$.

However, we came to realize that this potential was non-physical, because the bond that attaches the base to the sugar is a single bond and should therefore freely rotate. The effect that we observed in the all-atom potentials was not its own potential, but the result of the angle potentials and the pair potentials between the bases and phosphates. Also we found that these potentials disrupted the stacking of bases along the strand. Thus the base orientation potential was abandoned, although it was still implemented and tested.

3.3 Rigid Base-Sugar System

The all-atom simulations we conducted showed that the base and sugar formed a nearly rigid system. The spring constant of the bond between them is much higher than the soft bonds between the sugars and phosphates, shown in Table 2.1 and Figure 2.3. Additionally, the position of the sugar with respect to the ellipsoid axes of the base is tightly focused, shown in Figure 3.2. It is a common strategy in molecular simulation to freeze out stiff degrees of freedom so that the system becomes more stable and a larger timestep may be used. Typically an iterative algorithm such as SHAKE [49] is used, where atoms are iteratively perturbed back into their frozen configuration when they deviate from it. However, it is more computationally efficient to treat the system as a rigid body instead. The forces on each of the two components of the rigid body were calculated separately and only put together when the positions and orientations of the coarse-grained atoms were updated. The torques and forces on the whole system may be determined by the forces on the base and sugar and the torques on the bases. At

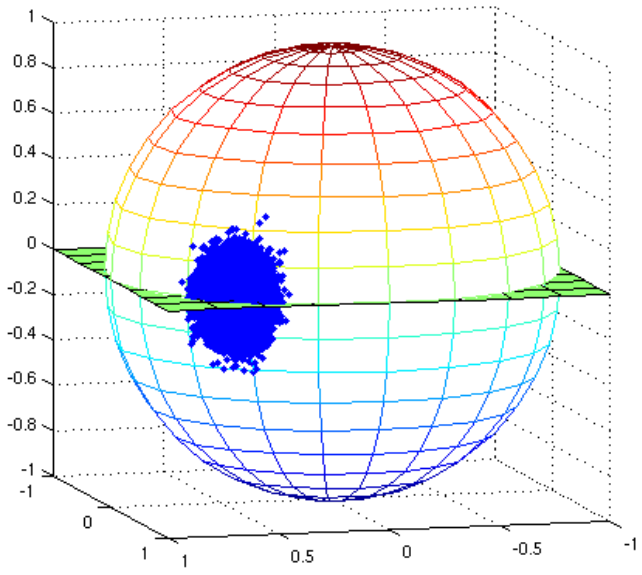

Figure 3.2: Distribution of the orientation of the sugar residue with respect to the principal axes of the base, taken from atomistic simulations. Also shown are the unit sphere and the $x - y$ plane of the base.

each timestep the positions of both the sugar and base and the orientation of the base are updated together as a rigid body, using the moment of inertia tensor of the rigid system. The position of the base is always updated so that the initial vector of the sugar position with respect to the principal axes of the base is held constant. This update to the LAMMPS code was made to the integration routine. Because it was a fundamental change to the way the integrator worked, it was hard-coded into the integration routine based on the way we designed the input files for the DNA coordinates: the sugar to which the base was bonded always had an index two lower than the index of the base. The sugars were precluded from their usual iterations of the position updates, and instead were treated during the position updates of their bases. While it is naturally preferable to create a general algorithm to treat rigid bodies, it would have taken significantly more time to implement. Since we were specifically working with DNA, we felt that hard-coding the algorithm into the routine was justifiable.

An instructive diagram of the rigid base system is shown in figure Figure 3.3. From this figure we see that the forces and torques on the rigid

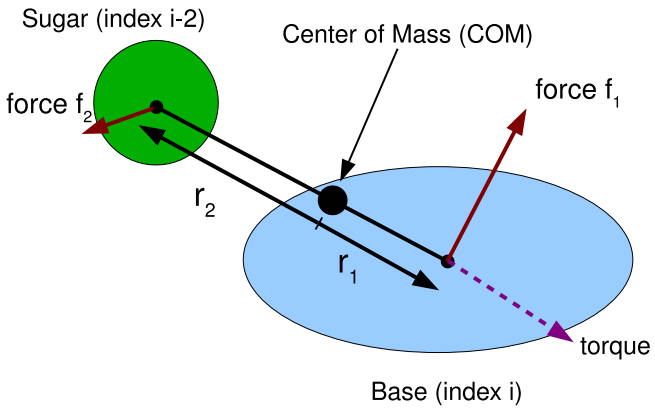

Figure 3.3: A diagram of the base-sugar system elucidating the method of treating forces and torques on the body that is described in Equations 3.4-3.8 and Program 3.1.

system are given by:

$$\mathbf{F} = \mathbf{f}_1 + \mathbf{f}_2 \tag{3.4}$$

$$\boldsymbol{\tau} = (\text{torque}) + \mathbf{r}_1 \times \mathbf{f}_1 + \mathbf{r}_2 \times \mathbf{f}_2 \tag{3.5}$$

The moment of inertia of a bare ellipsoid of radii a, b, c and mass m_i is given in the principal axis frame by:

$$I_0 = \frac{1}{5} m_i \begin{pmatrix} b^2+c^2 & 0 & 0 \\ 0 & a^2+c^2 & 0 \\ 0 & 0 & a^2+b^2 \end{pmatrix} \tag{3.6}$$

Since the center of mass is between the sugar and the base, we need to add an extra moment of inertia term for the base by the parallel axis theorem:

$$I_1 = m_i \begin{pmatrix} r_{1,y}^2 + r_{1,z}^2 & -r_{1,x}r_{1,y} & -r_{1,x}r_{1,z} \\ -r_{1,x}r_{1,y} & r_{1,x}^2 + r_{1,z}^2 & -r_{1,y}r_{1,z} \\ -r_{1,x}r_{1,z} & -r_{1,y}r_{1,z} & r_{1,x}^2 + r_{1,y}^2 \end{pmatrix} \tag{3.7}$$

Similarly we must add a term I_2 for the sugar of mass m_j which is given by Equation 3.7 making the substitutions $i \rightarrow j$ and $1 \rightarrow 2$ in the subscripts. The full moment of inertia is therefore:

$$I = I_0 + I_1 + I_2 \tag{3.8}$$

The algorithm to update the position and orientation of the rigid system is shown in Program 3.1. Note that it is hard coded to recognize the rigid system by the masses of the bases and that it assumes that the sugar has an index two less than the base. This algorithm is modelled after the position and orientation updates of the ellipsoids that is supplied by LAMMPS: a Velocity-Verlet method (see [22, 48]) whose half timestep update is given below.

1	$\mathbf{L} = \mathbf{L} + \boldsymbol{\tau} dt$	update angular momentum	
2	$\boldsymbol{\omega} = I^{-1}\mathbf{L}$	update angular velocity	
3	$\mathbf{q} = \mathbf{q} + d\mathbf{q}(\boldsymbol{\omega}, dt)$	update quaternion orientation	(3.9)
4	$\mathbf{v} = \mathbf{v} + \mathbf{F}dt/m$	update velocity	
5	$\mathbf{r} = \mathbf{r} + \mathbf{v}dt$	update position	

3.4 Thermal Rotations of the Bases

As noted in Section 2.3, an ellipsoid in a thermal bath will experience both translational and rotational diffusion, whose coefficients we determined by atomistic simulation. However, the Langevin temperature integration routine that LAMMPS has at the time of writing does not incorporate rotational degrees of freedom into thermal motions. Therefore we had to modify the Langevin integrator to implement Equations 2.30 and 2.31 in their entirety. Note that the thermal forces calculated for the rigid system of the base and sugar are calculated for each individually, as is the case for all other forces acting on the two objects. Since Equation 2.31 is dependent on the angular velocities, it is necessary to store the angular velocities so that the Langevin routine may read them. However, these are not variables that LAMMPS is presently set up to store outside of the primary integration routine. We needed to modify LAMMPS to store the angular velocities as part of the "atom" data which stores other data such as the positions, velocities and quaternion orientation each timestep. We stored the angular velocities for both the lab frame and the frame of reference of the base. Finally, we changed the way that LAMMPS calculated the temperature of the system, because now we included rotations as thermal motions. We modified the calculation of temperature using the equipartition theorem:

$$\frac{3}{2}Nk_BT + \frac{3}{2}N_bk_BT = \frac{1}{2}\sum_{i=1}^{N} m_i|\mathbf{v}_i|^2 + \frac{1}{2}\sum_{i\in\text{bases}} \langle\boldsymbol{\omega}_i|I_i|\boldsymbol{\omega}_i\rangle \tag{3.10}$$

Program 3.1 Part of LAMMPS code, written in C++, implementing the position/orientation update of the rigid base-sugar system

```
// hard coded to look for the mass of the bases
if (mass[type[i]] == 134.1  ||mass[type[i]] == 110.1  ||mass[type
    [i]] == 150.1  ||mass[type[i]] == 125.1 ){
  // --- CODE SUPPRESSED ... various initializations ... define
      r_1, r_2, center of mass COM, and torque1 and torque2 ...
  // Convert the torques to angular momenta: torque = dL/dt
  angmom_system[i][0] += dtf * (torque[i][0] + torque1[0] +
      torque2[0]);
  angmom_system[i][1] += dtf * (torque[i][1] + torque1[1] +
      torque2[1]);
  angmom_system[i][2] += dtf * (torque[i][2] + torque1[2] +
      torque2[2]);
  // Get the angular momentum of the system in the body frame
  MathExtra::transpose_times_column3(rot, angmom_system[i],
      angmom_body);
  // Convert the angular momentum to an angular velocity by
      multiplying by the inverse of the inertia tensor
  MathExtra::times_column3(inertia_inv[type[i]], angmom_body,
      w_body);
  // Convert the angular velocity back to an angular momentum.
      This time it is the angular momentum of the base only.
  w_body[0] = w_body[0]*inertia[type[i]][0]; w_body[1] = w_body
      [1]*inertia[type[i]][1]; w_body[2] = w_body[2]*inertia[type
      [i]][2];
  // --- CODE SUPPRESSED ... Update orientation by calling '
      richardson' function. Define r1p, r2p ... These are the new
       r_1 and r_2 vectors AFTER the rotation. Update COM
      velocity Using integration dv/dt = F/m. Update COM with dr
      = v*dt.
  // Remember the old positions before this function was called
  x1_0[0] = x[i][0]; x1_0[1] = x[i][1]; x1_0[2] = x[i][2];
  x2_0[0] = x[i-2][0]; x2_0[1] = x[i-2][1]; x2_0[2] = x[i-2][2];
  // Update the positions by adding r1p and r2p to the center of
       mass vector
  x[i][0] = COM[0] + r1p[0]; x[i][1] = COM[1] + r1p[1]; x[i][2]
      = COM[2] + r1p[2];
  x[i-2][0] = COM[0] + r2p[0]; x[i-2][1] = COM[1] + r2p[1]; x[i
      -2][2] = COM[2] + r2p[2];
  // Determine the velocities using v = dx/dt
  v[i][0] = (x[i][0] - x1_0[0] )/dtv; v[i][1] = (x[i][1] - x1_0
      [1] )/dtv; v[i][2] = (x[i][2] - x1_0[2] )/dtv;
  v[i-2][0] = ( x[i-2][0] - x2_0[0] ) / dtv; v[i-2][1] = ( x[i
      -2][1] - x2_0[1] ) / dtv; v[i-2][2] = ( x[i-2][2] - x2_0[2]
       ) / dtv;
}
```

where N is the number of atoms, k_B the Boltzmann constant, T the temperature, N_b the number of bases (ellipsoids), m_i the mass of coarse-grained atom i, $\mathbf{v}_i$ its velocity, I_i its moment of inertia tensor and $\boldsymbol{\omega}_i$ its angular velocity.

Program 3.2 shows part of the code which implements the Langevin thermostat. This is a fairly typical example of the kinds of modifications made to the LAMMPS simulation package. As with most of the LAMMPS additions we made involving ellipsoids, this code required conversions between the lab and ellipsoid's reference frames. In this case, Equations 2.30 and 2.31 are diagonal only in the principal axis frame of the ellipsoid. Updates to the positions and orientations are done in the lab frame, because one needs to express the coordinates of all-atoms in a common frame of reference, but calculating the thermal forces and torques on the base is done in the ellipsoid's frame. LAMMPS proved a number of linear algebra methods in the "MathExtra" package which was used to transform between reference frames.

3.5 Hydrogen Bonds

LAMMPS does not presently feature any potential which could serve to model hydrogen bonds between two ellipsoids. Like the RE^2 potential, it must be both orientation and distance dependent, except that it must be much more strongly directionally dependent, because for a hydrogen bond to form, the two bases must be facing each other. The form of the potential that we decided to use is given by Equation 2.27. The position-dependent part of the equation within the square brackets is commonly used as the distance dependent portion of the potential, e.g. [32, 50]. Because the model is coarse-grained, the distance over which pairwise potentials act is larger than one would find in an atomistic simulation. We required some way to insure that a base would not bond to several bases along the chain at once. We accomplished this by making it strongly orientation dependent, which is why the cosine is raised to the fourth power. The multiplier of 3 for the angles, as discussed in Section 2.2.9, ensures that the potential goes to zero exactly at the angle cutoff $\pi/6$. It is additionally convenient, as it reduces the computation time via the triple angle formula:

$$\cos(3\theta) = 4\cos^3\theta - 3\cos\theta, \tag{3.11}$$

Program 3.2 Part of LAMMPS code, written in C++, implementing Langevin thermostat (Equations 2.30 and 2.31)

```
double omega_body[3]; // Angular velocity in body frame
double rot[3][3]; // The orientation in matrix representation
// Converts the quaternion orientation 'q' into matrix
    representation 'rot'
MathExtra::quat_to_mat(q,rot);
// rot transpose multiplied by the angular velocity in the lab
// frame 'angmom' gives the angular momentum in the lab frame '
    omega_body'
MathExtra::transpose_times_column3(rot,angmom,omega_body);
//Convert the angular momentum in the body frame into an angular
     velocity by dividing by the moments of inertia.
omega_body[0] /= inertia[type][0];
omega_body[1] /= inertia[type][1];
omega_body[2] /= inertia[type][2];
// Get velocity in body frame, similar to angular momentum
double v_body[3];
MathExtra::transpose_times_column3(rot,v,v_body);
// Define torques and forces in lab and body frames
double torque_body[3], torque_lab[3], force_body[3], force_lab
    [3];
// Torques and forces are given by a drag term plus a random
    thermal term
torque_body[0] =gconst[0]*xiRotation[type][0]*omega_body[0] +
    gconst[1]*sqrt(xiRotation[type][0])*(random->uniform()-0.5);
torque_body[1] =gconst[0]*xiRotation[type][1]*omega_body[1] +
    gconst[1]*sqrt(xiRotation[type][1])*(random->uniform()-0.5);
torque_body[2] =gconst[0]*xiRotation[type][2]*omega_body[2] +
    gconst[1]*sqrt(xiRotation[type][2])*(random->uniform()-0.5);

force_body[0] = gconst[0]*xiTranslation[type][0]*v_body[0] +
    gconst[1]*sqrt(xiTranslation[type][0])*(random->uniform()
    -0.5);
force_body[1] = gconst[0]*xiTranslation[type][1]*v_body[1] +
    gconst[1]*sqrt(xiTranslation[type][1])*(random->uniform()
    -0.5);
force_body[2] = gconst[0]*xiTranslation[type][2]*v_body[2] +
    gconst[1]*sqrt(xiTranslation[type][2])*(random->uniform()
    -0.5);
//convert torques and forces back to lab frame
MathExtra::times_column3(rot,torque_body,torque_lab);
MathExtra::times_column3(rot,force_body,force_lab);
```

which expresses $\cos(3\theta)$ algebraically in terms of $\cos\theta$. This way, we can apply the chain rule when differentiating the potential to obtain the forces:

$$\frac{\partial \cos(3\theta)}{\partial \mathbf{r}} = \frac{d\cos(3\theta)}{d(\cos\theta)}\frac{\partial \cos\theta}{\partial \mathbf{r}}, \tag{3.12}$$

The following discussion is regarding the θ_1 term in Figure 2.13, but applies equally to θ_2 by symmetry. We see that $\cos(\theta_1) = \frac{-\mathbf{r}\cdot\mathbf{p}_1}{|\mathbf{r}||\mathbf{p}_1|}$. To obtain the forces and torques, we need to differentiate with respect to $\mathbf{r}_a$ and $\mathbf{r}_b$. We can now differentiate Equation 3.11 with respect to $\mathbf{r}_b$ to get:

$$\begin{aligned}
\frac{\partial \cos(3\theta_1)}{\partial \mathbf{r}_b} &= \left(\frac{-\mathbf{r}\ \mathbf{r}\cdot\mathbf{p}_1}{|\mathbf{r}|^3|\mathbf{p}_1|} + \frac{\mathbf{p}_1}{|\mathbf{r}||\mathbf{p}_1|}\right)(12\cos^2\theta_1 - 3) && (3.13)\\
&= \left(\cos\theta_1\frac{\mathbf{r}}{r^2} + \frac{\mathbf{p}_1}{|\mathbf{r}||\mathbf{p}_1|}\right)(12\cos^2\theta_1 - 3). && (3.14)
\end{aligned}$$

A similar calculation for $\mathbf{r}_a$ gives:

$$\begin{aligned}
\frac{\partial \cos(3\theta_1)}{\partial \mathbf{r}_a} &= \left(-\cos\theta_1\frac{\mathbf{r}}{r^2} - \cos\theta_1\frac{\mathbf{p}_1}{|\mathbf{p}_1|^2} - \frac{\mathbf{r}+\mathbf{p}_1}{|\mathbf{r}||\mathbf{p}_1|}\right)\\
&\quad\times(12\cos^2\theta_1 - 3). && (3.15)
\end{aligned}$$

Now we can define the forces induced on points a and b:

$$\begin{aligned}
\mathbf{f}_b &= -\frac{\partial V_{HB}^{(i)}}{\partial \mathbf{r}_b} && (3.16)\\
&= -\left[\frac{\partial V_{HB}^{(i)}}{\partial \mathbf{r}}\frac{\partial \mathbf{r}}{\partial \mathbf{r}_b}\right] - \left[\frac{\partial V_{HB}^{(i)}}{\partial(\cos\theta_1)}\frac{\partial\cos\theta_1}{\partial \mathbf{r}}\right] && (3.17)\\
&= \epsilon\Bigg\{\left[-60\left((r_o/r)^{12} - ((r_o/r))^{10}\right)\frac{\mathbf{r}}{r}\cos^4(3\theta_1)\right]\\
&\quad - \left[4\left(5(r_o/r)^{12} - 6((r_o/r))^{10}\right)\frac{\partial\cos(3\theta_1)}{\partial \mathbf{r}_b}\right]\Bigg\}\cos^4(3\phi_1) && (3.18)
\end{aligned}$$

using the chain rule several times. The $\mathbf{f}_a$ term can be calculated by substituting $\mathbf{r}_a$ for $\mathbf{r}_b$. The $\cos^4(3\phi_1)$ is not considered to induce a force on any of the three points, but is added purely as a torsional term to break a rotational symmetry of base i about the p_1 axis. The forces on bases i and j from the $V_{HB}^{(i)}$ term in the hydrogen bond is thus:

$$\mathbf{F}_i = -\mathbf{f}_b \qquad \mathbf{F}_j = \mathbf{f}_b \tag{3.19}$$

The torques are now straightforward to calculate as well:

$$\begin{aligned} \boldsymbol{\tau}_i &= \mathbf{p}_1 \times \mathbf{f}_a - \frac{\partial V_{HB}^{(i)}}{\partial \phi_1} \frac{\mathbf{p}_1}{|\mathbf{p}_1|} \\ \boldsymbol{\tau}_j &= \mathbf{p}_2 \times \mathbf{f}_b + \frac{\partial V_{HB}^{(i)}}{\partial \phi_1} \frac{\mathbf{p}_1}{|\mathbf{p}_1|} \end{aligned} \tag{3.20}$$

We implemented Equations 3.19-3.20 in LAMMPS to use for the hydrogen bond potential. For each hydrogen bond, these equations must be computed twice, once for the $i \rightarrow j$ direction, $V_{HB}^{(i)}$, and once for the reverse direction, $V_{HB}^{(j)}$. The implementation of the equations themselves are given in Program 3.3. This code forms part of the hydrogen bond function added to the RE^2 force calculation routine. Since the RE^2 routine already deals heavily with the features of the ellipsoids like quaternions, it was most efficient to add the hydrogen bond calculations in with the RE^2 calculations. Note that Program 3.3 does not include parts of the hydrogen bond code, such as the angle cutoffs, or the enforcing of Watson-Crick base pairing, the latter of which is accomplished by not calling the hydrogen bond function for pairs whose well depth is zero.

3.6 Testing LAMMPS Implementations

Since we made a number of changes to the LAMMPS code, we needed to ensure that those modifications worked correctly. We prepared a number of test simulations which could be compared to analytical solutions. We took our double well potentials and set up three atoms with a double well angle potential between them. By starting the system off slightly perturbed from its potential minimum, the harmonic approximation remained valid. Comparing the observed oscillation frequencies at zero temperature to those found analytically using the harmonic approximation, we found that they matched to within at least five significant figures. We did this for each of the two minima of the double well.

We also ran simulations with the three atoms starting at a point on the potential above the barrier height, so that the system would oscillate between both wells at zero temperature. Each timestep, we output the atomic coordinates of all three atoms, the energy as calculated by our implementation of the integrator, and the forces on all-atoms. Using this data we compared the energies and forces calculated by LAMMPS with a script written to calculate the forces and energies independently. Since outputs matched

Program 3.3 Part of LAMMPS code, written in C++, implementing the hydrogen bonds Equations 3.19-3.20

```
//Triple angle formula
cos3theta = 4*costheta3 - 3*costheta;
// Calculate powers of cos(3theta): cos3thetaM = cos(3theta)^M
cos3theta2 = cos3theta*cos3theta; cos3theta3 = cos3theta*
    cos3theta2; cos3theta4 = cos3theta2*cos3theta2;
// Various pieces of the calculation
fconst = cosphi4*E0a*cos3theta4*60*(rpr12 - rpr10)/rsq;
vconst = cosphi4*E0a*(5*rpr12 - 6*rpr10);
tauconst = 12*vconst*cos3theta3*(4*costheta2 - 1);
// Calculate f_a and f_b
f_b[0] = -fconst*r12[0] - tauconst*( costheta*r12[0]/rsq + pihat
    [0]/r_mag);
f_b[1] = -fconst*r12[1] - tauconst*( costheta*r12[1]/rsq + pihat
    [1]/r_mag);
f_b[2] = -fconst*r12[2] - tauconst*( costheta*r12[2]/rsq + pihat
    [2]/r_mag);
f_a[0] = fconst*r12[0] + tauconst*( costheta*(r12[0]/rsq + pihat
    [0]/pi_mag) + (r12[0] + pi[0])/(r_mag*pi_mag) );
f_a[1] = fconst*r12[1] + tauconst*( costheta*(r12[1]/rsq + pihat
    [1]/pi_mag) + (r12[1] + pi[1])/(r_mag*pi_mag) );
f_a[2] = fconst*r12[2] + tauconst*( costheta*(r12[2]/rsq + pihat
    [2]/pi_mag) + (r12[2] + pi[2])/(r_mag*pi_mag) );
// Add the forces found from the hydrogen bond
fj[0] += f_b[0]; fj[1] += f_b[1]; fj[2] += f_b[2];
fi[0] -= f_b[0]; fi[1] -= f_b[1]; fi[2] -= f_b[2];
// Update the van der Waals force on particle i
f_vdwl[0] -= f_b[0];
f_vdwl[1] -= f_b[1];
f_vdwl[2] -= f_b[2];
// The torque on base i and j from the theta term
MathExtra::cross3(pi, f_a, torque_i); MathExtra::cross3(pj, f_b,
     torque_j);
// update torque on bases i and j
tori[0] += torque_i[0]; tori[1] += torque_i[1]; tori[2] +=
    torque_i[2];
torj[0] += torque_j[0]; torj[1] += torque_j[1]; torj[2] +=
    torque_j[2];
// Torque from the phi term
E_r_theta = vconst*cos3theta4;
phi_torque_final = phi_torque*E_r_theta/cosphi4;
// Calculate torque from the phi angle
tori[0] += phi_torque_final*pi[0]; tori[1] += phi_torque_final*
    pi[1]; tori[2] += phi_torque_final*pi[2];
torj[0] -= phi_torque_final*pi[0]; torj[1] -= phi_torque_final*
    pi[1]; torj[2] -= phi_torque_final*pi[2];
```

to within four decimal places, we were satisfied that our implementation of the double well angle potentials was correctly integrated.

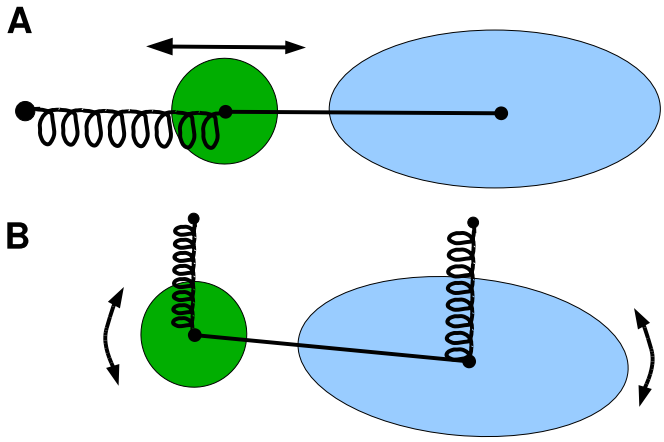

Figure 3.4: The setup to test the implementation of the base-sugar system. Figure A shows a test of forces on the system. Similar tests were also conducted with the force applied to the ellipsoid and with the colinear spring and base-sugar bond in each of the x, y and z directions. Figure B shows a test of torques induced on the system. For symmetry the springs are identical and the masses are identical so that the springs apply forces equidistant from the center of mass.

Treating the base and sugar as a rigid system was a fundamental change to the way the integrator algorithm functioned. It was tested in a similar way to the double well potentials: by making small oscillations about a minimum and comparing output energies with analytical values. A spring was attached to the sugar and the system was allowed to oscillate symmetrically, shown in Figure 3.4A. So that the system is simpler for analysis, the sugar and spring were placed on the base's x axis. The oscillation frequency was observed to agree with $\omega = \sqrt{k/m}$ to within four decimal places. In Figure 3.4B, the masses of the sugar and base are made equal. The correct frequency is determined by:

$$\text{torque} \approx -2RkR\theta \tag{3.21}$$

$$I\ddot{\theta} \approx -2kR^2\theta \tag{3.22}$$

$$\omega \approx \sqrt{\frac{2kR^2}{I}}, \tag{3.23}$$

for small oscillations θ. The spring constant is k and R is the distance from the center of mass to either of the springs. I is the moment of inertia of the rigid system. The analytic frequency was confirmed by simulation to four significant figures, confirming the correctness of our implementation of the rigid base-sugar system.

The first test of the Langevin integrator was to check that the temperature was being calculated correctly. To do this we took a system of 1000 ellipsoids held at various temperatures and found the mean temperature reported by LAMMPS. We found that the target temperature and the mean temperature agreed to within four significant figures, subject to fluctuations about the mean over time. Finally, to determine that the rotational diffusion was correctly implemented, we wrote a Matlab script to check that the rotational diffusion coefficients agreed with the expected value from the Stokes-Einstein relation:

$$\frac{\langle \theta^2_{x,y,z} \rangle}{2t} = \frac{kT}{\zeta_{x,y,z}}. \tag{3.24}$$

Here $\theta_{x,y,z}$ is the total rotation of the ellipsoid about each of its principal axes, x, y and z, over a time t. The rotation is necessarily expressed in the frame of reference of the ellipsoid. The average $\langle \cdots \rangle$ is an average over many independent trials. Since the mean squared of the angular diffusion from Equation 3.24 increases linearly with time, θ is not determined modulo 2π. The rotational diffusion was determined in Matlab, by adding up the rotations from each timestep. The rotations were expressed in (axis, angle) form where the first three entries of the 4-vector form a unit vector giving the axis of rotation and the angle is the rotation angle about that axis, expressed in the range $-\pi \leq \text{angle} \leq \pi$. The product of the axis of rotation and the angle then gives a vector $\boldsymbol{\theta}$ whose components are θ_x, θ_y and θ_z. Again the simulations were found to agree with the theory satisfactorily to about three significant figures.

Chapter 4

Definitions, Methods and Relevant Details

4.1 Temperature Scale

We express temperature in units of a natural energy scale ϵ in the system. We construct this energy scale by taking the minima of the RE^2 potential between identical bases given as V_{min} in table 2.5, and averaging this value over all bases A, C, G, T, and orientations x, y, z (an average of the 12 values appearing in table 2.5). This results in a value of $\epsilon = 1.45$ kcal/mol.

4.2 Persistence Length

The persistence length of a polymer is a measure of its rigidity, and is a useful physical property to test in a computational simulation. Mathematically, it is the decay constant for correlations in the tangent vector to the polymer over its length. If we start at a position s_o along the polymer and measure the tangent vector $\hat{\mathbf{t}}(s)$, where s is the arc length along the polymer, then tangent correlations decay as follows:

$$\langle \hat{\mathbf{t}}(s_o) \cdot \hat{\mathbf{t}}(s) \rangle = e^{-(s-s_o)/l_p}, \tag{4.1}$$

where l_p is defined as the persistence length of the polymer and $\langle \cdots \rangle$ is a thermal average. As $l_p \to 0$ then the tangent vector decorrelates very quickly, meaning that there are many random bends in the strand over a short length scale comparable to l_p. As $l_p \to \infty$ then the correlations go to 1, which is the case for a perfectly straight polymer.

For DNA, there is some ambiguity of how to define the tangent vector $\hat{\mathbf{t}}(s)$ and likewise the arc length s. For single stranded DNA, we defined the tangent vector as being between phosphate residues five base pairs apart, and the arc length as the sum of the distances between neighbouring phosphate residues spanning s_0 to s. For double stranded DNA, a similar method was used, except that we took tangent vectors ten base pairs apart, and instead

of using the phosphate positions, we used the midpoint of the phosphate on one strand and the corresponding phosphate on the other strand.

There are two possible length scale difficulties that we need to avoid: the persistence length becoming less than one phosphate-phosphate distance, and the persistence length becoming much longer than the total length of the strand. The first type of length scale conflict did not occur, as the DNA cannot bend that quickly. The second type proved difficult to avoid in the double stranded DNA, because we were only able to simulate a strand of length 60 base pairs for the time scales necessary to get a reliable thermal average. We were careful to use strands long enough that the decay of tangent vector correlations was statistically significant. A thorough discussion of persistence length and the difficulties inherent in computing the quantity is found in [44].

Computing the thermal average is somewhat ambiguous as well, since we are measuring s continuously in units of Å. If at one timestep $s - s_o = 9.88632$Å, it is unlikely we will reproduce that exact measurement again to average over. There are three obvious ways around this problem, and we used all three to ensure consistency. The first is to use binned values for $s - s_o$; the second is to measure the distance in discrete units of base pairs; and the third is to plot all the distances as data points, then perform an exponential fit to the very noisy data instead of taking an average. We feel that the third method is the most accurate, because it best captures the idea that internal distances in DNA can expand or contract, since the distance between base pairs does not remain constant. It was also the most straightforward method to implement. During testing, however, we compared all three methods and found that they gave equivalent answers to about 0.2Å, which is within the error ranges for the estimates of l_p. The first 2 methods additionally have the advantage that they take the average properly, confirming that the assumption of an exponential decay is valid. The method used to calculate the persistence is shown explicitly in Program 4.1.

4.3 Radius of Gyration

The radius of gyration, R_G, is a measure of the compactness of a molecule and is defined by:

$$R_G = \sqrt{\frac{\sum_{i=1}^{N} m_i(\mathbf{r}_i - \mathbf{r}_{\mathrm{CM}})^2}{\sum_{i=1}^{N} m_i}} \tag{4.2}$$

Program 4.1 Matlab script to obtain the persistence length of double stranded DNA. Uses the Matlab tool available from the LAMMPS website (*http://lammps.sandia.gov/doc/Section_tools.html*)

```
% Alex Morriss-Andrews
% Wed Apr  8 17:34:08 PDT 2009

% --- CODE SUPPRESSED. Various initializations and definitions.
% Stores tangent vector correlations and arclengths
correlation = []; arclen = [];
% Run through starting positions s_0 along strand and times t
for start=(start0):(start0+num),
  for time=t(1:100:(length(t)-1)),
    % Read the information at that timestep
    S = readdump_one('quat.dump',time,9);
    data = S.atom_data;
    r = data(data(:,2)== 6 ,3:5);
    % Get position of average between base pairs
    v = [];
    for i=1:bp,
      v = [v;( r(i,:) + r(i+bp,:) )/2];
    end
    dv0 = v(start+5,:) - v(start,:);
    arcleni = norm(dv0); % Arclength between two pairs
    % Run through (s-s_0)
    for i=(start+10):5:(stop),
      dvi = v(i,:) - v(i-5,:);
      % Get correlation and arclength for this (s-s_0) and t
      correlation = [correlation, dot(dv0, dvi)/(norm(dv0)*norm(
          dvi))];
      arclen = [arclen, arcleni];
      arcleni = arcleni + norm(dvi);
    end
  end
end
% Apply some smoothing to the noisy data
NL = length(arclen);
sorted = sortrows([arclen', correlation'],1);
scor = fnval(1:NL, csaps(1:NL, sorted(:,2),0.2))';
slen = sorted(:,1);
% Perform exponential fit on smoothed data to obtain persistence
    length and confidence interval
[c,coef] = fit_plot(slen,scor, 'exp(-x/d)', [100])
pm = range(confint(c))/2;
% --- CODE SUPPRESSED. Output results and plots.
```

where the sums are over all particles in the molecule, $\mathbf{r}_i$ is the position of the i^{th} particle and $\mathbf{r}_{\text{CM}}$ is the molecule's center of mass.

4.4 Helical Twist

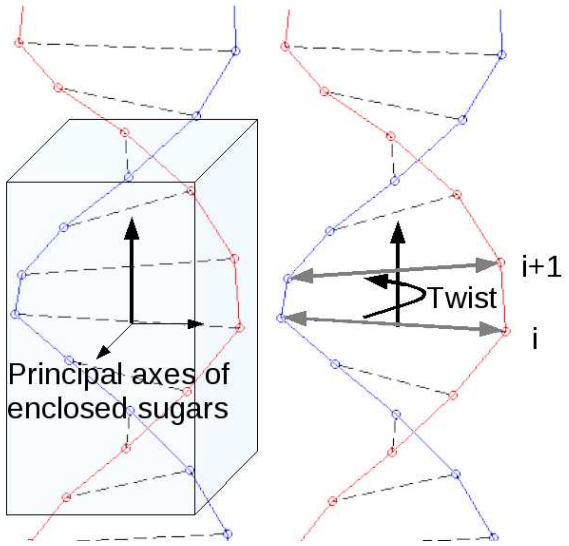

Figure 4.1: Calculation of local helical axis and twist

We define the helical twist of a piece of double stranded DNA as the average angle that the backbone of the double helix rotates about the helical axis for each successive base pair, depicted on the right of Figure 4.1. To obtain the helical axis at a given position along the double helix, we take the positions of the two sugars opposite the hydrogen-bonded bases at that position, as well as the sugar pairs up to three bases above or below that position. Using the sugar coordinates, we compute the principal moments of inertia, and take the moment of least rotational inertia to be the helical axis (as shown in the left of Figure 4.1). While this method would fail if the double helix persistence length decreases to the order of three base pairs or if the helix dehybridizes, neither of these scenarios occurs with the parameters that we use to simulate a stable double-helix. The twist is computed as an average along the strand, as well as a time average over the simulation. The pitch of the DNA is similarly defined where pitch $= 2\pi/\text{twist}$, or the number of base pairs one must traverse for a full revolution of the helix. Note that

the observed crystal structure value for the pitch of B-DNA is 10 [3], giving an expected twist of 36°.

4.5 Base Stacking

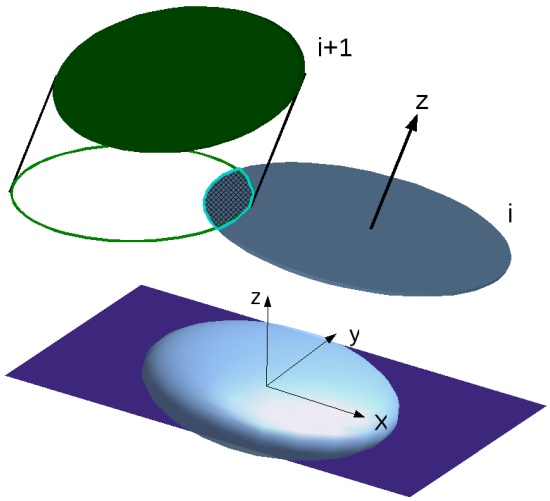

Figure 4.2: Calculation of the stacking fraction for a pair of bases

We have also developed an order parameter defining how well a given base pair is stacked. Taking the dot product of the z principal axes of the ellipsoids has an undesirable translational symmetry in the $x - y$ plane of either ellipsoid, and so does not fully capture the concept of stacking. The method we employ uses area projections, depicted in Figure 4.2. For each ellipsoid, we take the cross-section of the ellipsoid in its principal axis $x - y$ plane. To calculate the stacking fraction between bases i and $i + 1$, we project this elliptical cross-section $i + 1$ down onto cross-section i. We then take the average of this value with the equivalent projection of i onto $i + 1$ in order that our definition of stacking be symmetric. We normalize the projected area by the area that would be obtained from the B-DNA crystal structure, where the typical overlap between bases is about 60%. They do not overlap 100% in the "ideal" case because of the twist in the crystal structure. Therefore one observes stacking fractions greater than 1.

This definition of stacking mathematically represents the intuitive notion of stacking as the degree to which the areas of the objects' flat parts over-

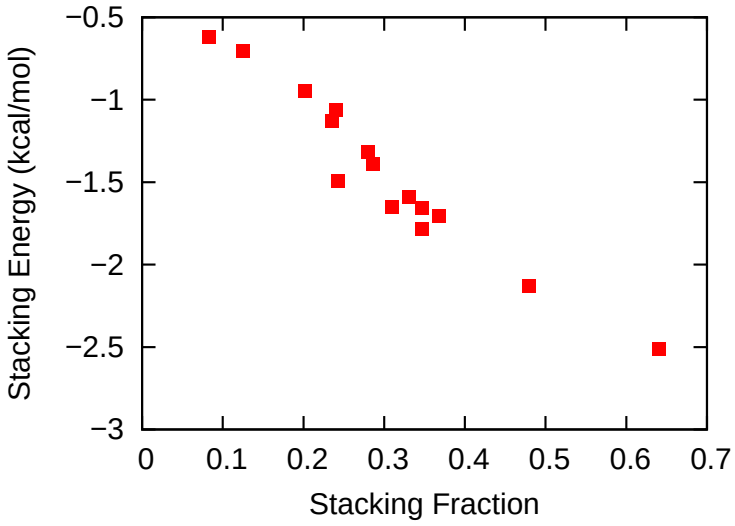

Figure 4.3: Stacking fraction plotted as a function of nearest neighbour stacking energies. Results obtained from a number of simulations run at temperatures varying from 0 to 0.34ϵ. These results were taken from the same simulations as the random sequence from Figure 5.9.

lap. The stacking fraction is a geometrical order parameter, but correlates strongly with the RE^2 energies between neighbouring bases on either of the two strands of the DNA ($r = -0.9690$, see figure 4.3). Thus our stacking fraction accurately captures the base-base van der Waals stacking energies in addition to quantifying a structural feature of DNA.

4.6 Contour Length

Similar to the persistence length we take the midpoints of the two phosphate sites on opposite strands at a position i along the strand. The vector between two neighbouring midpoints is the unnormalized tangent vector. Taking the dot product of this with the local helical axis (defined in Figure 4.1) gives the contour length per base. The total contour length is the sum of the contour lengths per base.

4.7 Definition of Sequences Used

Sequences used in our simulations will be referred to with the following convention. We will use C_N-G_N to mean a strand of poly(C) of length N which is hydrogen bonded to its complementary poly(G) strand of the same length. A_N-T_N is similarly defined. We denote by HET several heterogeneous sequences, defined in the $5' \rightarrow 3'$ direction as:

$$\begin{aligned} HET_{SS,25} &= CAGGATTAATGGCGCCTACCTTACC \\ HET_{SS,30} &= CATCCTCGACAATCGGAACCAGGAAGCGCC \\ HET_{SS,60} &= HET_{SS,30} - CCGCAACTCTGCCGCGATCGGTGTTCGCCT\ . \end{aligned}$$

Finally, $HET_{DS,N}$ is a strand of $HET_{SS,N}$ that is hydrogen bonded to its complementary strand.

Chapter 5

Results

5.1 Persistence Length and Radius of Gyration

5.1.1 Varying Ionic Concentration

As we increase the ionic concentration of the system, we increase the degree of Coulombic screening that occurs between the phosphates. Since the phosphates tend to repel each other, they cause the DNA to stiffen and therefore increase the persistence length. Therefore increasing ionic concentration will weaken the phosphate repulsion and decrease the persistence length of the strand. Simulation results for the persistence length versus ionic concentration are shown in the red triangles of Figure 5.1A for single stranded DNA and in the data points of Figure 5.1B for double stranded DNA. The functional dependence on concentration is captured by a constant bare persistence length added to a term inversely proportional to concentration [44], shown in the dotted lines in Figure 5.1. The single stranded DNA has a much lower persistence length since it lacks the stabilizing structure of the double helix. Its persistence length is on the order 1 nm, which corresponds to 2-3 bases. We find that the distance between successive base pairs is approximately 0.4 nm at typical cellular ionic concentrations of 0.04 mol/L. The single stranded DNA was simulated at a temperature of $310K$, where stacking and hydrogen bonds are the same order as k_BT. This temperature is above the dehybridization temperature in our model, and also above the collapse temperature for ssDNA. We find that a single stranded DNA strand collapses at a higher temperature than the double stranded DNA dehybridizes. Therefore, to obtain physical results, we must simulate the single stranded DNA at temperatures of about 0.4ϵ for the ssDNA and about 0.1ϵ for the dsDNA.

The values for the persistence length found in our simulations are generally less than those observed experimentally, but it is worth noting that the experimental measurements themselves are highly variable. This can be understood from the high degree of sensitivity of the persistence length to other factors, such as base sequence [25] and the experimental set-up used to

measure the persistence length, for example fluorescence spectroscopy [42], laser tweezers [52], hairpin loops [33] and gel electrophoresis [54].

The double helix is inherently more stable than a single strand, due to the added stability of two backbones wound around each other. As can be seen from Figure 5.1B, the persistence length of the double strand has the same functional dependence on concentration as ssDNA, but is roughly 55 times stiffer at $0.02M$, 25 times stiffer at $0.04M$ and 52 times stiffer at $0.13M$. The ratio found experimentally for the persistence lengths is 66 [52].

Plots of the radius of gyration versus ionic concentration are shown in Figure 5.2. This gives a general idea of how the DNA collapses when the strength of the Coulomb repulsion is weakened.

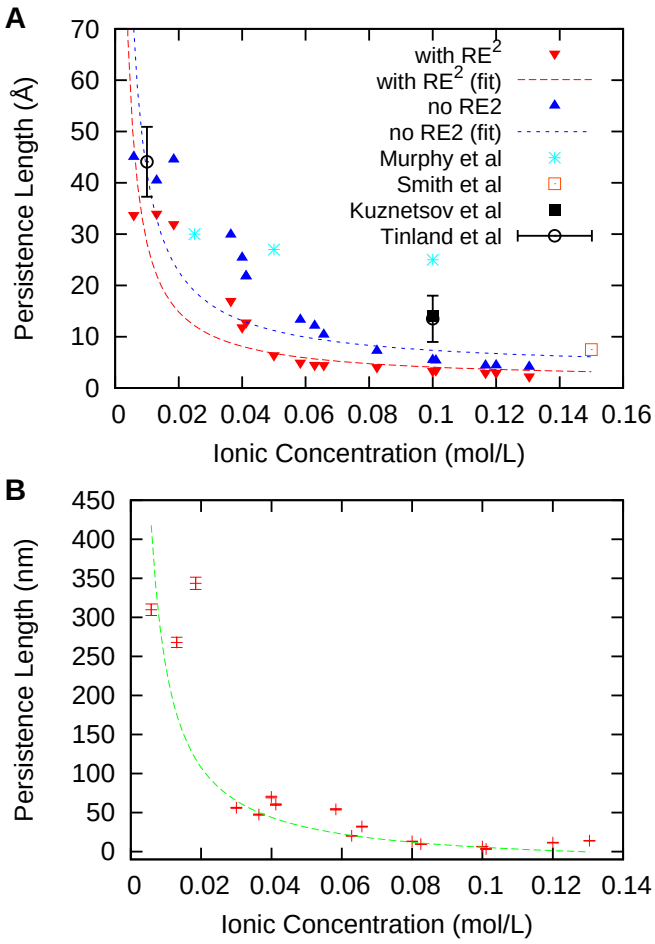

Figure 5.1: Persistence length as a function of concentration of ions, for simulations of A) ssDNA, B) dsDNA. Also plotted as dashed lines is the theoretical model of Nguyen et. al. [44] for the persistence length of a polyelectrolyte, whose functional form consists of an intrinsic persistence length plus a term inversely proportional to ionic concentration. A) Persistence length of a random 25bp sequence ($HET_{SS,25}$) at a temperature of 0.42ϵ, simulated for 300 ns. Also shown are simulation results for ssDNA with the RE^2 potential turned off (blue triangles). The remaining data sets are from various experiments: Murphy *et al* [42], Smith *et al* [52], Kuznetsov *et al* [33] and Tinland *et al* [54]. B) Persistence length vs. ionic concentration for 60bp dsDNA (sequence $HET_{DS,60}$) at a temperature of 0.10ϵ, simulated for 600 ns.

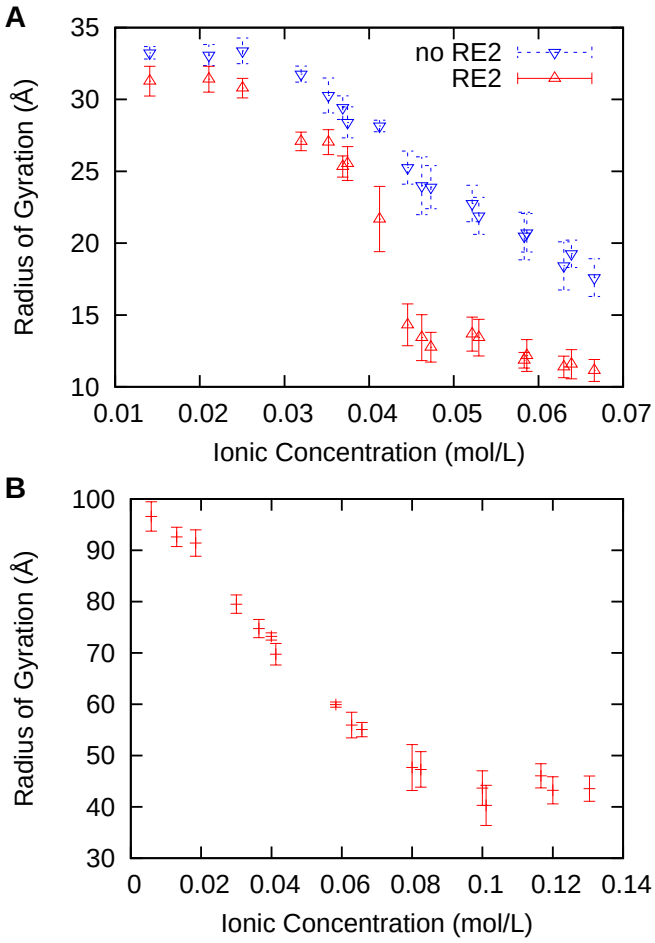

Figure 5.2: Dependence of the radius of gyration on ionic concentration. Figure A shows the results for a single strand of DNA and Figure B for a double helix. The results are taken from the same simulations as shown in Figure 5.1 A and B respectively.

5.1.2 Varying RE^2 Energy

Figure 5.1A also shows that turning off the RE^2 potential, i.e. eliminating the base-base stacking interactions, increases the stiffness of the single-stranded DNA. This occurs because the RE^2 potential is an attractive interaction, which tends to collapse the strand. According to Figure 5.3A, our model predicts a larger persistence length for a homogeneous single-strand of adenine (as large as ~ 50 bases or more) than the corresponding homogeneous strand of thymine bases ($l_p \approx 2$ bases) at low temperatures. These results are consistent with the conclusions of Goddard *et al.* [25], who found larger enthalpic costs for hairpin formation in poly(A) than in poly(T). However, at high temperatures the situation is reversed, and ss-poly(A) has a *smaller* l_p (≈ 1.5 bases) than ss-poly(T) ($l_p \approx 4$ bases), see Figure 5.3B. Adenine, being a purine, has a stronger RE^2 stacking interaction (see the z-minima in Table 2.5), however all $A-A$, $A-P$ interactions are generally stronger than $T-T$, $T-P$ interactions, and at high temperature this induces a greater degree of collapse due to non-local self interactions of DNA strand. The persistence length shows an increasing trend with the length of the strand, at the temperatures and ionic concentrations that we studied (Figures 5.3 A and B). This is due to the exaggeration of end effects on shorter strands. The persistence length converged to its infinite length value for stands longer than about $7l_p$ at high temperature.

Additionally, simulations were run where the strength of the RE^2 interaction was scaled by a global factor, which is seen in Figure 5.4. Here it is evident that the attractive interaction is causing the collapse of the helix. The DNA undergoes a collapse at a value very close to our parametrized strengths of the DNA, which makes the ssDNA highly sensitive to collapse.

5.1.3 Varying Temperature

The persistence length and the radius of gyration are also functions of temperature, seen in Figure 5.5. At low temperatures, we expect the polymer to collapse as favoured by the attractive pair interactions. As we raise the temperature, however, a transition occurs where the DNA extends as a result of thermal motions. We find that our collapse temperature is 0.41ϵ.

5.2 Helical Twist and Base Stacking

Here the twist and stacking properties of double stranded DNA are treated together, because their dependence on salt concentration, RE^2 energy and

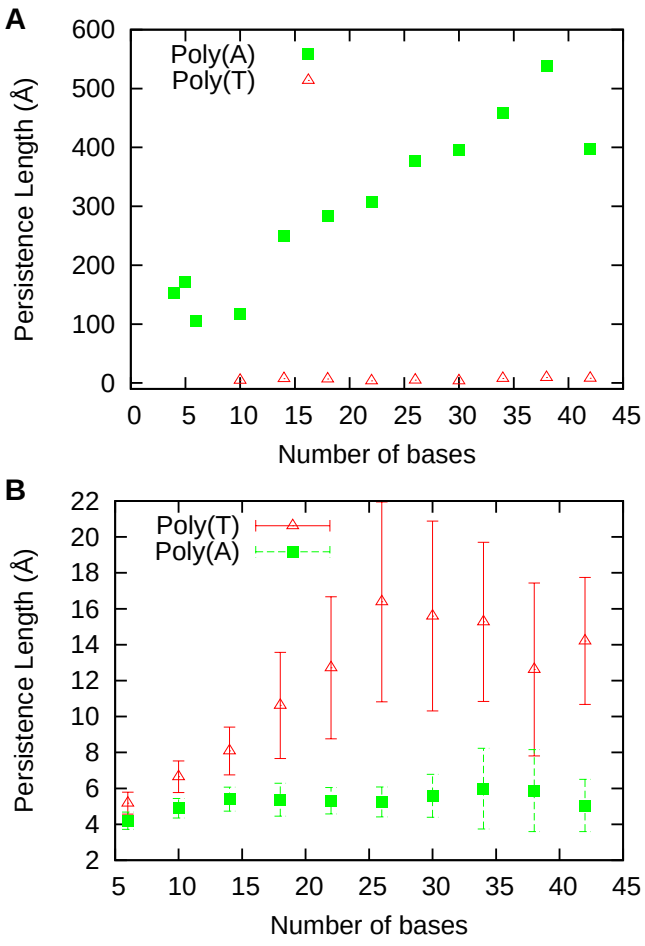

Figure 5.3: Persistence length of ssDNA for poly(A) and poly(T) as a function of the number of bases in the polymer chain. Simulations were conducted at temperatures of (A) 0.03ϵ and (B) 0.42ϵ at ionic concentration of 40 mM for 240 ns. At these temperatures and ionic concentrations the distance between stacked bases is about 4Å. Note that the temperature in (B) is above the hybridization temperature for dsDNA.

temperature is qualitatively very similar. These two DNA properties quantify how similar the model dsDNA is to its PDB structure. Recall from Chapter 4 that the expected twist is 36° and the expected stacking fraction is defined to be 1.

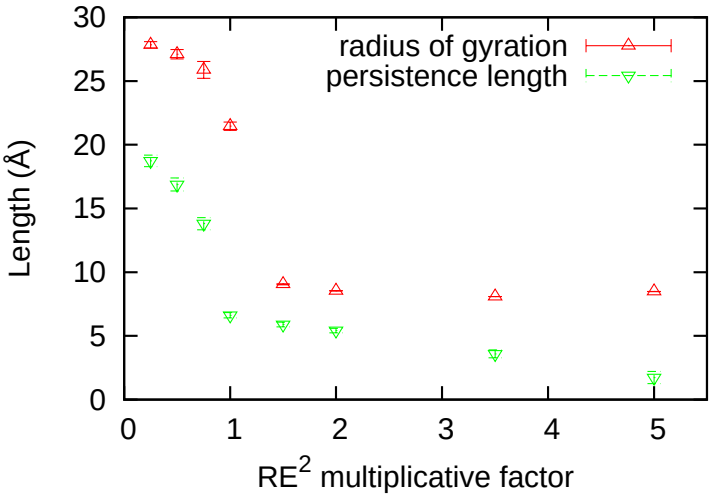

Figure 5.4: Single stranded DNA, with the same sequence as in Fig. 5.1A. Shows the behaviour of the two lengths scales, radius of gyration and persistence length, as a function of the strength of the RE^2 potential. Simulations were conducted for 240 ns. An ionic concentration of 0.04 mol/L is used and the temperature is 0.42ϵ.

5.2.1 Stacking Along Homogeneous Strands

The behaviour of base stacking as a function of position on the strand is nontrivial. Traces of periodicity occur with respect to position, over approximately four to five bases (see Figure 5.6). Bases appear to stack well in small groups, at the expense of poorer stacking seen in bases situated nearby along the strand. This kind of phenomenon is indicative of kinks in the stacking structure of the strand, where the distance between kinks is only a few base pairs. An interesting trend is that purines, the larger bases with greater stacking interactions, fluctuate far less and show more consistent stacking. Conversely, pyrimidines exhibit greater extremes in their decimated stacking pattern, intermittently stacking more strongly than purines, and completely unstacking. The stronger stacking interactions of the purines apparently induce convergence to the average value.

The decimation pattern that we observe for the bases is due to frustration between stacking interactions and the other potentials. As Figure 5.7 shows, reducing the potential results in less sudden oscillations in the stack-

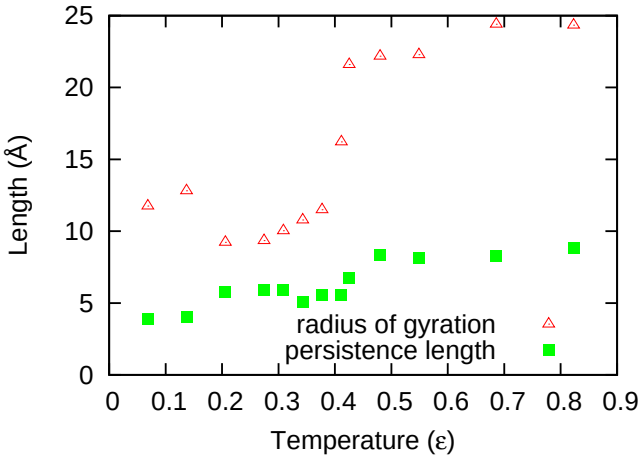

Figure 5.5: Single stranded DNA, with the same sequence as in Fig. 5.1A. Shows the behaviour of the two lengths scales, radius of gyration and persistence length, as a function of temperature. Simulations were conducted for 240 ns. An ionic concentration of 0.04 mol/L is used.

ing fraction and results in a much weaker amplitude in the existing oscillations. This situation is reminiscent of a Frenkel-Kontorova model, where competition between two incommensurate length scales, corresponding to the equilibrium separation of a 1D chain of oscillators and a periodic underlying potential, results in frustration-induced domains of oscillators [5].

5.2.2 Varying Ionic Concentration

Figure 5.8 shows the twist and stacking of the $HET_{DS,60}$ sequence as a function of ionic concentration. Increasing ionic concentration reduces the repulsive Coulomb energy by decreasing the Debye length in Equation 2.8.

As the repulsive Coulomb force becomes more screened, the RE^2 potential will begin to win out over the Coulomb potential, and this will have the effect of compressing the double helix by reducing the mean distance between neighbouring phosphates along the strand as in Figure 5.9. In order to prevent the other potentials in the model from becoming too frustrated by this compression, the helix must twist to compensate. Thus we expect an increase in concentration to increase helical twist in an indirect way. Indeed this is the general trend observed in Figure 5.8A. Since the repulsion of the phosphates does not directly drive helix formation, there occur fluctuations

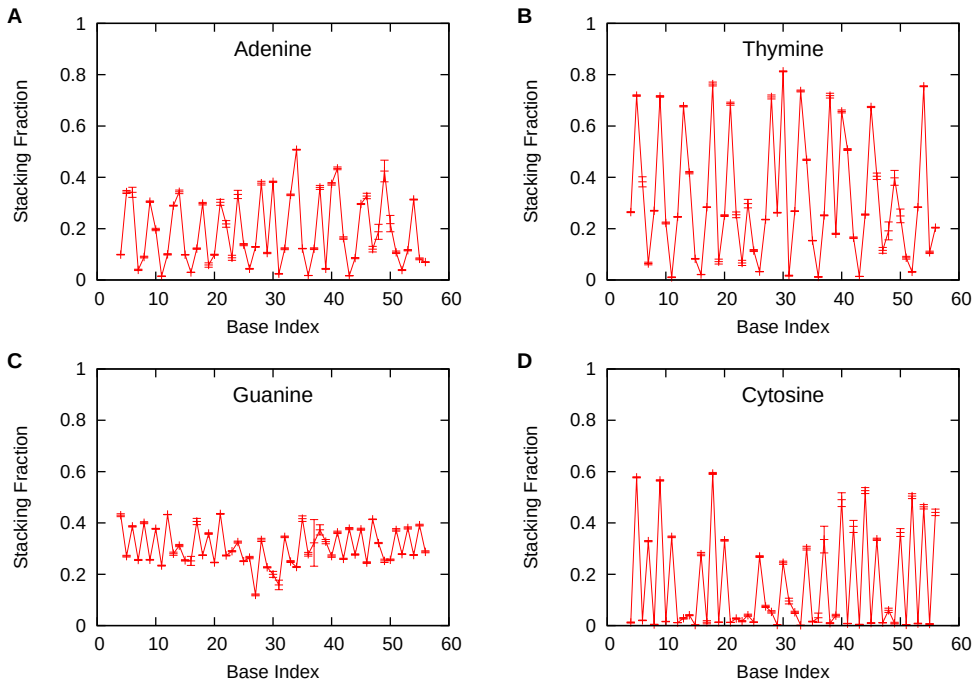

Figure 5.6: Time averaged stacking fractions of homogeneous strands in a helix with their complementary strand. Figs. A and B are obtained from simulations of 60 base dsDNA, consisting of a homogeneous strand of poly(A) hydrogen bonded to a similar strand of poly(T). The figures show the stacking of the adenine bases with themselves and the thymine bases with themselves respectively. Figs. C and D are obtained similarly to A and B except for guanine and cytosine bases. All simulations were conducted at $T = 0.07\epsilon$ and an ionic concentration of 0.04 mol/L. The total simulation time is 800ns. The average stacking fraction for each of these simulations is 0.1889 ± 0.0185 for adenine, 0.3178 ± 0.0343 for thymine, 0.3065 ± 0.0099 for guanine and 0.1593 ± 0.0271 for cytosine.

due to the complexities of competing interactions. It is also interesting that the effect appears to saturate as we raise the concentration high enough. This is intuitive, since increasing the concentration weakens the Coulomb interaction. It would not make sense if a hypothetical DNA, without any residual charges on its backbone, was infinitely tightly wound. Thus it is necessary that the helicity saturate at some point.

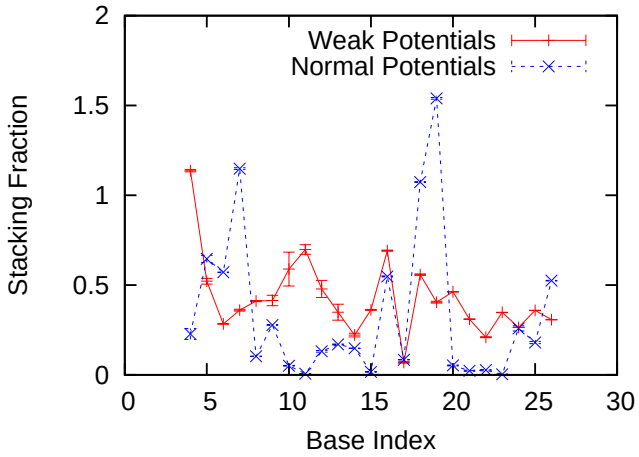

Figure 5.7: This figure shows the time-averaged stacking fraction of the bases along a strand for both the all-atom parametrized potentials and another simulation with weakened parameters. Here the bond potentials were weakened by a factor of 0.5, the angle potentials by 0.1 and the dihedral potentials were disabled. These simulations were done with $HET_{DS,30}$ at an ionic concentration of 0.04 mol/L and a temperature of 0.07ϵ.

In Figure 5.8B we can see how well a double helix stacks in varying concentrations of ionic solution. As the concentration of ionic solution serves to screen the Coulombic repulsion of the phosphate residues, the DNA will tend to compress in higher concentrations due to the fact that the repulsive forces give a rigidity to the helix. While it may not be immediately obvious how this affects stacking between bases, many of the properties of DNA result from competition between the attractive RE^2 potential and the repulsive Coulomb force. As we increase ionic concentration, the RE^2 force begins to win out, so we might expect that in this limit stacking will be highest. Additionally, as the helix compresses, there will be less room for the bases to move about, which would again favour stacking. We can see the simulations of the DNA at finite temperature in Figure 5.8B clearly obey the trend of increased concentration leading to increased stacking. The fact that zero temperature stacking decreases past roughly 0.08M can be rationalized by observing that at this point, with no temperature to help extend the DNA, the backbone begins to crumple and some structure is lost. For this reason, several kinks are seen to form along the backbone which hinders

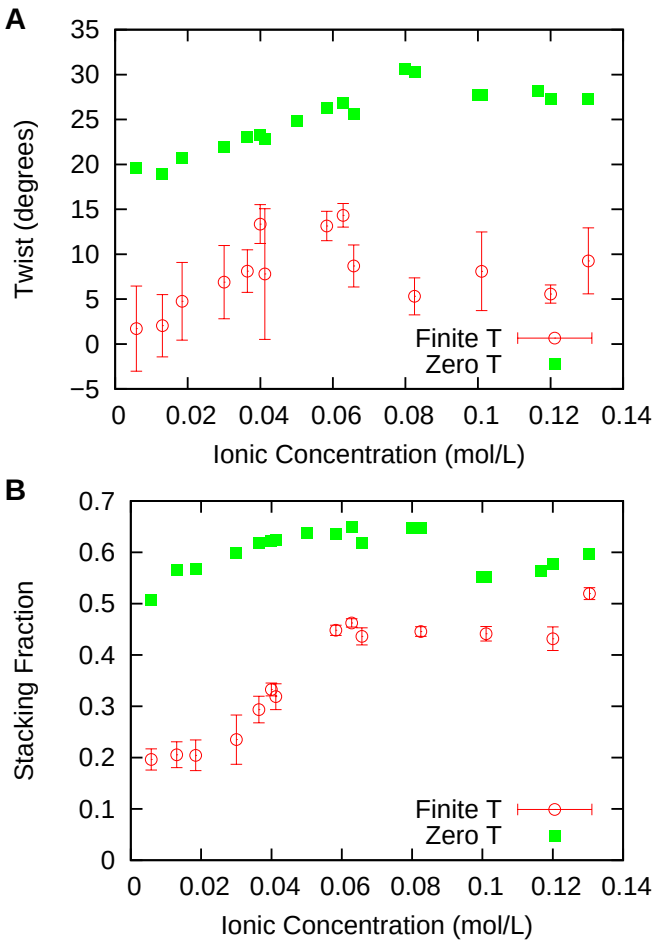

Figure 5.8: Twist and stacking of double stranded DNA vs. ionic concentration. Finite temperature data is taken from a 600ns simulation of $HET_{DS,60}$ at $T = 0.10\epsilon$, and thermal averages are shown. Zero temperature data is taken from an 80ns simulation of $HET_{DS,60}$, and the final equilibrated state is shown.

base stacking.

5.2.3 Varying RE^2 Energy

Surprisingly we find that the base-base attraction also appears to induce

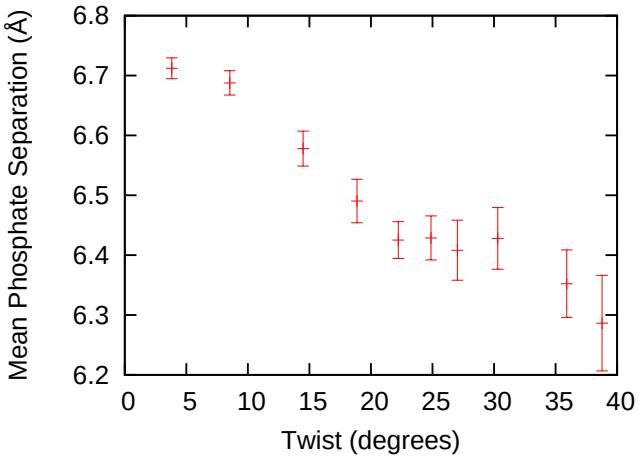

Figure 5.9: The mean distance between phosphate groups as a function of helical twist. for a $HET_{DS,30}$ sequence of at zero temperature and 40mM ionic concentration. Simulations were started from the PDB structure [3] and allowed to equilibrate by energy minimization to structures having the values shown at $T = 0$. The equilibration time was 80ns, implying shallow energetic gradients of collective modes.

twist of the double helix. A picture that emerged from computational models of DNA structure [40] is that twist results from the competition between electrostatic repulsion of phosphate groups and favourable base stacking interactions. Stronger stacking interactions would favour alignment of bases and tend to straighten the helix, and indeed the RE^2 potential is minimized when the relative twist between stacked bases is zero. Figure 5.10A plots helical twist as a function of RE^2 interaction strength which modulates stacking interactions. This clearly shows a monotonic increase in twist with stacking interaction strength. These results were obtained by a zero temperature energy minimization, starting from the expected crystal structure of B-DNA. During the simulations, helical twist relaxed to a degree determined by the overall strength of the RE^2 potentials. At the all-atom parametrized values the twist is 21.2°, whereas in the crystal structure it is 36.0°.

The ladder conformation is preferred when the RE^2 potential is scaled down to zero, where Coulombic interactions are competing solely with local bond, angle, and dihedral potentials. As seen in Figure 5.9, the tightly

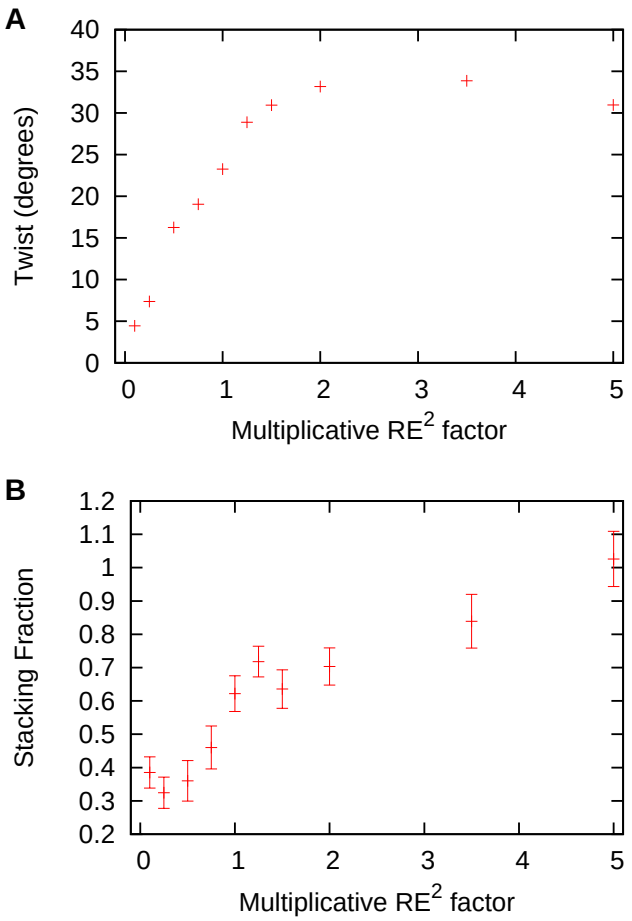

Figure 5.10: The effect of scaling the RE^2 energies on the (a) twist and (b) stacking of double stranded DNA, from the same simulations as Figure 5.9.

twisted conformations of the helix bring the phosphate residues closer together. The RE^2 potential serves to compress the DNA as seen from Figure 5.12. The direct compression of a ladder configuration would increase the Coulomb energy in inverse proportion to the contour length, as well as frustrate local potentials such as angular potentials (see Figure 5.11 and description below). To avoid this energetic cost as DNA is compressed, the system can lower its energy by structurally relaxing into a helical conforma-

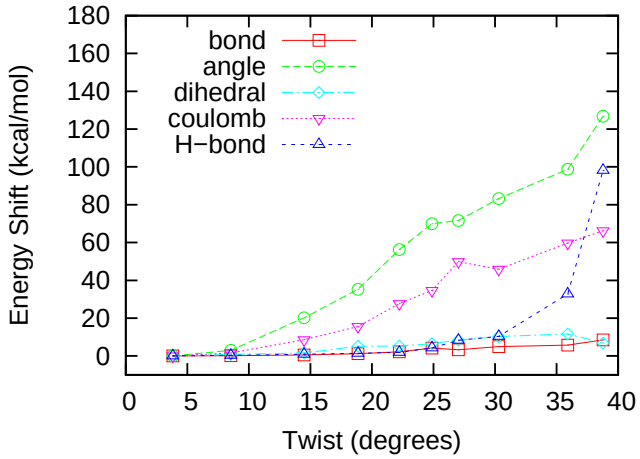

Figure 5.11: Behaviour of energy scales as functions of helical twist, as induced by increasing stacking interactions. Data plotted here are taken from the same simulations as in Figure 5.10. Energies have been shifted to start from zero to depict the relative effects of increasing twist more clearly.

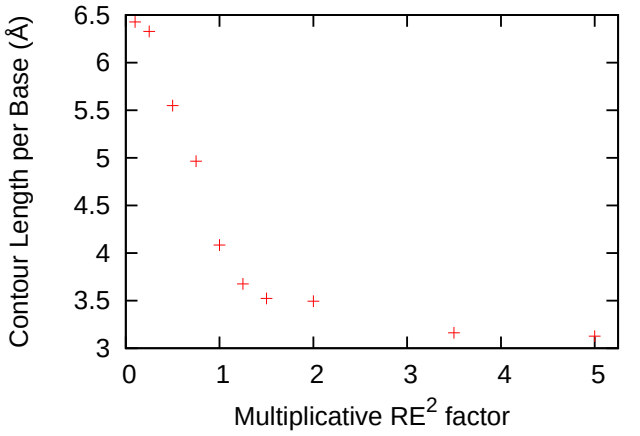

Figure 5.12: The effect of scaling the RE^2 energies on the contour length of double stranded DNA. Simulation parameters are the same as in Fig. 5.1B.

tion.

As expected, the dominant interaction governing the stacking fraction is the strength of the base-base RE^2 potential. There was significant relaxation of the PDB structure at the all-atom parametrization values. The twist and stacking fraction obtain their PDB values only when base-base interactions are magnified by a factor of 5, as can be seen from Figure 5.10. It is worth noting that large dynamical fluctuations in the DNA structure are also observed in all-atom simulations [36].

5.2.4 Varying Temperature

Increasing temperature tends to reduce the twist of DNA (see Figure 5.13A). The sequence G_{60}-C_{60} maintains a twist of approximately 29° until $T = 0.05\epsilon$ at which point it drops sharply, suggesting a phase transition. Large fluctuations manifest themselves by a sudden increase in the statistical error at $T = 0.05\epsilon$. The helicity then fluctuates about zero for higher temperatures. The stacking fraction for G_{60}-C_{60} shows similar behaviour to the twist (Fig. 5.13B), with a phase transition around $T = 0.05\epsilon$. This transition involves a reduction of order within the double-stranded structure, and is distinct from dehybridization. The $HET_{DS,60}$ double strand was observed to dehybridize at a temperature around 0.31ϵ. It took approximately 330 ns to dehybridize completely. Above the transition temperature, stacking is still appreciable, while the twist does not appear to be significant. We found that A_{60}-T_{60} maintains very little twist, even at zero temperature, where it has a weak twist of 12°. This sequence likewise showed the least stacking order among the sequences we studied. The sequence $HET_{DS,60}$ displays much smoother behaviour with changing temperature. At zero temperature, it has a twist of 26°, slightly less than homogeneous G-C DNA, and its stacking fraction is intermediate to G-C and A-T. The phase transition behaviour present in G_{60}-C_{60} is smoothed out for the random sequence, yet curiously maintains helicity until a higher temperature. Its stacking is intermediate to G-C and A-T sequences at all temperatures, and also has a broadened transition.

5.3 Helix Chirality

The chirality of the helix in our model comes from the dihedral potentials which are not symmetric about an angle of zero, and therefore break chiral symmetry. We find the following potentials bias toward a right-handed helix as found for B-DNA (strength of bias given in brackets in units of

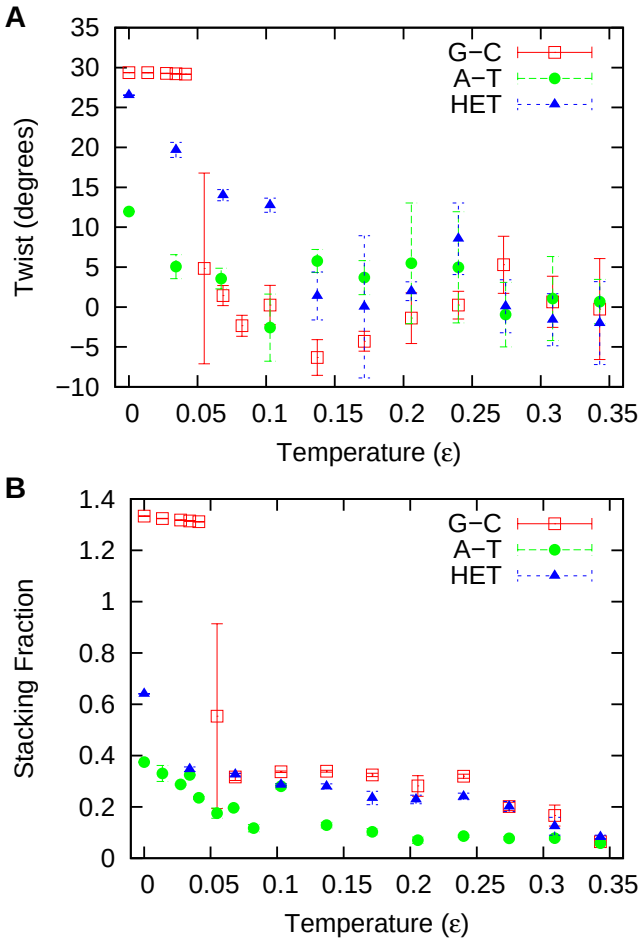

Figure 5.13: Behaviour of twist and stacking versus temperature. This is a simulation of sequences G_{60}-C_{60}, A_{60}-T_{60}, and $HET_{DS,60}$ that are simulated at an ionic concentration of 0.04 mol/L for 800 ns.

kcal/mol): $S_{5'}P_{5'}SP_{3'}$ (0.25), $P_{5'}SP_{3'}S_{3'}$ (0.06), $S_{5'}P_{5'}SA$ (0.40), $S_{5'}P_{5'}SC$ (0.18), $S_{5'}P_{5'}SG$ (0.62). The potential $S_{5'}P_{5'}ST$ biases toward a left-handed helix, but at a strength of only 0.03 kcal/mol. The remaining potentials are chiral symmetric. In our simulations, the local minimum of the right-handed helix is favoured over the left-handed helix by 57.7 kcal/mol.

5.4 Major and Minor Grooves

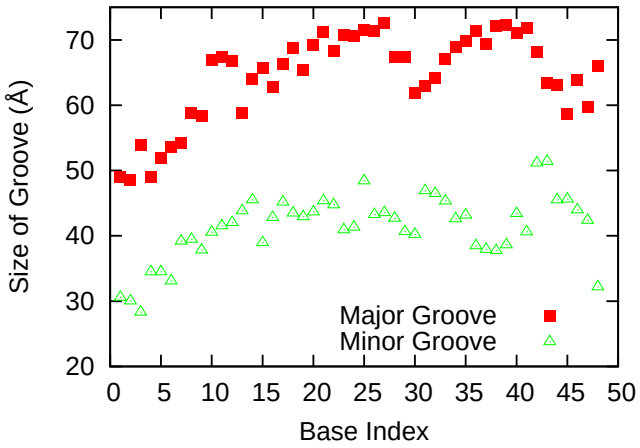

Figure 5.14: The size of the major and minor grooves as a function of position along the double helix. This is a finite temperature simulation of ionic concentration 0.04 mol/L as in Figure 5.8. The errors are of the order of the symbol size. The double helix makes about two turns over the length of the strand.

Our simulations show a major and minor groove in double stranded DNA. This effect is observed at zero and finite temperature, provided that the twist of the DNA is at least 10°. A typical example is plotted in Figure 5.14.

Chapter 6

Conclusion and Future Work

We have developed a coarse-grained model of DNA where we represent the bases as ellipsoids, which are better representations of the steric shape of bases than traditional spherically symmetric coarse-graining. Our parametrization procedure used a first-principals approach where we determined our local potentials through statistics gained from atomistic simulations (for the bond, angle and dihedral potentials), analytical forms (for the Yukawa Potential) and fits to all-atom potentials (for the base-base interactions).

We have simulated both single and double stranded DNA molecules of various lengths and analyzed various properties of the model, and determined how these properties varied as functions of salt concentration, base-base RE^2 interactions and temperature. This analysis gave us insight into the effects of the parametrization and external conditions on our coarse-grained DNA model.

We found that our model gave realistic persistence lengths as a function of ionic concentration, within the noise of the experimental data, and was reasonably well modelled by the functional form of [44] at concentrations of about 0.04 mol/L and higher.

Our results show that the properties of stacking and twist in a double helix are primarily a competition between the RE^2 energy and the repulsive Coulomb interaction. The Coulomb forces try to push the DNA into an unwound helix and the attractive RE^2 interaction tries to compress it. We believe that the compression of the helix is strongly disfavoured by the other potentials, which forces the DNA into a helix. The chiral symmetry of the helix is broken by the dihedral potential, which we determined to bias the dsDNA to a right-handed helix, as seen in B-DNA. This bias was not deliberately put into our model, but came out of all-atom statistics for a piece of DNA too small to form a helix itself. We also find that the helicity and stacking of the DNA are highly sequence dependent: phase transitions occurred upon changing the temperature in a G_{60}-C_{60} sequence, but not in a A_{60}-T_{60} sequence.

Finally, our model also clearly shows the existence of major and minor grooves, as seen in real DNA.

Since the base-base interaction so strongly affects the properties of DNA, it would be fruitful to continue to search for coarse-grained models which accurately represent the steric effects of the bases.

We have refrained from tweaking the parameters to optimize the results of our simulations, as is typically done, favouring a systematic approach. Our model, in large part due to the complex form of the base-base RE^2 potential, has a large number of parameters that could be tweaked which made it very difficult to optimize in such a high dimensional parameter space. Scaling a number of the potentials and analyzing the results provides hints on how to find a more optimal parameter set. However, this optimization is still extremely involved and thus requires future research.

There are a number of virtual experiments to which one could subject our model that have not been done. It would be interesting to force a strand of DNA through a nanopore, as discussed in Section 1.3. Such an experiment would give insight into dehybridization in our model as well as clarify the sequence dependence of the current through the pore, which is the heart of the experiment. Our model, with its bases represented as ellipsoids, would be ideal for this experiment because it the steric shape of the base determines the amount of current it can block. An ellipsoid much better represents the steric shape of the base than a spherical coarse-grained model does. Although not having the resolution of an all-atom simulation, our coarse-grained model is much more computationally inexpensive, enabling the simulation of much longer strands passing through the pore.

For longer strands, the persistence length of the DNA becomes a relevant parameter governing the resistance of the DNA to passing through the pore. The biggest difficulty in preparing this experiment with our model is to make the nanopore itself. One would need to determine a potential function for the walls of the pore and how they interact with the DNA and ions passing through it. The use of implicit water in a crowded region such as the nanopore may also be a poor approximation.

Another experiment, popular with protein researchers (e.g. [8]), is to determine the force extension curves by stretching the ends of the molecule with, for example, an atomic force microscope. By finding the nonequilibrium properties of the force extension, one can determine equilibrium free energy differences between collapsed and extended states using the Jarzynski relation [30]. This experiment is especially amenable to computational simulation, since all it requires is to attach two new potentials to the ends of the molecule forcing them apart.

It would also be useful to study hairpin formation in single stranded DNA, where a hairpin structure forms by the forming of Watson-Crick pairs

involving the strand with itself. This would likewise be simple to implement, but would require a very long simulation time in order to get a reasonable free energy surface. One would hope to see at least two states in the hairpin: an open state and closed state. Reasonable order parameters to use for the free energy surface would be the total hydrogen bond energy or the root mean square displacement from a "perfect" hairpin structure. It also would be interesting to analyze the sequence dependence of hairpin free energy surfaces. In particular, a homogeneous sequence such as $A_N T_N$ could have nearly N adenines in the first half of the strand bonded to nearly N thymines in the second half. There would be a low energetic cost for the hairpin to "slip" by having one half of the strand creep up the other half. The minimal slip cost would be the hydrogen bond energy of one A-T pair. In such a case, one might expect less cooperativity in the hairpin formation and a smoother free energy surface.

Bibliography

[1] Aleksij Aksimentiev, Jiunn B. Heng, Gregory Timp, and Klaus Schulten. Microscopic kinetics of dna translocation through synthetic nanopores. *Biophysical Journal*, 87:2086–2097, 2004.

[2] BJ Alder and TE Wainwright. Phase transition for a hard sphere system. *The Journal of Chemical Physics*, 27:1208, 1957.

[3] S. Arnott, P. J. Campbell-Smith, and R. Chandrasekaran. Atomic coordinates and molecular conformations for dna-dna, rna-rna, and dna-rna helices. In G. D. Fasman, editor, *Handbook of Biochemistry and Molecular Biology.*, volume 2, pages 411–422. CRC Press, Cleveland, 3 edition, 1976.

[4] Neil W. Ashcroft and David N. Mermin. *Solid State Physics*. Thomson Learning, Toronto, January 1976.

[5] S. Aubry. Solitons and condensed matter physics ed. ar bishop and t schneider. *Springer Series in Solid State Sciences*, 8:246–277, 1978.

[6] M. Babadi, R. Everaers, and M. R. Ejtehadi. Coarse-grained interaction potentials for anisotropic molecules. *J. Chem. Phys.*, 124(17):174708, 2006.

[7] R. Berardi, C. Fava, and C. Zannoni. A gay-berne potential for dissimilar biaxial particles. *Chem. Phys. Lett.*, 297:8–14, 1998.

[8] C. Bouchiat, MD Wang, J.F. Allemand, T. Strick, SM Block, and V. Croquette. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophysical journal*, 76(1):409–413, 1999.

[9] B. R. Brooks, R. E.Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[10] D.A. Case, T.A. Darden, T.E. Cheatham, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, M. Crowley, R.C. Walker, W. Zhang, et al. Amber 10. *University of California, San Francisco*, 2008.

[11] C.J. Cramer. *Essentials of computational chemistry: theories and models*. Wiley, 2004.

[12] W.L. DeLano. *The PyMOL Molecular Graphics System*. DeLano Scientific, Palo Alto, CA, 2002.

[13] K. Drukker and G.C. Schatz. A model for simulating dynamics of DNA denaturation. *J. Phys. Chem. B*, 104(26):6108–6111, 2000.

[14] B. Dünweg and W. Paul. Brownian dynamics simulation without gaussian random numbers. *Int. J. Mod. Phys. C*, 2:817–827, 1991.

[15] Adrian H. Elcock and J. Andrew McCammon. Sequence dependent hydration of dna: Theoretical results. *Journal of the American Chemical Society*, 117(40):10161–10162, 1995.

[16] D.J. Evans. On the representatation of orientation space. *Molecular Physics*, 34(2):317–325, 1977.

[17] D.J. Evans and S. Murad. Singularity free algorithm for molecular dynamics simulation of rigid polyatomics. *Molecular physics*, 34(2):327–331, 1977.

[18] R. Everaers and M. R. Ejtehadi. Interaction potentials for soft and hard ellipsoids. *Phys. Rev. E*, 67(4):041710, Apr 2003.

[19] M. Feig and B.M. Pettitt. Structural equilibrium of dna represented with different force fields. *Biophysical journal*, 75(1):134–149, 1998.

[20] Michael Feig and B. Montgomery Pettitt. Sodium and Chlorine Ions as Part of the DNA Solvation Shell. *Biophys. J.*, 77(4):1769–1781, 1999.

[21] A.R. Fersht. The hydrogen bond in molecular recognition. *Trends Biochem. Sci*, 12:301–304, 1987.

[22] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*. Academic Press, 2002.

[23] Jose Garcia de la Torre, Maria L. Huertas, and Beatriz Carrasco. Calculation of Hydrodynamic Properties of Globular Proteins from Their Atomic-Level Structure. *Biophys. J.*, 78(2):719–730, 2000.

[24] J. G. Gay and B. J. Berne. Modification of the overlap potential to mimic a linear site–site potential. *The Journal of Chemical Physics*, 74(6):3316–3319, 1981.

[25] Noel L. Goddard, Grégoire Bonnet, Oleg Krichevsky, and Albert Libchaber. Sequence dependent rigidity of single stranded dna. *Phys. Rev. Lett.*, 85(11):2400–2403, Sep 2000.

[26] Bertil Halle and Monika Davidovic. Biomolecular hydration: From water dynamics to hydrodynamics. *Proc Nat Acad Sci USA*, 100(21):12135–12140, 2003.

[27] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.

[28] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[29] Thomas E. Cheatham III and Matthew A. Young. Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. *Biopolymers*, 56-4:232–256, 2001.

[30] C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693, 1997.

[31] J.E. Jones. On the determination of molecular fields. ii. from the equation of state of a gas. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, pages 463–477, 1924.

[32] Thomas A. Knotts, IV, Nitin Rathore, David C. Schwartz, and Juan J. de Pablo. A coarse grain model for dna. *J. Chem. Phys.*, 126(8):084901, 2007.

[33] S.V. Kuznetsov, Y. Shen, A.S. Benight, and A. Ansari. A semiflexible polymer model applied to loop formation in dna hairpins. *Biophysical Journal*, 81(5):2864–2875, 2001.

[34] H. Lamb. *Hydrodynamics*. Dover Publications, 1932.

[35] F. Lankaš, J. Šponer, P. Hobza, and J. Langowski. Sequence-dependent elastic properties of dna. *Journal of Molecular Biology*, 299(3):695–709, 2000.

[36] M. Levitt. Computer simulation of dna double-helix dynamics. In *Cold Spring Harb. Symp. Quant. Biol*, volume 47, pages 251–262, 1983.

[37] Jenn Huei Lii and Norman L. Allinger. Molecular mechanics. the mm3 force field for hydrocarbons. 3. the van der waals' potentials and crystal data for aliphatic and aromatic hydrocarbons. *Journal of the American Chemical Society*, 111(23):8576–8582, 1989.

[38] Sairam S. Mallajosyula, Ayan Datta, and Swapan K. Pati. Structure and electronic properties of the watson-crick base pairs: Role of hydrogen bonding. *Synthetic Metals*, 155(2):398 – 401, 2005. Proceedings of the Sixth International Topical Conference on Optical Probes of Conjugated Polymers and Biosystems, Bangalore-INDIA, January 4-8th, 2005.

[39] K.J. McConnell and D.L. Beveridge. Dna structure: what's in charge? *Journal of Molecular Biology*, 304(5):803–820, 2000.

[40] Boris Mergell, Mohammad R. Ejtehadi, and Ralf Everaers. Modeling dna structure, elasticity, and deformations at the base-pair level. *Phys. Rev. E*, 68(2):021911, Aug 2003.

[41] Y. Mo. Probing the nature of hydrogen bonds in dna base pairs. *J. Mol. Model.*, 12:665–672, 2006.

[42] MC Murphy, I. Rasnik, W. Cheng, T.M. Lohman, and T. Ha. Probing single-stranded dna conformational flexibility using fluorescence spectroscopy. *Biophysical Journal*, 86(4):2530–2537, 2004.

[43] Jonathan J Nakane. Nanopore sensors for nucleic acid analysis. *Journal of Physics Condensed Matter*, 15(32):R1365, 2003.

[44] T. T. Nguyen and B. I. Shklovskii. Persistence length of a polyelectrolyte in salty water: Monte carlo study. *Physical Review E*, 66:021801, 2002.

[45] S.O. Nielsen, C.F. Lopez, G. Srinivas, and M.L. Klein. Coarse grain models and the computer simulation of soft materials. *Journal of Physics Condensed Matter*, 16(15):481–512, 2004.

[46] NIH/NHGRI. International human genome sequencing consortium announces "working draft" of human genome `http://www.nih.gov/news/pr/jun2000/nhgri-26.htm`. 2000.

[47] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[48] D.C. Rapaport. *The art of molecular dynamics simulation*. Cambridge University Press, 2004.

[49] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327–41, 1977.

[50] E.J. Sambriski, D.C. Schwartz, and J.J. de Pablo. A mesoscale model of dna and its renaturation. *Biophysical Journal*, 96(5):1675–1690, 2009.

[51] B. Schneider and H.M. Berman. Hydration of the DNA bases is local. *Biophys. J.*, 69(6):2661–2669, 1995.

[52] S.B. Smith, Y. Cui, and C. Bustamante. Overstretching b-dna: The elastic response of individual double-stranded and single-stranded dna molecules. *Science*, 271(5250):795, 1996.

[53] Harald L. Tepper and Gregory A. Voth. A coarse-grained model for double-helix molecules in solution: Spontaneous helix formation and equilibrium properties. *J. Chem. Phys.*, 122(12):124906, 2005.

[54] B. Tinland, A. Pluen, J. Sturm, and G. Weill. Persistence length of single-stranded dna. *Macromolecules*, 30(19):5763–5765, 1997.