Classical and Quantum Simulations of DNA/Spermine Systems

Jack William Shepherd

MSc by Research

University of York

Physics

January 2015

Abstract

A study of the effects of the presence of multiple spermine molecules in the major groove of poly-d(A)$_{20}$.poly-d(T)$_{20}$ on the A-B transition has been carried out, alongside a survey of the long-term effect of a single spermine molecule diffusing into various DNA sequences. The transition from A- to B-form was found to be slowed by spermine, and further slowed as more spermine was added. In some cases, the presence of spermine induced an unstable backbone shift away from both the canonical A- and B-forms, which eventually decayed rapidly into previously seen behaviour. The interactions between various sequences of DNA and spermine indicated that the strength of the interaction can be affected by the DNA sequence.

The efficacy of the molecular dynamics force field was assessed using *ab initio* quantum mechanics as a reference. The classical approximation was found to give results which were consistently lower in magnitude than those from the quantum calculations, and in thymine a regular site of considerable inaccuracy was observed. This represents a novel finding and a target for future development of molecular dynamics techniques.

Contents

List of Figures

List of Tables

Acknowledgements

I would like firstly to thank Dr Matt Probert and Dr Robert Greenall, who supervised this work, not only for their time, scientific knowledge, and ideas but for their patience with and belief in me. I am also indebted to Dr Phil Hasnip, who has been a great source of information and pointers, and many times helped me to make sense of things and devise intelligent strategies.

Dr James Ramsden has been a truly great friend to me and helped keep my spirits up through the hard times - thank you. I also thank the rest of N102 for helping me with physics and making me feel at home, especially Ed for his relentless stream of "jokes", Aaron for his inspiring photographs, and Matt for regular speculation on general relativity.

Dr Iain Johnston of Imperial College was originally responsible for my being interested in biological and computational physics, so I would like to thank him for all those interesting conversations in the King's Arms which set me on this path. I am also grateful for the help he rendered during my application for postgraduate study - his advice was invaluable.

Finally, my deepest thanks go to my parents for their amazing support - without them, none of this would have been possible.

Author's Declaration

I declare that the work presented in this thesis, except where otherwise stated, is based on my own research and has not been submitted previously for degree in this or any other university. No parts of the work reported in this thesis have appeared elsewhere.

1 Introduction

1.1 History of DNA

Since Mendel, it has been known that there is some system in all living organisms whereby information which determines various properties is passed from parent to child. How this vital mechanism operates and what it consists of is a question which has only begun to be answered in the last hundred and fifty years.

The molecule which we have come to know as Deoxyribonucleic Acid (or, more commonly, DNA) was first observed by Friedrich Miescher in 1871, purified from the nucleus of cells. Miescher termed it "nuclein". While his discovery went on to enjoy very little celebrity, his work on the chemical properties of DNA were to be of great importance when the function of DNA became apparent thanks to the work of Albrecht Kossel, who described the 5 bases of DNA - adenine (A), thymine (T), cytosine (C), guanine (G) and uracil (U). In 1910, Kossel was awarded the Nobel Prize "in recognition of the contributions to our knowledge of cell chemistry made through his work on proteins, including the nucleic substances"[1]. However, due to technological limitations, the physical structure remained unknown.

The great leap forward which enabled direct observation of DNA was X-ray crystallography. Using diffraction techniques, Rosalind Franklin and colleagues at Kings College London were able to produce striking images of DNA molecules in two conformations which have gone on to be known as the A- and B-form. From these photographs, it was possible for Franklin to construct the enduring model of DNA - an outer phosphate backbone, with bases inside, arranged in a double helix.

Unfortunately, a large amount of Franklin's work was not published immediately, and thus it was James Watson and Francis Crick - building on the work of Franklin and others - who first published the double helical structure of DNA in their 1953 Nature paper "A structure for Deoxyribose Nucleic Acids"[2]. A runaway success, their paper energised the field of molecular biology and allowed considerable advances in the study of DNA to be made, as well as helping to secure them and Maurice Hugh Frederick Williams the 1962 Nobel Prize. It

finishes with an insightful sentence which predicts work that continues even to this day: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material". Since then, the mechanism behind DNA replication has indeed been the focus of a great deal of research.

However, the replicative function is not the only one which enables DNA to control the properties of a cell. Via interactions with proteins and enzymes in the cell, protein expression and cell death are also mediated. The way in which these interactions can take place is affected by the shape of the DNA - if the double helix took on an unexpected structure, a binding site may not be recognised by a ligand and hence not bound to, for example. Because of this, the effect of various factors on the conformation of DNA has also been of considerable interest.

Predicting the conformation which it is energetically preferable for a stretch of DNA to assume is a non-trivial problem, and structure prediction in general remains an active area of biological research. Analytically, it is virtually impossible, so simulations of DNA and ligands are instead performed and analysed. As the phosphate backbone of DNA is electronegative, positive counterions are found *in vivo* (or placed *in silico*) around the DNA to keep it stable. Which counterions are used does not appear to help determine the overall structure of the molecule, but may affect some conformational parameters[3]. Similar studies have also been performed of DNA binding of a polyamine, a larger charged molecule, which can induce significant changes including bending of the helical axis and a delay in A to B transitions[4].

Despite the large amount of effort and resources devoted to it, DNA remains an exciting area of research, with progress still being made. The emergence of new and more powerful techniques and equipment for simulation and experimentation continues to open up new possibilities, and with continuing interest in the causes of and possible treatments for genetic and DNA-based diseases, the field shows no sign of diminishing.

1.2 The A- and B-Form

When discussing the conformation of a stretch of DNA, it is useful to have well-defined reference points with which to make comparisons. The structures first observed by Franklin and her team represent two bounds, within which other naturally occuring conformations in standard conditions lie[5], and so have become the standard to refer to. They are termed the A- and B-form, and each has sub-conformations - AI, AII, BI, and BII. The BI form is by the most

common conformation to find[5], while the pure A-form is generally exhibited by partially hydrated DNA duplexes, RNA duplexes, and DNA-RNA hybrids. A great many conformations are between the two, and can be thought of as a mix of A- and B-DNA, while thermal motion and the effect of other molecules can help DNA assume otherwise energetically unfavourable conformations for short stretches of time. Indeed, a major problem in the field of molecular biology is understanding how large local deviations from canonical structures affect DNA function.

In order to identify the different conformations of DNA, various quantities and measurements can be compared, each of which has a definite value in the canonical forms. To discuss the different terms, it is necessary first to define the reference frame and axes which will be used during comparison.

1.2.1 The DNA Reference Frame

Figure 1.1 shows visually the definitions of the x, y and z axes with respect to an arbitrary base pair. As can be seen in the box marked 'Coordinate frame', the z axis is aligned with the helical axis, and is aligned with the direction of the phosphate chain which lies along the major groove. By convention, the phosphate backbones' direction runs from the end of the chain with a final phosphate group to the end which finishes with a hydroxyl group, these ends being denoted 5´ and 3´, respectively.

The Figure also usefully has separated the different quantities into similar groups for ease of comparison. From the top left and working clockwise, these are intra-base pair, inter-base pair, and base pair axis parameters.

1.2.2 RMSd

The simplest means to evaluate a conformation is to take the value of the root-mean-square deviation of the positions of the atoms on the phosphate backbone of the system of interest from the positions of the atoms on the phosphate backbone of the canonical forms were they to be overlaid. Whilst this provides a useful, and computationally cheap, measure of the state of the system, it is a coarse-grained measure only, and relies purely on atoms on the outer backbone to give information about the positions of the bases. In fact, the differences between conformations are much more subtle than this, and the backbone itself is the location for distinguishing between the substates of the A- and B-form.
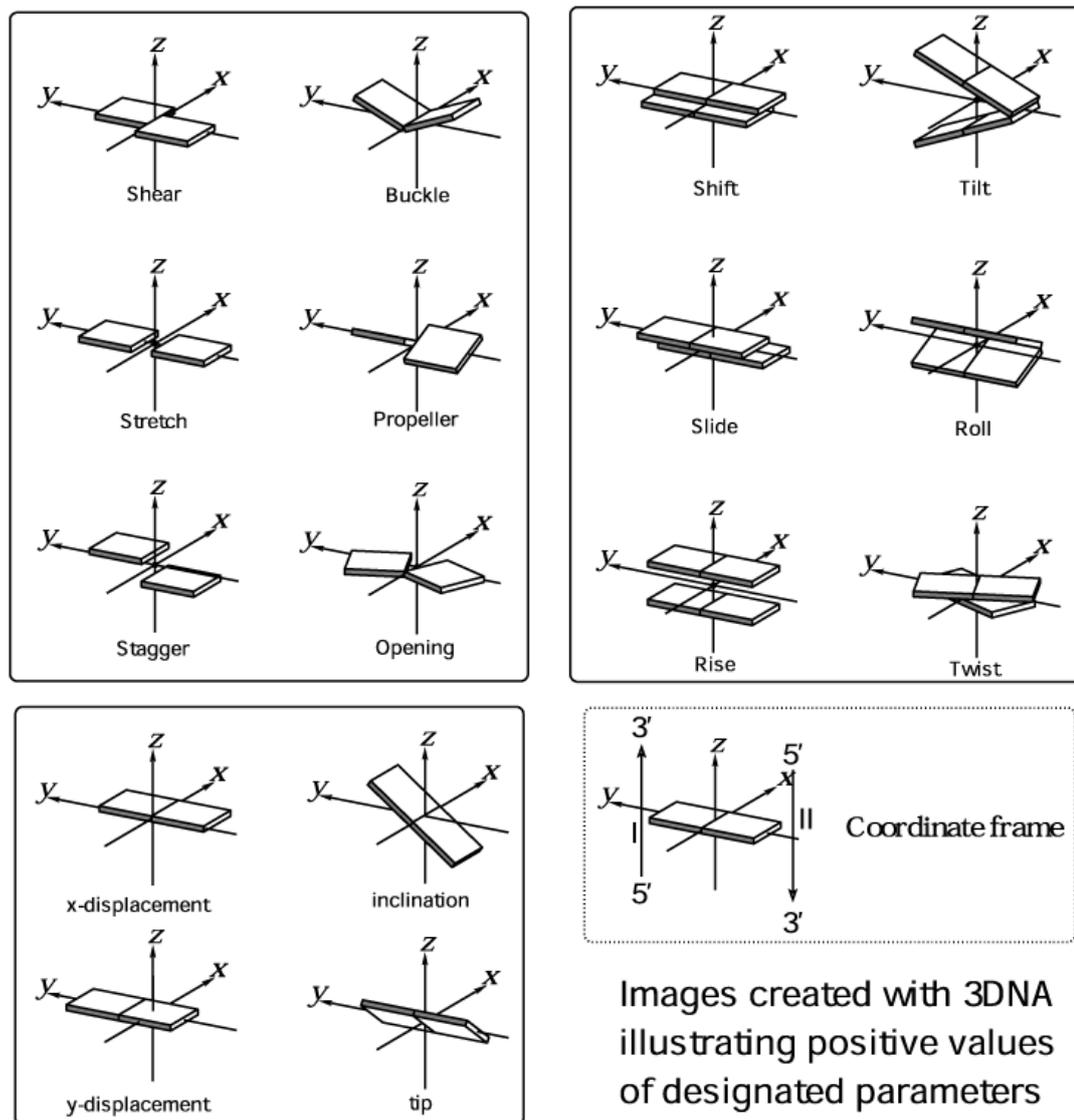
Figure 1.1: Graphical representations of various DNA parameters.
Taken from
http://rutchem.rutgers.edu/~xiangjun/3DNA/images/bp_step_hel.gif

1.2.3 Intra-Base Pair Measurements

The intra-base pair quantities are those in the top left box of Figure 1.1. These quantities are functions of the relative positions and rotations of the bases in a Watson-Crick base pair with respect to each other. Briefly defined in words, they are:

- *Shear:* The relative displacement along the x axis of the two bases, $x_{base1} - x_{base2}$

- *Buckle:* The angle each base makes with the y axis. The outer edge of the base being further along z than the centre is a positive value

- *Stretch:* The distance along the y axis between the two bases, $y_{base1} - y_{base2}$

- *Propeller:* The relative rotation of each base around the y axis

- *Stagger:* The relative displacement in the z direction, $z_{base1} - z_{base2}$

- *Opening:* The angle between the bases due to rotation about the z axis

The values of most of these parameters are not useful when evaluating conformations, as the value is the same for either A- or B-DNA. However, the propeller twist is different between the two canonical forms, and can be used a measure. In A-form DNA, the mean propeller twist is 18, while in B-form it is 16.

1.2.4 Inter-Base Pair Measurments

The inter-base pair parameters are of more utility than the intra-base pair ones - both twist and rise are different in A- and B-DNA, so can be used in analysis of a structure. They refer to differences in rotation and position between two adjacent Watson-Crick base pairs, and can be described as:

- *Shift:* The relative x displacement of successive pairs, $x_{bp(i+1)} - x_{bp(i)}$

- *Tilt:* The relative rotation about the x axis between pairs

- *Slide:* The relative displacement along the y axis of adjacent base pairs, $y_{bp(i+1)} - y_{bp(i)}$

- *Roll:* The relative rotation about the y axis

- *Rise:* The distance along the z axis between successive base pairs

- *Twist:* The rotation about the z axis between successive base pairs

The rise per base pair and twist angle are both crucial measurements used when identifying conformations of DNA. In A-DNA they are $0.26nm$ and 32.7° respectively, whilst in B-form DNA they are $0.34nm$ and 35.9°. This is a large factor in the different appearance of A- and B-form DNA. As B-DNA has a slightly higher twist per base pair, it has fewer base pairs per complete turn of the helix than A-form - 10.5 to A-DNA's 11. However, these are close enough that the significantly higher rise per base pair leads to a long stretch of DNA per complete turn of the helix for B-DNA, while A-DNA remains fairly squat and compact by comparison. The difference is significant: A-DNA has a pitch of just $2.82nm$ compared to B-DNA's $3.32nm$, making one complete turn of B-DNA 27% longer than that in A-form DNA.

1.2.5 Base Pair Axis Parameters

The base pair axis parameters are the values which specify the location and rotation of a base pair with respect to the helical axis. The x-displacement and inclination combine with those discussed in Sec 2.4 to give the A- and B-form their distinctive appearance. Summarising them:

- *x-displacement:* The distance along the x axis the base pair is found

- *Inclination:* The rotation about the x axis of the base pair

- *y-displacement:* The distance along the y axis the base pair is found

- *Tip:* The rotation of the base pair about the y axis

The tip and y-displacement are the same for both A- and B-DNA. However, the x-displacement has a dramatic effect on the appearance of the DNA structure, and the inclination is also a useful quantity to look at when comparing conformations. In B-DNA, the x-displacement is around -0.8Å, while in A-DNA, it is around $-5.5\mathring{A}$. This leads to the characteristic 'hole' seen when observing A-DNA down the helical axis, as seen in Figure 1.2.

1.2.6 Phosphate Backbone Parameters

Having now fully described the positions of the bases with respect to one another and with respect to the canonical values, we can look more closely at the crucial phosphate chain which holds the bases together, and allows for the DNA to be soluble. There are two parameters which are of interest in this area of the DNA double helix, and each is very helpful in the analysis of conformations.
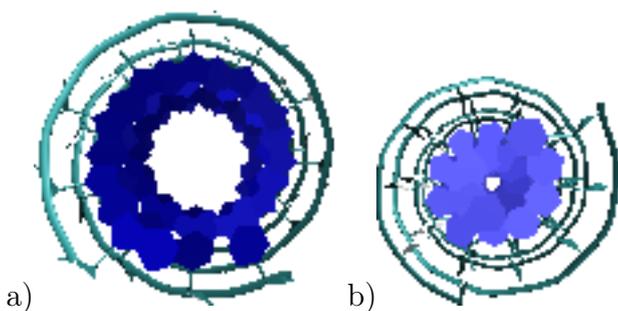
Figure 1.2: Axial view of a) A-DNA and b) B-DNA
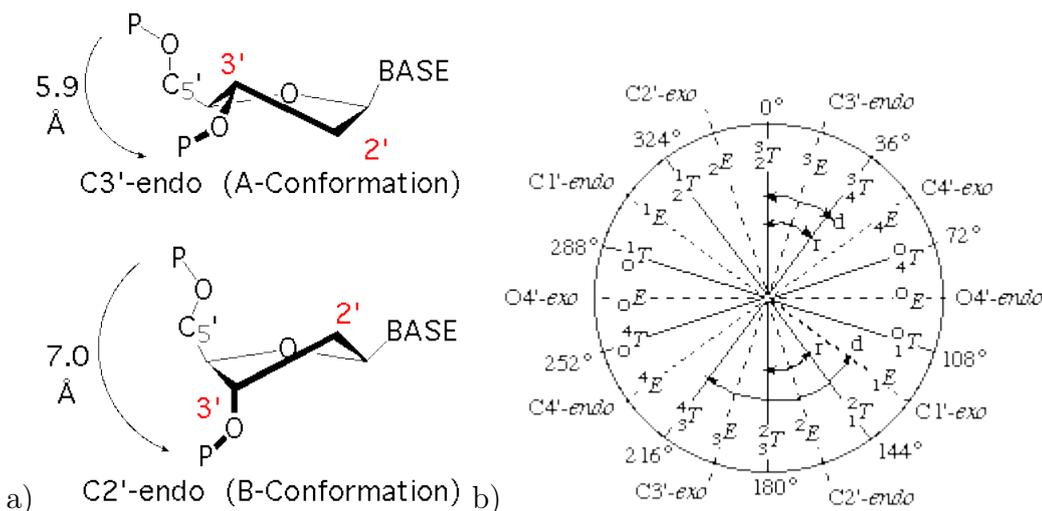
Figure 1.3: a) Diagram representing the C2´-endo and C3´-endo sugar puckers
of DNA.
b) Pseudorotational classification of sugar puckers.
a) Taken from
http://ww2.chemistry.gatech.edu/~lw26/structure/nucleic_acid/ribose/pucker.gif;
retrieved 2/1/2015
b) Taken from http://www.chem.qmul.ac.uk/iupac/misc/pnuc2.html; retrieved
2/1/2015

1.2.6.1 Sugar Puckers

The phosphate backbone is not linked directly to the bases, but instead that link
is mediated by a pentose sugar called 2-deoxyribose. 2-deoxyribose is not nor-
mally planar - instead, two carbon atoms lie outsie of the plane of the remaining
three. C2´ and C3´ are the two carbons which lie outside of this plane, and
their positions define the sugar pucker. Whichever one of them is down is said
to be -endo and the upper is said to be -exo. The -endo suffix is conventionally
used when describing the puckers, so that the two are C2´-endo (C2´ below
the plane, C3´ above) and C3´-endo (C3´ below the plane and C2´ above).
Generally, A-DNA's sugar puckers will be C3´-endo, and for B-form DNA, the
C2´-endo wins out. These two configurations are shown in Figure 1.3a.

a)   b)
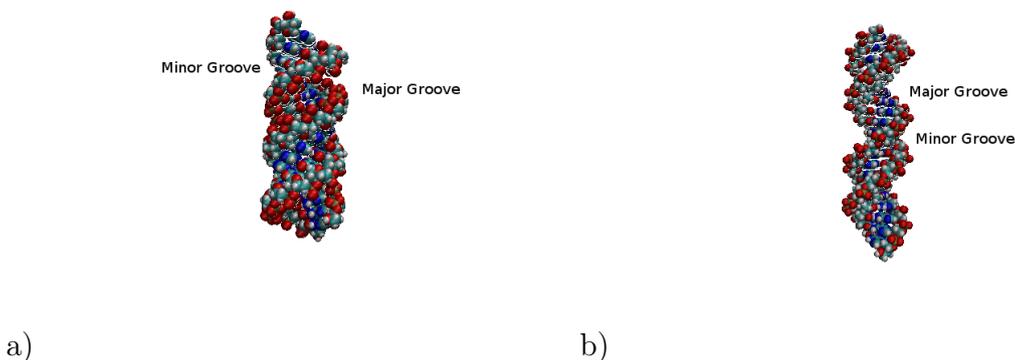
Figure 1.4: Canonical A- and B-forms (subfigures a and b respecively) of DNA with labelled major and minor grooves

However, the -endo/-exo notation neglects a good deal of information which may be of use. Knowing that the C2′ or C3′ is above the plane tells us that a sugar is in a more or less A- or B-like state. Quantifying how close or far away from one of those states the sugar is can be very helpful when dealing with a structure which is a mixture of A- and B-like states. The way that the puckers are numerically classified is known as the pseudorotational angle, and allow quantitative comparison of sugar puckers. ϕ_m is also defined to measure the amplitude of the puckering[6]. Between them, the pseudorotational angle and ϕ_m can unambiguously describe all sugar puckers. The pseudorotational classification of sugar puckers is shown in Figure 1.3b.

1.2.6.2 The Major and Minor Grooves

In both the A- and B-form of DNA, major and minor grooves are described by the phosphate chains, as can be seen in Figure 1.4. These grooves are where ligands bind, and so their geometries are of great interest. In order to characterise the grooves fully, the widths and depths of the grooves are defined. Whilst the groove width is trivially found, the depth is found in terms of specific atoms, and which atoms are used depends on the bases present. The groove widths and depths are defined as follows:

- *Groove width:* Distance from P $\longrightarrow$ P

- *Minor groove depth:* Distance from P $\rightarrow$ N2$_{Guanine}$; or P $\rightarrow$ O2$_{Thymine}$

- *Major groove depth:* Distance from P $\rightarrow$ O6$_{Guanine}$; or P $\rightarrow$ O4$_{Thymine}$

These distances may be reported either as is, or the distance minus the van der Waals radii of the atoms.
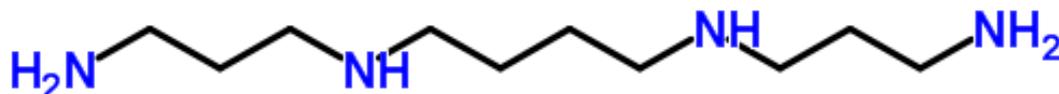
Figure 1.5: Skeleton formula of spermine.
Taken from www.chemspider.com/Chemical-Structure.1072.html
Retrieved 19th September 2014

1.3 Spermine

Spermine is a corrosive biological molecule with the molecular formula $C_{10}H_{26}N_4$ and structural formula
$NH_2.CH_2.CH_2.CH_2.NH.CH_2.CH_2.CH_2.CH_2.NH.CH_2.CH_2.CH_2.NH_2$[7], depicted skeletally in Figure 1.5. Belonging to the class of molecules known as the polyamines, and a polycation under physiological conditions, it is important for many reasons. Necessary for eukaryotic cell growth[8], it is also known to interact with ion channels, specifically those of K^+ and Ca^{2+}[9], and can be used as part of an ion conductor[10]. Spermine is also able to prevent misacylation of transcription RNA under physiological conditions, possibly by binding with the tRNA to form a structure which is easily and correctly aminoacylated[11].

Alongside these roles, spermine is also implicated in caspase (cysteine-aspartic protease) activation and hence apoptosis[12, 13], and may have a role in the generation of a cytotoxic component of seminal plasma[14]. Spermine also has a role in defending DNA from free radicals[15] and lone oxygen[16, 17].

Associated with DNA replication, cell death, and cell growth, the polyamines are of interest to researchers in cancer, specifically in chemoprevention[18].

Besides its functions within and without cells, spermine has also been implicated in various unexpected processes, including the formation of nanostructured titanium oxide[19] and the maturation of the intestines of of sea bass larvae[20] and suckling rats[21].

The primary effect of spermine on DNA which will be explored further in this thesis is its effect on conformation, and conformational stability. It is known that spermine can stabilise otherwise uncommon DNA structures, and induces bending of DNA helices even on ps timescales[22, 23]. Preferentially binding in the major groove[24, 25], spermine is also able to affect sugar puckers[25]. Of primary interest to this work is the previously demonstrated property of sper-

mine that major groove binding by spermine in adenine-rich A-DNA can delay the A→B transition[4]. The change in this phenomenon due to multiple spermines in the major groove over a timespan ten times that in [4] is investigated here.

1.4 A Very Brief History of Molecular Dynamics Simulations of DNA

As computers have become more powerful and sophisticated, simulations of molecular systems have become invaluable in exploring the world on a scale beyond the diffraction limit. Classical molecular dynamics, making use largely of electrostatic interaction and Newton's laws, has proven to be relatively computationally cheap, and tolerably accurate.

Beginning life as an exploration of hard spheres in 1957[26], molecular dynamical methods quickly went on to describe fluids, and by 1977 scientists were able to simulate proteins[27]. In the modern age, using techniques to simulate systems such as the very successful TIP3P model of water and variants thereon, the SHAKE algorithm to constrain atoms, various thermostats, and the Particle Mesh Ewald method to deal with long-range electrostatic interactions, biological simulation has become ever more accurate and capable of dealing with larger and more complex systems.

As one of the molecules most important to life as we know it, DNA has understandably been the focus of much simulation work to date. Beginning from picosecond scale simulations of DNA systems of only 5 or 8 bases in length, the state of the art now includes multiple hundred nanosecond simulations of duplex DNA[28], with 50ns simulations of DNA-ligand systems being commonplace at present.

Simulating DNA and DNA-ligand systems, while important, does not tell the full story of DNA or protein mechanics. Many processes remain difficult or impossible to investigate by means other than simulation. Whilst some processes, such as DNA stretching or protein folding, have been possible to explore via molecular dynamics techniques for some time[29, 30], others require longer stretches of DNA, longer timescales, or more complex treatment of the forces involved. As such, there has recently been a growth of MD research into complicated behaviours such as DNA supercoiling, which are critical to life but difficult to elucidate.

1.5 Previous Simulations of DNA/Spermine Systems

Polyamines are essential to the workings of all eukaryotic cells, and so have enjoyed sustained interest from the scientific community for many decades. Spermine, being one of the more interesting polyamines, has been the focus of individual studies, as well as being included in broader investigations.

Initially, as the simulations undertaken had to be performed on small systems for short periods of time, the focus was on the bending of the DNA induced by interaction with spermine. This met with considerable success, demonstrating that alternating purine-pyramidine stretches of DNA would be distorted, while homopolymers would remain largely unaffected[23]. These simulations also gave evidence towards spermine's preference for major groove binding, which had been theorised thanks to previous data taken from crystallised DNA. Spermine's role in DNA condensation was also of interest, with one paper managing a small-scale simulation to support its largely experimental findings[31].

As the technology and techniques matured, focus was shifted from the bending of DNA to its conformational changes. B→Z transitions were simulated over 75ps in 1992[32], but otherwise examining the effects on transitions between canonical conformations was largely the remit of the experimentalists.

Ten years later, the simulation of spermine/DNA complexes remained sluggish, but had changed in character somewhat: the focus more on the details of binding and interaction, rather than looking at the large-scale bending phenomena. Simulating solvation of DNA, and the effect of ligands on DNA's relationship to water, has been of interest since the first DNA trajectories were calculated, and using the techniques developed it was possible to assess the effect of spermine on DNA hydration and its interactions with counterions[33]. The specific binding sites of polyamines was also determined via simulation[34], looking further into the atomic-scale interactions.

In 2002, the work which directly precedes that presented here was published[4]. A study of A $\rightarrow$ B transitions in the presence of spermine, it demonstrated that one spermine bound in the major groove of A/T-rich DNA could delay the A to B transition. However, due to the technology available at the time, multiple spermine molecules were not considered, and the simulations themselves were run for a maximum of $5ns$.

2 AMBER Simulations of Spermine-DNA Complexes

2.1 The History of AMBER

AMBER is a very successful suite of applications which "... evolved from a program that was constructed in the late 1970s to do Assisted Model Building with Energy Refinement"[35] at UCSF. Paul Weiner, and the head of the group Peter Kollman, aimed to break free of the contemporary specialised molecular modelling techniques by developing a code which was able to perform general molecular mechanics for biological systems[36]. Several programs were developed, each of which took on one of the tasks required when undertaking molecular simulation. Code to analyse topology, link molecules together, alter the co-ordinates of atoms, relate the atom types and their properties, minimise the system's energy, and analyse the results was written, along with a graphical user interface. These utilities, which had been built to communicate with each other effectively via original file formats, formed the intial release of AMBER, and were the foundations upon which the modern software was built.

Having written these primitive tools, the developers were joined by others who wrote code able to calculate and visualise normal modes of molecules, and collaborated with others who wrote general visualisation software, without which the utility of the package was limited. A new force field had also been developed in parallel in Kollman's group, and this was incorporated into AMBER[37], and molecular dynamics capability was added[36]. These improvements and new additions came together to make the second version of AMBER.

AMBER was by that time proficient in deterministic molecular methods - energy minimisation, basic molecular dynamics, and normal mode analysis. However, new techniques were being developed at a rapid pace, and to remain at the cutting edge of the field, AMBER would need to adapt and extend yet further. To this end, Monte Carlo methods and statistical mechanics for free energy calculations were worked into code and included in the AMBER repertoire. Quantum mechanics, another area of nature hitherto ignored by the AMBER developers, was also considered, and a hybrid quantum/classical mechanical

program was written[38], whereby *ab initio* quantum mechanics and molecular mechanics were combined in an attempt to efficiently find optimal geometries and the energy of systems. Although not as accurate as some previous *ab initio* calculations, the code was declared successful by the authors thanks to its relative computational cheapness and qualitative replication of features of the systems studied. AMBER 3 was born.

Aside from a major rewrite and various improvements to the individual components of the AMBER suite, the major addition to AMBER in version 4 was that of SANDER (Simulated Annealing with NMR-Derived Energy Restraints). Initially joining the lineup of codes as a means of performing restrained molecular dynamics, SANDER (later sander) became in later versions the AMBER molecular dynamics workhorse.

In the years, and ten versions, since AMBER 4 was released, the code has undergone significant upgrading and refinement. However, in general, the philosophy of the developers (i.e. providing a suite of programs which can be interlinked to create and analyse simulations, as opposed to one program that tries to include everything) has remained consistent. There are a few exceptions to this - for example LEaP and xLEaP were written to combine three of the vital codes: Link, Edit, and Parm into both a command-line version and one with an X-Windows-based graphical interface.

Whilst there have been many new tools and programs written which efficiently carry out certain tasks, only a few were employed in the course of the work presented in this thesis, and therefore require mention by name. The Nucleic Acid Builder (nab) program - or more accurately, language - has also been created to ease the making of input files for LEaP. Nab also allows the building of novel structures from basic components, including triplex DNA, and proteins which do not yet have ready-made structures available available. An ubiquitous trajectory analysis tool known as cpptraj is now also freely available as part of AmberTools. Its purpose is to give users a tool to perform simple analysis of coordinates on trajectories which have been generated by sander, and it includes functionality to remove water molecules and reimage the system according to specifiable parameters. Written in C++, it is generally the initial tool used by scientists using AMBER upon completion of a simulation to link individual simulation stages into one trajectory file, find conformational parameters, and remove explicit water molecules for easier visualisation.

When discussing the AMBER suite, it is vital to be clear about what AMBER is and does. It is a very effective suite of programs which can combine together within a project to allow both simulation and powerful analysis thereof. However, little of this would be possible were it not for a key component: the

force field. AMBER does not have any force field parameters built in: they are included by reading them from a file when building structures with LEaP. As such, any force field can be used (as long as it is in the right format), although few use force fields which are not developed by the AMBER community. The distinction between AMBER simulation/analysis software and the force field used by them is, however subtle, nonetheless present, and so any discussion of AMBER's history would not be complete without a similar discussion of the force field.

2.1.1 Modern History of AMBER Force Fields

The original AMBER force field was conceived at a time when computational limitations had to be at all times borne in mind when writing software. As such, along with the majority of the force fields then available, certain approximations had to be used. In particular, hydrogen was targeted as an element which would need explicit treatment in some cases - for example when the hydrogen was polar - and in the case of it being bound to another atom, specifically carbon, a combined parameter set could be employed. Within the popular 1984 forcefield[37], hydrogen was treated in this manner, which was both satisfactorily accurate, and suitably computationally cheap. This 1984 force field - known as ff84 - was parameterised against quantum chemistry calculations where possible, largely fitting partial atomic charges, and otherwise parameterised against data obtained experimentally.

Ff84 enjoyed considerably success, popularity, and longevity. As a forerunner in a new class of force field it worked as hoped - accurate, cheap, and general. However, it was not without problems. The dual treatment of hydrogen meant that systems which are very sensitive to hydrogen's placement were poorly parameterised - for example, the sugar puckers in DNA obtained via simulation were not reliable[39]. Additionally, computers were getting quicker, and so the compormises which had needed to be made were no longer necessary or acceptable. The simulation community wanted instead an all-atom force field. In 1986, ff86 was released as an all-atom update to ff84, and surpassed the popularity even of its parent force field.

As time went on, the technology improved ever more quickly. Soon, the efficacy of ff86 was also in question: devised largely for the gas phase, it coped poorly with liquid phase simulations of certain systems, for example pure biomolecules[39]. Further, new developments meant that more computational power could be expended on the quantum chemistry calculations, generating more accurate results for the force field developers to fit to. In 1994 the ff94,

also known as Cornell et al., force field was released. It went significantly beyond the scope of ff86, and was found to address, at least to some extent, the problems that the previous force fields had had. It was not, however, as long-lived as ff86, and was in 1999 surpassed by ff99, which went on to be one of the most successful force fields ever devised, and continues to be the basis of the force fields in use today. Another increase in computing technology led to ever longer simulations, and it was found that the contemporary force fields broke down in their handling of conformation parameters in proteins and DNA over long time scales.

ff99 was taken up by, amongst others, two main groups of developers as a foundation upon which to improve. The two groups - the AMBER group in Stony Brook, New York, and the other working largely in Spain - were motivated to improve the force field in different ways. The so-called "Stony Brook" improvements targetted largely the problems with simulating protein backbones, while the Spanish group found ways to target the breakdown of the α/γ helical parameter in the phosphate backbone of DNA over long simulations[40]. These improvements were made available to the general community as ff99SB and ff99bsc0, respectively.

Clearly, it is undesirable to have separate force fields which are optimised for different systems, especially systems which may be expected to interact in a simulation. Although other force fields such as ff03 were released before ff99bsc0, the force field which has become the most influential of recent years was ff10, which combined the improvements to protein and DNA modelling from ff99SB and ff99bsc0 in one force field. It is ff10 which became the basis of the most recent updates - ff12SB, ff13SB, and ff14SB are all ff10 with further Stony Brook modifications, meaning that while protein handling has been updated since, the way DNA backbones are treated has remained the same since the release of ff99bsc0.

2.2 How AMBER works

2.2.1 General Principles of Molecular Dynamics

Molecular dynamics in its most basic form is derived from very simple principles - Newton's laws of motion and electrostatic interactions. For a system composed of interacting hard spheres, the scientist must simply sum the individual forces on each sphere thanks to pairwise interactions from the other spheres, find from that total force the acceleration, and hence move the particle an appropriate distance based on the time step desired. The new velocity of the sphere is then

stored, ready for use in the next timestep.

Sadly, the universe is not composed of hard particles experiencing simple interactions thanks to their environment. Simulating molecules, for example, requires knowledge of the potential energy of - and hence force due to - the bonds which hold the molecule together. For a single covalent bond, is may be a reasonable suggestion to treat the bond as a spring with a force constant dependent on the atoms which are bonded. However, to do so it is necessary to know both the equilibrium distance - the separation of the atoms where there is no force between them thanks to their bond - and the spring constant of the bond itself. These can be found via *ab initio* quantum mechanics, but such calculations can be very expensive computationally. Further, the atoms are not simply spheres with springs sticking out of them - there is more complex geometry at work which will affect the energies and hence forces. The atoms are also not of fixed orientation - they may twist, introducing a torsion energy. This is virtually impossible to calculate or derive, and so refining the parameters which describe it must be done qualitatively[41].

When considering electrostatic interactions, it is important to place the charges on molecules in appropriate places. Nuclei and electrons which are not chemically relevant present the least challenging problem (at least, in a naive implementation): one could simply combine the nucleus with its full electron shells to give a total charge. This is then positioned at the nucleus' location and moved according to the forces on it. However, valence electrons are somewhat more difficult. The tactic employed by AMBER is to perform an *ab initio* quantum mechanical simulation of the bond, and to assess the charge distribution as a result of it. Having done this, partial electron charges are assigned to the respective nuclei and used in the energy and force calculation. A tradeoff must be found here as well. It would be desirable to have as many loci of electron charge as possible, especially in the case of nonbonded electrons - for example in water. Increasing the number of these will make the calculations more accurate, but as there are more interactions to take in to account the cost of computation will rise. It is a problem which has been solved for water by creating many different models, with different numbers of charged points. Then it is for the scientist to decide which level of accuracy is most appropriate for the simulation they are performing.

Finally, real life systems have other properties which need to be faithfully reconstructed if a physically realistic simulation is to be obtained. Keeping the temperature of the system constant is a problem which is of fundamental importance, and is the subject of considerable research in its own right. Various thermostats are in existence, of which the most popular for the type of work

done with AMBER is the Langevin thermostat. Taking into account stochastic processes, the Langevin thermostat can be used both to raise the temperature of a system, and to keep the temperature constant. A more thorough discussion of it is in 7.5

2.2.2 The AMBER force field

As a set of parameters which have been refined for thirty years, the force fields used in AMBER simultions are considerably more complex than the treatment described above. They are, however, something of a misnomer - the force fields in fact give the potential energy thanks to interactions, the gradient of which must then be found to reach a force.

Each AMBER force field since ff94 has had the same functional form, which then takes parameter values from the input files to calculate the total energy and force on an atom. A coarse-grained descritption of the force field would be to say simply that

$$E_{Total} = E_{Bonds} + E_{BondAngles} + E_{Unbonded} \tag{2.1}$$

The true functional form of the force fields is[42]

$$
\begin{aligned}
E_{Total} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \\
\sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \gamma)] + \\
\sum_{i<j} \left\{ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_1 q_2}{\epsilon R_{ij}} \right\}
\end{aligned}
\tag{2.2}
$$

The physical meaning of these terms is as follows:

- $\sum_{bonds} K_r(r - r_{eq})^2$ represents the energy of covalent bonds as simple springs with spring constant K_r. This K_r necessarily depends on the displacement of the two atoms - a constant K would behave well at small displacement but get progressively worse as it increased.

- $\sum_{angles} K_\theta(\theta - \theta_{eq})^2$ is the energy due to the atomic orbitals' geomtery and the angle between bonds.

- $\sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \gamma)]$ gives the energy caused by the torsion of the bonds - i.e. the relative rotation of the atoms.

- $\sum_{i<j} \left\{ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_1 q_2}{\epsilon R_{ij}} \right\}$ treats the electrostatic energy between atoms as being a van der Waals interaction (calculated as a modified Lennard-Jones

potential) combined with the simple electrostatic energy. This representation is able to make sense of both normal electrostatcs and hydrogen bonds.

The apparent simplicity of the force field belies its subtlety and complexity. The equilibrium values of parameters are dependent on the atoms involved in the interaction, and so there are thousands of values to be parameterised. Further, changing one parameter often necessitates changing the others to retain agreement with experiment, and often it is only by exhaustively testing the force field over long time scales that small errors in it parameterisation become apparent.

The force field as described above would work, in theory, for a system of any size and any number of atoms. However, calculating the electrostatic energy thanks to thousands of atoms thousands of times at each time step quickly becomes very expensive. Further, most biological simulations make use of periodic boundary conditions, so that there are effectively infinite electrostatic interactions to calculate. In order to be able to include long-range electrostatic interactions in the calculations, the Particle Mesh Ewald technique has been developed, which has a good level of accuracy combined with a relatively low cost of computation. Using it, it is possible to calculate the electrostatic energies quickly and accurately. Developed for other uses prior to the advent of computers, PME has become standard practice when performing simulations of large systems.

2.2.3 Solvation Models

One of the most important and common compounds on the planet, water is of fundamental significance to life. Its inclusion in serious simulations is a necessity to achieve cellular or otherwise physiological conditions. Although many early simulations were carried out *in vacuo* (modified by an aproximate dielectric term), the time of such limited computational resource is over, and now all work done on biological systems has some implementation of a solvent included. There are three basic types of solvation scheme, including the *in vacuo* method, which will here be briefly described.

2.2.3.1 Dielectric Model

The simplest way to include an approximation of water is simply to include its effect on the permittivity of free space in the electrostatic energy calculations. This is done by applying a distance-dependence to the ϵ value used when calculating van der Waals and Coulomb interactions. Whilst appropriate in the

past, when it was used to combat very restrictive technology, this approach has now been superseded by far more accurate, though computationally expensive techniques.

2.2.3.2 Implicit Solvent

More complex than the dielectric model is the implicit solvent. AMBER's implicit solvent of choice is the Generalised Born Solvation (GB) model, and its efficient parallelisation and inclusion in AMBER is still under development; it remains used in simulations today. The basic principle of the method is to separate the solute from the solvent, and model it as its own region, inside which the dielectric constant is different from that outside. The GB model makes extensive use of electrostatic theory and statistics, aiming to replace the water molecules with a continuum with the same properties as water. Beginning from the Poisson Equation, the electrostatic contribution to the free energy of solvation is approximated[43]. Rather than the analytical solution[43]

$$\Delta G_{cl} = \frac{1}{2} \sum_i q_i (\phi_{sol}(\vec{r_i}) - \phi_{vac}(\vec{r_i})) \tag{2.3}$$

an approximate form is used[44]:

$$\Delta G = -166 \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{q_i q_j}{f_{GB}} \tag{2.4}$$

where

$$f_{GB} = \left(r_{ij}^2 + \alpha_{ij}^2 e^{-D}\right)$$
$$\alpha_{ij} = \sqrt{\alpha_i \alpha_j}$$
$$D = \frac{r_{ij}^2}{(2\alpha_{ij})^2}$$

Here, α_i is the effective Born radius, which must be carefully calculated for each atomic species in solution. Physically, it has the dimension of length, and represents the distance from the atom to the surface dividing the 'inner' dielectric constant of the solute and the 'outer' dielectric constant of the solvent.

This treatment goes well beyond what was used for the dielectric method: it treats atoms as individuals with individual parameters, and it fit the data it was compared to at the time very well, for little extra computational cost. It is not as successful at treating large systems as was originally hoped, however, and so for most applications an explicit solvent technique is put to use instead.

2.2.3.3 Explicit Solvents

A great many explicit solvents have been produced which replicate water's interactions with molecules well. The most popular of these is the TIPxP family of solvents, where x is a number which denotes the number of loci of charge on the model. For example, TIP3P[45], one of the most popular models, has three charged points: one each at the locations of the oxygen and hydrogen atoms. TIP5P[46] has 5 points of charge - one each for the atoms, and two more which model the two unbonded electrons in the outer shell of the oxygen atom. As the number of points of charge increases - and bearing in mind that there may well be tens of thousands of water molecules in a simulation cell - the time taken to calculate the energy of all the solvate molecules increases. As in the parameterisation of other atoms in AMBER, the charge locations are found via quantum mechanical methods, and verified through simulation testing.

Once again, the user of AMBER must make a decision about the level of accuracy which is desired. For an exhaustive and accurate investigation into the interactions of water molecules with biological structures, TIP5P or higher may be wanted, as the accuracy may be improved at the expense of computational efficiency. For general biological simulations on multi-nanosecond scales, TIP3P is generally found to be a good balance.

2.2.4 The Langevin Thermostat

There are many algorithms which may be used to control temperatures within molecular dynamics simulations. Of these, one of the most common is the Langevin thermostat. Based on Langevin dynamics, it includes Brownian motion via a stochastic force applied to each atom at every timestep. In AMBER, the implementation of the Langevin thermostat is based on that in a 1992 paper on molecular dynamics by Loncharich, Brooks, and Pastor[47], an overview of which is presented here.

The derivation of the thermostat begins from the Langevin equation

$$m\frac{d^2x}{dt^2} = F(t) - \xi\frac{dx}{dt} + R(t) \tag{2.5}$$

Here, F is the force due to the surroundings (given by the AMBER force field), ξ represents a frictional force on the particle, and $R(t)$ is a random force, which must be uncorrelated with time, or the velocity or position of the particle. Its values take the form of a Gaussian with a mean of 0.

This equation is used in AMBER by integrating it with the leapfrog algorithm, a system of three equations which between them propagate the system in time.

They take the form[47]:

$$x_{n+1} = x_n + u_{n+\frac{1}{2}} \tag{2.6}$$

$$u_{n+\frac{1}{2}} = u_{n-\frac{1}{2}} + (x_n - x_{n-1})\frac{1 - \frac{1}{2}\gamma\Delta}{1 + \frac{1}{2}\gamma\Delta} + \left(\frac{\Delta^2}{m}\right)\left(\frac{F_n + R_n}{1 + \frac{1}{2}\gamma\Delta}\right) \tag{2.7}$$

$$v_n = \sqrt{1 + \frac{1}{2}\gamma\Delta}\left(\frac{u_{n+\frac{1}{2}} + u_{n-\frac{1}{2}}}{2\Delta}\right) \tag{2.8}$$

Here, $\gamma = \frac{\xi}{m}$, x_n, v_n, and R_n are the position, velocity, and random force at timestep n, and Δ is the timestep. The random force R_n is, as before, a Gaussian centred on 0, and for this algorithm its variance must take the form

$$< R_n^2 >= \frac{2m\gamma k_b T}{\Delta} \tag{2.9}$$

k_b is the Boltzmann constant, and T the temperature of the system.

By tweaking the values of temperature and friction constant in the variance of the random force, it is possible to either cancel out the effect of the drag force - thus keeping the system at a constant temperature - or to increase or decrease the temperature of a system.

2.2.5 The AMBER Barostat

In order to perform NPT simulations, there must be a mechanism which ensures the pressure remains constant, otherwise fluctuating temperatures may lead to non-physical behaviour. In AMBER, a barostat is implemented which at each timestep scales the volume of the simulation cell and co-ordinates of the atoms in order to keep the pressure constant.

First introduced by Berendsen *et al.* in 1984, the method is a weak-coupling method, and works by coupling the system to a pressure bath in a similar way to the Langevin thermostat. In to the equation of motion is inserted an additional term which mediates a pressure change in the bath[48]:

$$\left(\frac{dP}{dt}\right)_{bath} = \frac{P_0 - P}{\tau_P} \tag{2.10}$$

The pressure is then expressed as

$$P = \frac{2}{3V}(E_k - \Xi) \tag{2.11}$$

where Ξ is the pairwise virial

$$\Xi = -\frac{1}{2}\sum_{i<j}\vec{r_{ij}}.\vec{F_{ij}} \tag{2.12}$$

$\vec{r}_{ij}$ and $\vec{F}_{ij}$ are the vector betwen particle i and j, and the force due to particle
j on particle i, respectively. Thus it is that a pressure change may be effected
by changing the intermolecular distances and hence the virial (forces within
molecules do not contribute to the pressure). Scaling all positions as well as
the volume of the cell is the easiest and most stable means of going about this,
but requires a known scaling factor. This can be found usnig known quantities,
however.

Firstly, the definition of $\dot{x}$ must be updated: instead of the usual velocity, it
now represents a change due to the velocity and the scaling:

$$\dot{x} = v + \alpha x \tag{2.13}$$

whilst the volume changes according to

$$\dot{V} = 3\alpha V \tag{2.14}$$

The pressure change can be expressed via thermodynamic relations as

$$\frac{dP}{dt} = \frac{1}{\beta V}\frac{dV}{dt} \tag{2.15}$$

which simplifies with Equation 2.14 to

$$\frac{dP}{dt} = \frac{3\alpha}{\beta} \tag{2.16}$$

Using this expression in Equation 2.10, an expression for α is found:

$$\alpha = -\frac{\beta}{3}\left(\frac{P_0 - P}{\tau_P}\right) \tag{2.17}$$

This can now be put into Equation 2.13, which becomes

$$\dot{x} = v - \left\{\frac{\beta(P_0 - P)}{3\tau_P}\right\}x \tag{2.18}$$

Equation 2.18 represents scaling of positions and box sizes by a common
factor μ[48]

$$\mu = 1 - \frac{\beta\Delta t}{3\tau_P}(P_0 - P) \tag{2.19}$$

which is in first order of Δt, essentially an Euler integration scheme. This is,
however, true only in an isotropic system[48], and so must be cast to a tensor

form for more complex simulations. In that case, Equation 2.19 retains the same form, except that μ, 1, P_0 and P take the form of tensors.

2.3 Systems Studied and Methods

A wide variety of systems were simulated for the purposes of this study into spermine and DNA, covering a wide temporal range as well as investigating the effect of DNA sequence on the DNA/spermine interaction. Below is an overview of the systems studied, and the simulation methodolgy.

2.3.1 DNA Sequence Notation

In order to describe the systems examined, it is first necessary to define the notation which will be used in its elucidation. In general, a DNA sequence will be given a description of the form

$$poly - d(B_1)_n.poly - d(B_2)_n$$

where $poly - d$ indicates DNA ($poly - r$ would be used for RNA), B_1 is one of the four bases (A, C, T, or G) and B_2 its complement, n gives the length of the strand of DNA in base pairs, and the . shows that there are two strands which are hydrogen bonded. The above then shows two strands of DNA hydrogen bonded together, one consisting of n bases of B_1 and the other of n bases of B_2 - i.e. a standard Watson-Crick double helix of n base pairs.

In some cases, a sequence will need to be defined where the two strands do not consist of one base each. In this case, there will be notation of the form

$$poly - d(B_1 - B_2)_n.poly - d(B_1 - B_2)_n$$

Here, the bases inside the brackets give the repeating base motif, and n gives the *total* length of the DNA strand - not the number of repeats of the motif. As is usual for DNA, these motifs are specified from the 3´ to the 5´ end of the strand.

2.3.2 Systems

The simulations performed can be categorised into two classes. In the first, three independent simulations of a length of DNA with fixed sequence beginning with multiple major-groove located spermine molecules were carried out for 50ns each. The second set of simulations represent an effort to quantify the difference in interaction between DNA and spermine due to no factor other than the DNA

Sequence	No. spermine	Spermine placement	No. prepared	Simulation length
poly-d(A)$_{20}$	0	N/A	1	50ns
poly-d(A)$_{20}$	1	Major groove	3	50ns
poly-d(A)$_{20}$	2	Major groove	3	50ns
poly-d(A)$_{20}$	3	Major groove	3	50ns
poly-d(A)$_{20}$	1	Outside DNA	1	5.8ns
poly-d(A-T)$_{20}$	0	N/A	1	5.8ns
poly-d(A-T)$_{20}$	1	Outside DNA	1	5.8ns
poly-d(G)$_{20}$	0	N/A	1	5.8ns
poly-d(G)$_{20}$	1	Outside DNA	1	5.8ns
poly-d(G-C)$_{20}$	0	N/A	1	5.8ns
poly-d(G-C)$_{20}$	1	Outside DNA	1	5.8ns

Table 2.1: Summary of prepared DNA/spermine systems. The sequence for one of the two strands is given; the other strand took the complement.

sequence. The simulated time for these simulations was considerably shorter, and so for convenience the simulations performed will be called 'long' and 'short' simulations. A summary of the lengths of simulations, sequences, and presence of spermine is found in Table 2.1.

For all systems, an initial A-form Watson-Crick DNA duplex of the appropriate sequence was produced using the AMBER utility nab. The resulting PDB file (a file containing the positions and types of each atom) was then loaded into xleap, along with the parameters for the force field of choice: ff10. The choice of force fields was between ff10 and ff12SB: other force fields did not include the ffbsc0 updates which would be critical for the long simulations of DNA undertaken. Since everything in these simulations is treated the same by both force fields, as a more longstanding force field ff10 was chosen from the two options.

If applicable, spermine was then added to the structure by hand. The systems were then made electrically neutral with the addition of Na+ ions using xleap, which also automatically placed the ions according to the minima of the elctric field produced by the DNA. It was important to place the spermine prior to the counterions as the counterions and spermine carry the same charge - thus the overall system would not be electrically neutral had the DNA been counterionised first. Finally, the whole system was solvated - again using xleap - with TIP3P water molecules, with a buffer of 15.0 angstroms. This means that the minimum distance from the DNA to the edge of the simulation cell was 15.0Å. The simulation cell took the shape of a truncated octohedron.

2.3.2.1 Spermine Placement

As previously stated, the spermine molecules were placed in and around the DNA by hand. There were two reasons for this - a lack of methodology existing for automating the spermine placement, and to ensure that the maximum amount of configuration space possible was sampled. Placing the molecules by hand meant that no two would be in exactly the same place, and so the resulting simulations would also differ in outcome. All spermine was placed using the program xleap, and placement was undertaken by visualising the system and moving the spermine appropriately.

For the long simulations, various numbers of spermines were placed at various depths in a randomly selected part of the major groove of the DNA molecule, while in the short simulations the spermine was allowed to diffuse by itself towards the DNA molecule - the spermine was placed in parallel to the DNA a suitable distance from it. Had the distance been too great, the water would effectively screen the two molecules, meaning that the spermine would take a very long time to diffuse - if it did at all. However, a suitable placement range was found, so the diffusion of the spermine was reliable and reasonably rapid.

2.3.3 Minimisation Procedure

Having built and solvated a system, it was then almost ready to undergo some molecular dynamics. However, having only built the systems and placed molecules, there was a need to relax the system to allow any unfavourable (or physically impossible) configurations of atoms to be removed. To this end, a two-stage minimisation routine was used, keeping the system at a constant volume.

In the first stage of minimisation, the DNA and spermine atoms were held fixed while the positions of the water molecules was allowed to change and find the lowest-energy configuration. 10Å was the cutoff value used - beyond that the electrostatic interactions are ignored. A total of 1000 steps of minimisation was undertaken, the first half being done through steepest-descent, and the other half using a conjugate gradient method. Although the sander default is just 10 steps of steepest descent, the systems used are large enough and robust enough that it is desirable to use more in order to cut the calculation time. The residues to be restrained - the DNA and any spermines present - were held in place via virtual springs with a force constant 500kcal/mol/Å^2, which for this system is effectively infinite. Periodic boundary conditions were imposed.

The second round of minimisation allowed the DNA and spermine to move - all restraints were lifted. The steepest descent method was used for 1000

steps, followed by 1500 steps of conjugate gradient, for a total of 2500 steps of minimisation. Again, periodic boundary conditions and a constant volume were employed, along with a 10Å cutoff.

2.3.4 Heating the System

The systems were then fully relaxed, and as such not in danger of exhibiting artifacts or unphysical behaviour if molecular dynamics was to be performed. The atoms would not, however, exhibit very much behaviour at all: the minimisation routine used left all the atoms stationary in their positions and at zero kelvin. To replicate physiological conditions, it was necessary to increase the temperature of the system to something of the order of body temperature - for these simulations 300K was decided upon.

In order to heat the systems, molecular dynamics was performed using the Langevin thermostat to increase the termperature over 20ps, or 10000 2fs timesteps.

As we are now performing molecular dynamics, there are certain other considerations which must be taken into account. For example: in order to use the Langevin thermostat properly, the collision frequency between particles should be known. This value is often, for stability purposes, set to much lower than its "true" value in nature. For this heating, the collision frequency was set to 1ps^{-1}.

In order to make computation more efficient, and to allow a longer timestep, the SHAKE algorithm was also employed. This constrains the motion of all bonds involving hydrogen, and so reduces the amount of work needing to be done in each timestep. Since the molecular dynamics timestep choice is informed by the quickest motion in the system, and the stretching of chemical bonds with hydrogen in is the fastest motion in these systems, the timestep can then be safely increased. As the bonds with hydrogen atoms are constrained, we can also safely omit calculating this energy with the AMBER force field. A random seed was also generated for use in the Langevin thermostat.

The DNA was held weakly in place to allow the solvent to heat unhindered and to avoid beginning production molecular dynamics before the system was fully prepared.

2.3.5 General Molecular Dynamics Protocol

The constant-volume heating having taken place, the input file needed to be adapted to maintain a physiologically reasonably set of parameters. Keeping the volume constant as the system was heated will have had an effect on the pressure , a quantitiy which it is important to keep within acceptable bounds.

As such, we now switch from NVT molecular dynamics to NPT, in which the temperature of the system, number of atoms in it, and the pressure are all kept constant. In order for this to be possible, the volume of the simulation cell is permitted to change, and the bond lengths are changed correspondingly, according to the scheme in Section 2.2.5. For a NPT simulation, the pressure relaxation time must also be specified. AMBER recommends values between 1.0 and 5.0ps for stability; in these simulations 2.0ps was used and a physically realistic simulation was obtained.

Aside from this change in ensemble, the input parameters for production MD remain very similar to those used in the heating phase. The major changes are to the initialisation of the run - sander is instructed to read in positions and velocities for each atom so that it can restart where the previous stage of molecular dynamics left off. This is necessary as a multi-nanosecond simulation would not complete in the maximum runtime on any of the clusters which the work was carried out on. Thus, the molecular dynamics was run in many shorter stages which were stitched together at the end to produce one trajectory. Each stage was 200ps, or 100000 timesteps, and a total of 250 such stages were run for the long simulations, and 29 run for the short simulations.

2.3.6 Analysis Techniques

Initially, each trajectory stage had to be tested for being physically reasonable or otherwise. An error may lead to an increase in temperature, potential energy, or a spike in pressure, for example, and all of those cases would lead to meaningless data. Fortunately, AMBER comes packaged with a script called process_mdout.perl, which reads in the output files for each of the stages and gives data for various physical properties at each timestep. These values are put into separate files so that they can be plotted easily, and the data from all the separate stages is put together, so that only one graph needs to be viewed.

Assuming all the properties remained stable and that the simulation was therefore trustworthy, each trajectory was then combined using cpptraj, a trajectory analysing program which is included in AMBER. This served to stitch together the individual stages of the molecular dynamics, and to remove the water molecules from them so that the system could easily be visualised. Cpptraj also produced graphs of the root mean square deviation of certain important backbone atoms from the initial state as the simulation progressed, giving a coarse measure of the conformation of the DNA. By using a reference structure instead of the initial conformation of the DNA, it is easy to also generate a graph of the root mean square displacement from the reference structure - so

for a trajectory where the DNA began in the A-form, graphs of the RMSd from both the A- and B-forms could be obtained.

Having found the RMSd from the canonical forms, and removed the water molecules, finer details of the conformations could then be found. Using Curves+[49] and its sister code canal[50], the waterless trajectories were analysed, and the values of the conformational parameters at each base pair were obtained. These values were averaged, so that at each timestep there was a single value representing the mean of that value along the DNA duplex. This could then be plotted, and trends looked for. However, due to fraying in the long simulations at the end of the DNA double helix, the first and last two base pairs were necessarily neglected to reduce noise in the final values. Further, as could be expected from a biological simulation, the data remained very noisy, so a running average was also applied to obtain a more tractable graph.

Finally, it is also desirable to know how the stable values for the conformational parameters was changed as a result of spermine's presence. To calculate this, the average structure of the DNA over the simulation was calculated with cpptraj, and the resultant DNA configuration analysed as a single frame with Curves+. This allows numerical comparison of quantities, rather than relying on approximating differences on graphs. In order not to skew the data, only frames after the initial transitions had occurred were used, so that a true representation of the average could be obtained.

2.4 Results of AMBER Simulations

In this section, the results of both the long and short simulations are presented, as analysed with process_mdout.perl and Curves+/canal.

In each case, only certain conformational parameters are of interest. Some parameters are not well defined for the A- and B-form of DNA and so do not give a useful indication of the state of the system. As a coarse measure of the state the system is in, the root-mean-square deviation of key phosphate backbone atoms from the canonical A- and B-forms will be assessed. For more detail into the conformation, the parameters which will be studied are the twist, rise, and x-displacement. Combined, these give a great deal of insight into the mechanics of the DNA systems.

2.4.1 Validity of Simulation Technique

Prior to performing any analysis of the DNA conformations, it was of foremost importance to demonstrate that the simulations carried out remained physi-

a)

b)

c)
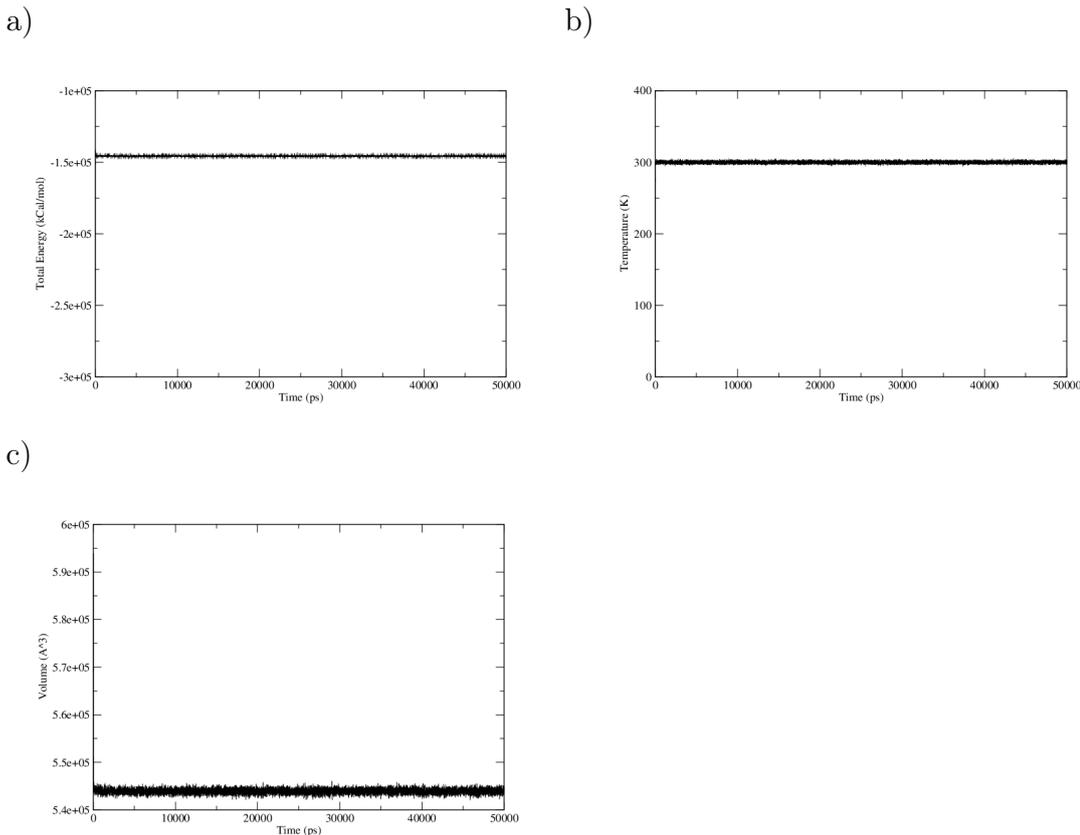
Figure 2.1: Properties of a simulation of 1 spermine molecule in the major
            groove of
            poly-d(A)$_{20}$.poly-d(T)$_{20}$
            a) The total energy b) The temperature c) The volume

cally reasonable throughout. This was done both qualititavely and quantita-
tively: firstly by simply visualising the trajectories and looking for artifacts or
any unphysical behaviour, and secondly by graphing various thermodynamic
properties of the systems. A breakdown in the simulation would be indicated
by instability in the total energy of the system, the volume, or the temperature.

As there are ten long simulations, and eight short ones, the number of graphs
produced was very high - to look at temperature, volume, and total energy re-
quires 54 graphs in total. Rather than include them all here, one representative
set of graphs are presented.

It can easily be seen from Fig 2.1 that the properties of interest were stable.
Minor fluctuations are to be expected, as in experiment, but in all cases these
remained small compared to the average values. It can, therefore, be reasonably
said that the simulations did not break down over the timescale simulated, and
so the next stage of analysis can be undertaken with some confidence.

2.4.2 The A-B transition of poly-d(A)$_{20}$.poly-d(T)$_{20}$

Aside from the control case of DNA without spermine, each case was simulated
from initial conditions three times. The data recovered from each simulation
was somewhat noisy, so for each case a running average over 1000 snapshots,
or 100ps was applied so that broad features which may have been lost in noise
could be seen. In order to make the data more tractable, for a given parameter
the time evolution for each trial will be given on one set of axes, along with a
second graph giving the mean over the three trials.

2.4.2.1 Spermine-free DNA

For the case of DNA in water without spermine only one simulation was carried
out as the behaviour would be very similar for multiple trials - there is no
factor which may affect the motion of the DNA except for thermodynamic
fluctuations, which in this case will make very little difference to the large-scale
structual changes involved in a transition away from the A-form.

It can be seen in Figure 2.2 that the DNA strand quickly moves away from
the A-form, as would be expected. Having reached roughly equilibrium values
for the rise, x-displacement, and RMSd, the DNA remains in a similar, B-like
conformation for the remainder of the simulation. It is not unexpected that the
DNA parameter values would not quite reach the canonical B-form values, as
the temperature and solvation make the situation somewhat different from the
crystallised state in which the pure B-form exists.

The twist is some way from the B-form value, and remains close to the A-
form value for the whole simulation, at times showing a value even below that
of the accepted A-form. This is to be expected - previous work showed that the
DNA twisting is more relaxed in solution than in the experiments performed by
Franklin et al. - so the graph shown is not unphysical, and does not represent
an error either in simulation technique or in the AMBER forcefield used.

2.4.2.2 Mono-spermine Systems

The RMSd plots for the systems with one spermine initially bound in the ma-
jor groove offersubstantial information. In the first two graphs in Figure 2.3,
the movement of the backbone from an A-like conformation to a B-like one is
slowed by the presence of the spermine, a property which is visible by eye when
comparing the plots to the RMSd plot in Figure 2.2. However, in the third
simulation, the RMSd plot is very similar to that in Figure 2.2, which implies
that it is not only major groove binding which is important in this process -
other factors such as the depth of the spermine in the major groove, or the

a)   b)

c)   d)
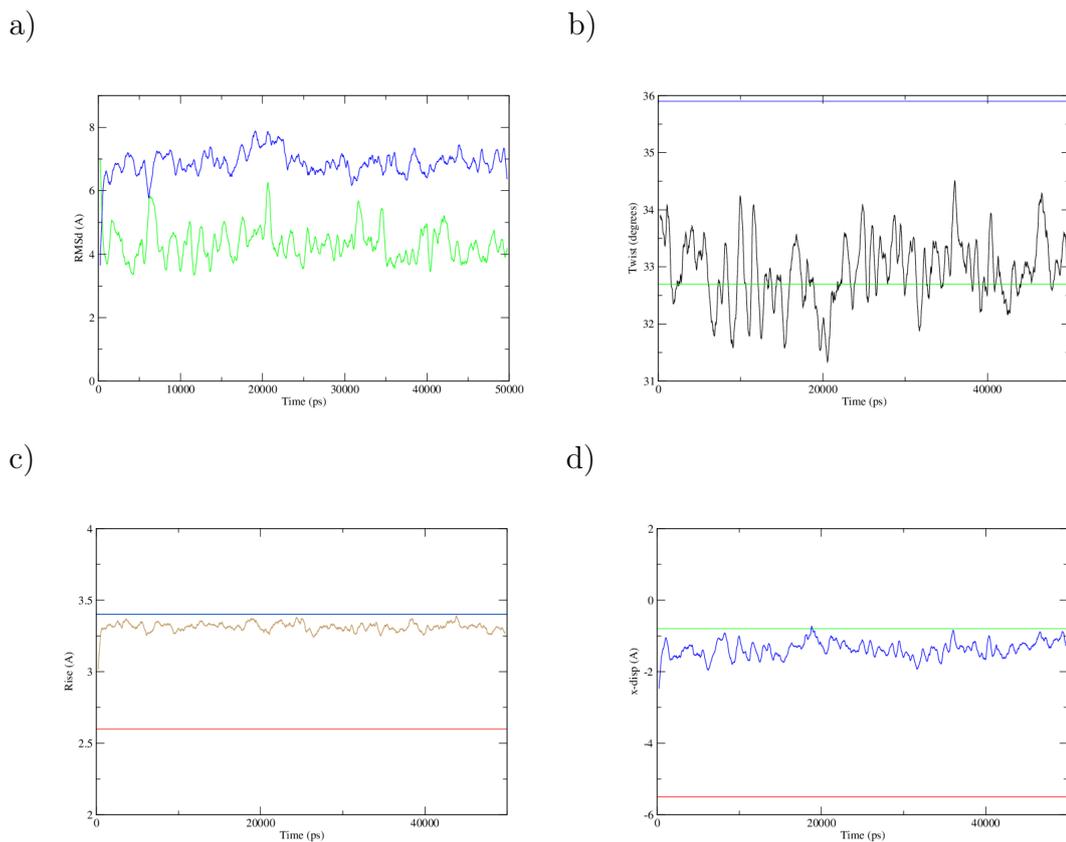
Figure 2.2: The a) RMSd between the canonical A-form (blue) and B-form
(green), b) twist between base pairs, c) rise between base pairs
and d) x-displacement of bases for spermine-free poly-d(A)$_{20}$.poly-
d(T)$_{20}$. On each graph, the lower horizontal line represents the
canonical A-form value, while the upper line is the canonical B-form
value.
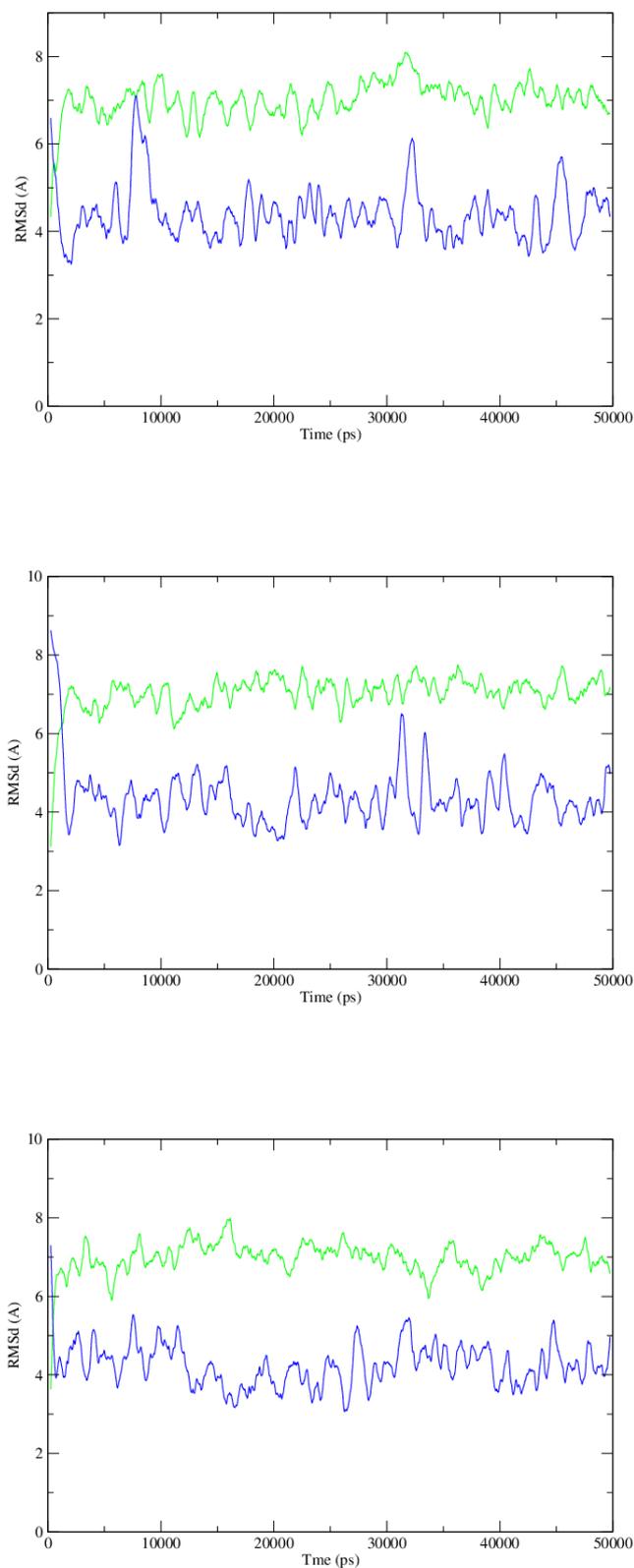
Figure 2.3: The root mean square displacement of backbone atoms from the A-
form (green) and B-form (blue) for the a) first simulation, b) second
simulation, and c) third simulation of poly-d(A)$_{20}$.poly-d(T)$_{20}$ with
one spermine initially in the major groove

proximity of certain areas of the spermine to certain atoms in the DNA may also be of great importance.

Comparing the average values for the rise and x-displacement in this case with their counterparts in the spermine-free case once again demonstrates that the spermine has had an effect on the rate of transtition from A to B. Once again, however, it is shown that when the DNA has reached a stable conformation, it remains there for the remainder of the simulation. The shift in average parameters will be explored as part of later analysis.

Finally, it is noted that when looking at the plots with the values for all three trials, the values change at different rates initially. Once again, as seen in the RMSd graphs, this demonstrates that the presence of spermine alone is not sufficient to produce a strong response - its location in the major groove is also important.

2.4.2.3 Di-spermine Systems

Two features are immediately aparrant when looking at the RMSd graphs in Figure 2.5. On comparison with the equivalent graphs for the one spermine and spermine-free cases, it can be seen that the initial transition from A- to B-form denoted by the steep gradient at the beginning of the simulation is slowed - that is, the gradient is less steep than previously, and so the transition takes longer on average.

However, this is only true for the first two simulations. In the third, the DNA backbone takes on a conformation further from both the canonical A- and B-form, before rapidly collapsing back into a B-like state. This is a situation which was not observed either in the spermine-free or single-spermine simulations, suggesting that the presence of the spermine has in this case induced a temporary shift in the DNA to an unstable conformation away from those which have been compared against.

Analysing the values of conformational parameters for these systems through time paints a similar, and more compelling, picture. Whilst there is little evidence on the surface of a totally novel conformation being formed, it is more striking in Figure 2.6 than 2.5 that the rate of the initial transition is reduced - both when compared to a spermine-free control system, and the simulations incorporating one spermine molecule.

2.4.2.4 Tri-spermine Systems

The three spermine systems offer another insight into DNA/spermine interactions. The RMSd graphs in Figure 2.7 follow the general pattern seen previ-
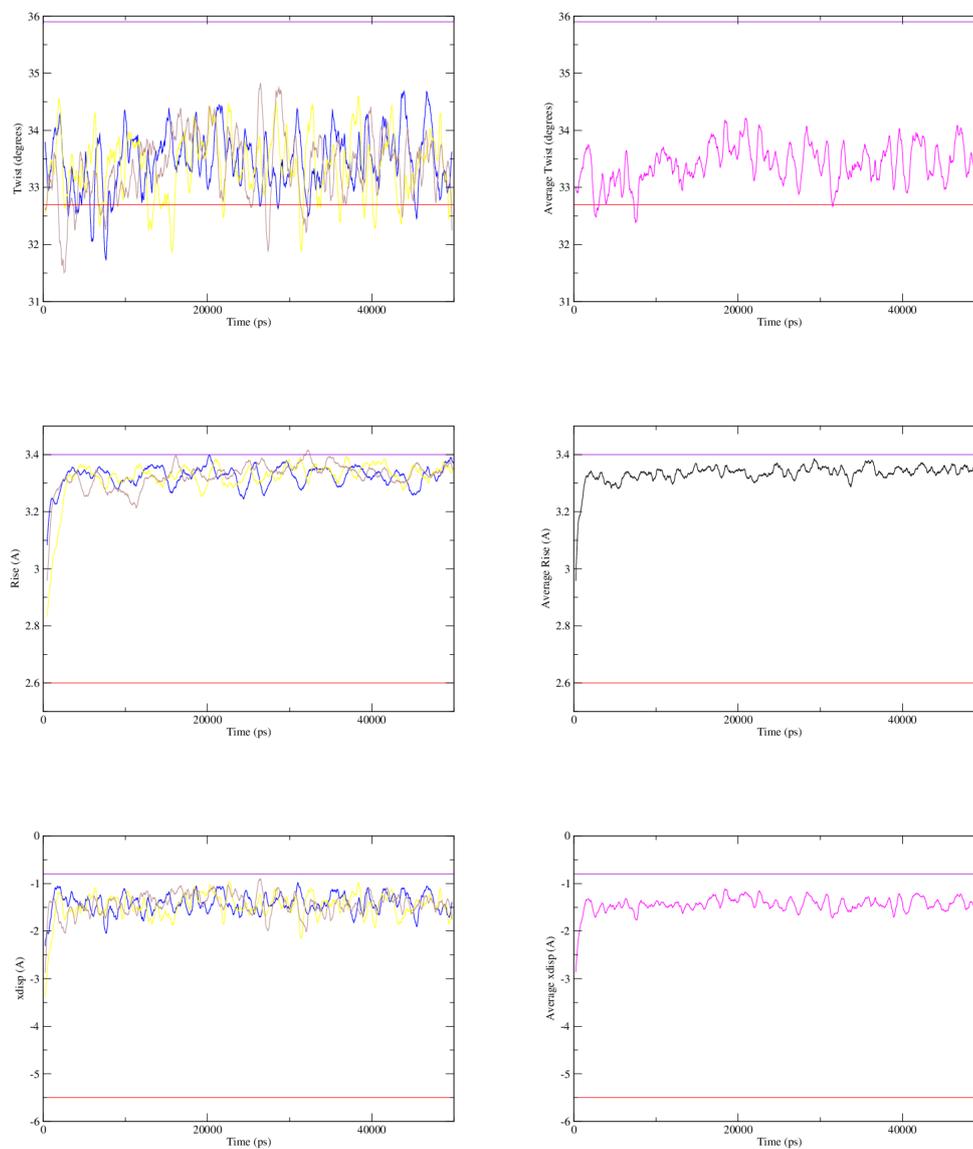
Figure 2.4: The a) twist between base pairs, b) rise between base pairs and c) x-displacement of bases for poly-d(A)$_{20}$.poly-d(T)$_{20}$ with one spermine intially placed in the major groove.

The values for each trial of the simulation (with a 1000 timestep running average applied) is plotted on the left, and the average of the trials on the right. B-form values are shown by the violet line, and A-form values by the red.
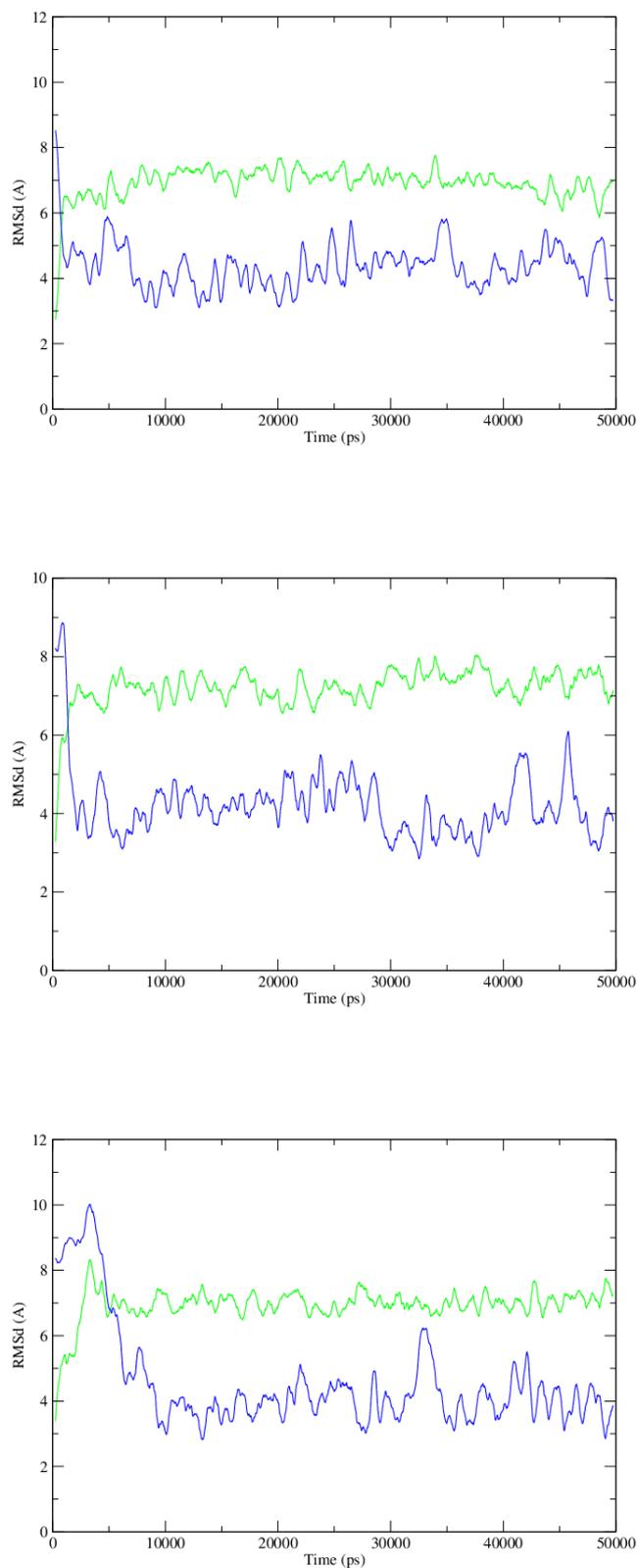
Figure 2.5: Root mean square deviation from the canonical A-form (green) and B-form (blue) for the three simulations involving two spermines in the major groove of poly-d(A)$_{20}$.poly-d(T)$_{20}$
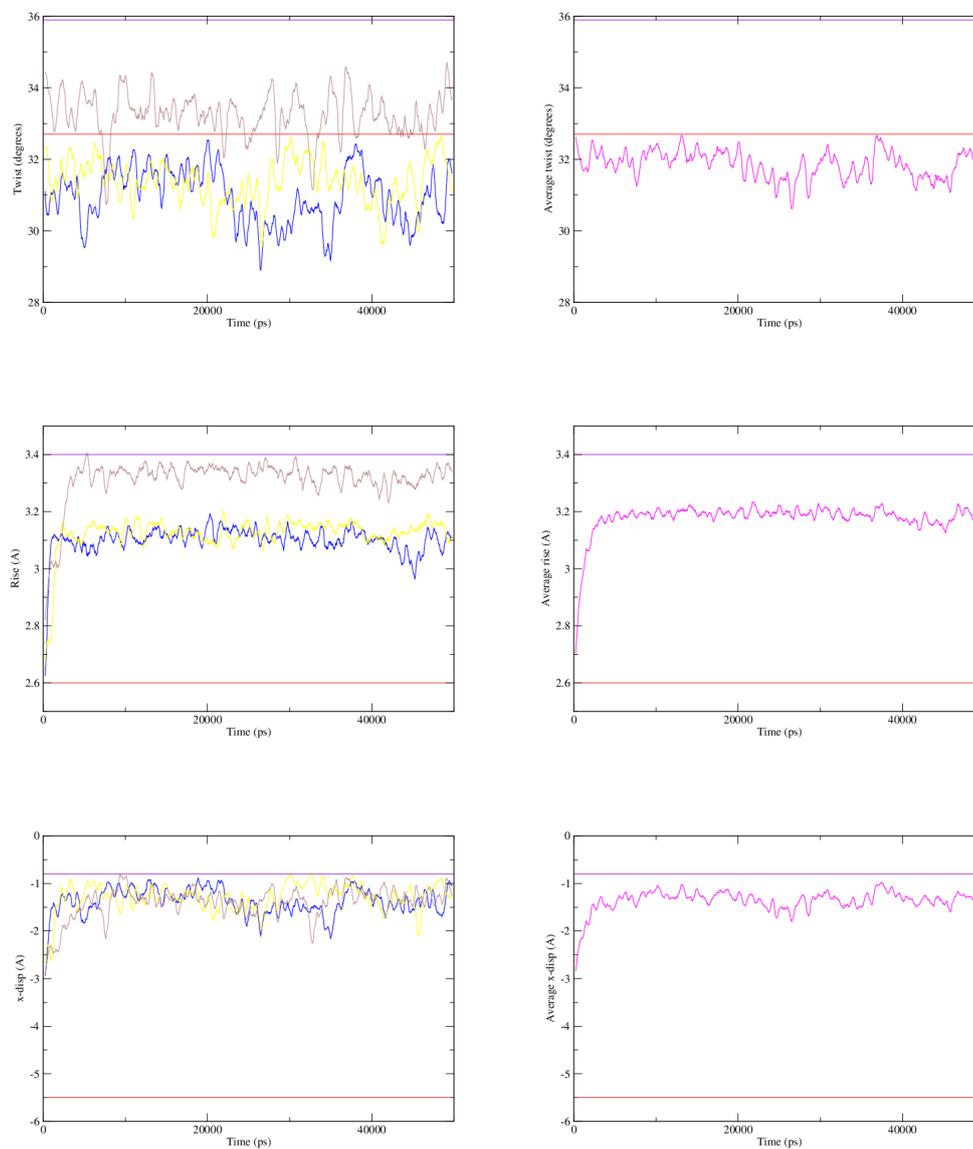
Figure 2.6: The twist, rise, and x-displacement for each simulation with two spermine molecules, and the average graph for each of these values.

The canonical A- and B-form values are shown by the red and violet straight lines, respectively
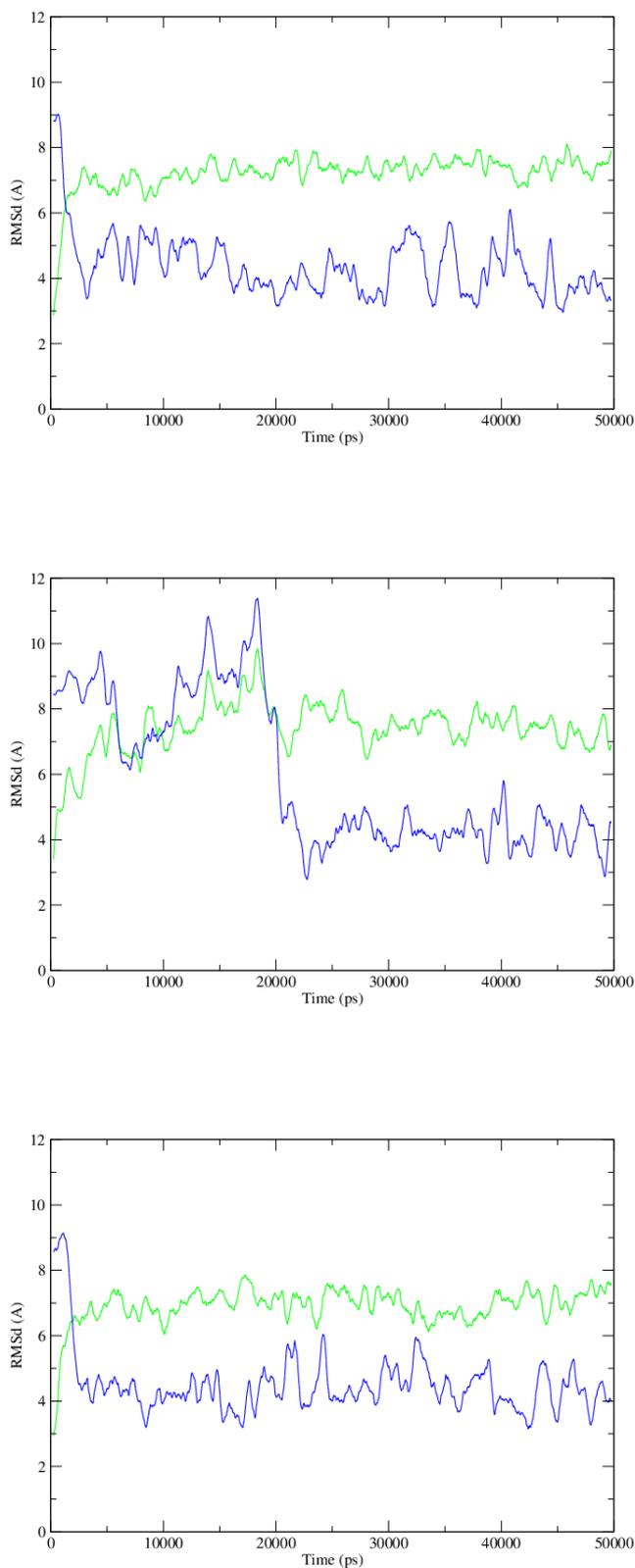
Figure 2.7: Root mean displacement from the canonical A-form (green) and B-form (blue) for the three 50ns simulations of poly-d(A)$_{20}$.poly-d(T)$_{20}$ with three spermines initially in the major groove

ously. Again, the addition of another spermine molecule has slowed the initial conformational transition.

This reduction in transition rate is not the most striking feature of the graphs, however. Once again, one of the graphs shows a clear movement away from both A and B forms - this time for a period covering around 20ns. During this time, a relatively stable state appears to have been reached on two occasions - once between 0-5ns, and again between around 12 and 17ns. The scale of this deviation from both canonical forms is in this instance greater in magnitude both temporally and spatially.

Whilst the second graph shows this behaviour dramatically, the other two do also in a more modest way. At the beginning of each simulation, it is seen that there is an initial movement away from the B-form which quickly is lost as the structure reverts to the expected motion.

The parameters of the DNA in the three spermine system are also noticeably changed. The initial gradient of the lines in Figure 2.8 is clearly shallower, and the average behaviour over the three simulations plainly demonstrates that the interactions between the DNA and polyamines is affecting the stability of the initial conformation, and making the transition less rapid.
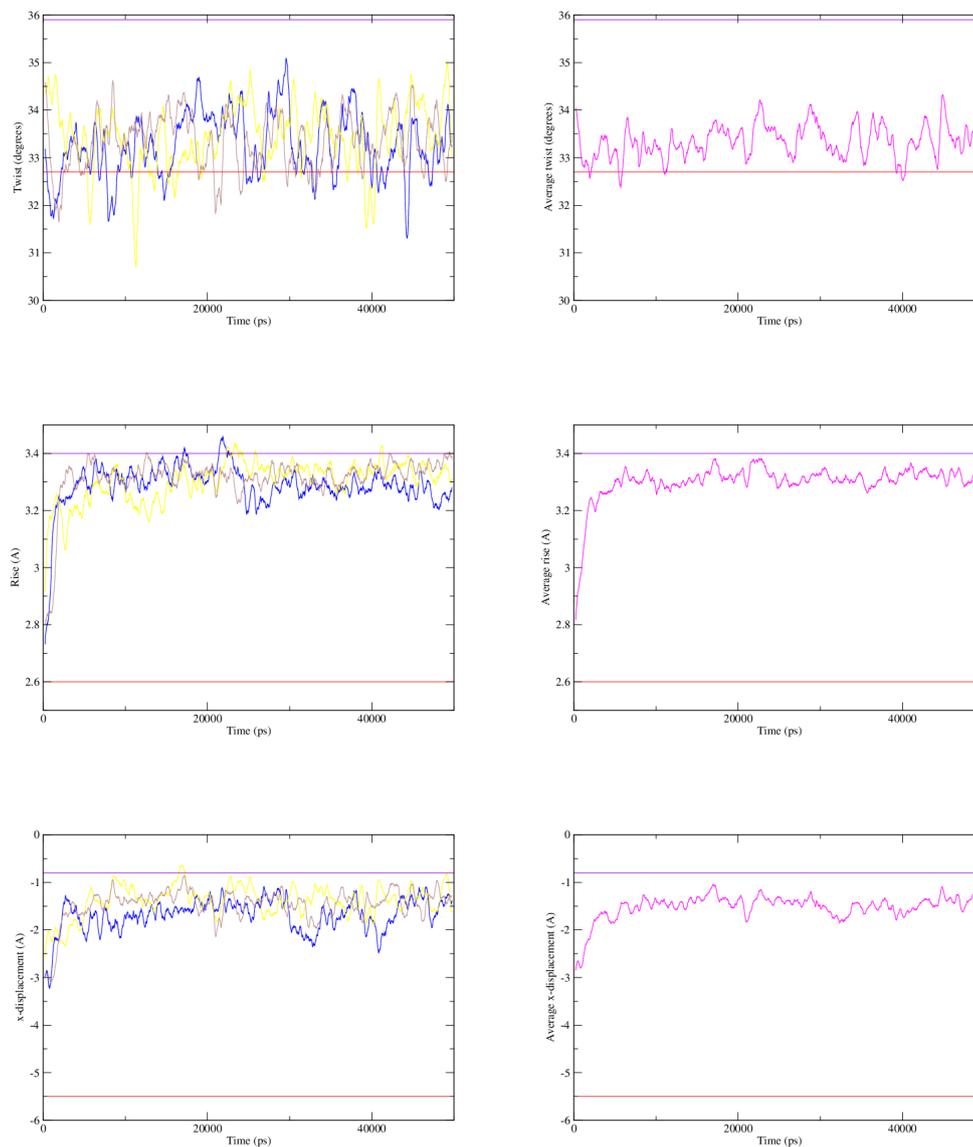
Figure 2.8: Conformational parameters for the three spermine case, with their
average separately plotted.
A- and B-form values are shown by the red and violet straight lines, respectively

2.4.3 Spermine Diffusing into DNA Sequences

For the short simulations which deal with the case of a spermine molecule
diffusing into a length of DNA with variable sequence, a detailed analysis of the
time-dependence of conformational parameters is not warranted: during the
timetaken for the spermine to diffuse, the intial transition between states will
already have occurred, and there will be little left which can be discerned by
eye.

To be able to sensibly discuss the phenomena simulated, therefore, analysis
will be limited to observing the root mean square displacement graphs, and
average structure analysis. However, for a complete picture of the simulations,
the RMSd graphs between the DNA and canonical forms will be presented.

2.4.3.1 Spermine-free Systems

Once again, each system was begun in the canonical A-form. The root mean
square data for the poly-d(A)$_{20}$.poly-d(T)$_{20}$ is in Figure 2.2. The features shown
in Figure 2.9 show the predicted features: in the guanine-rich simulations, the A-
form was preferred, whilst the poly-d(AT)$_{20}$.poly-d(AT)$_{20}$ simulations exhibited
the oscillation between preferred configurations which is characteristic of the
sequence. Further analysis is delayed to the average structure discussion in
Section 2.4.4.
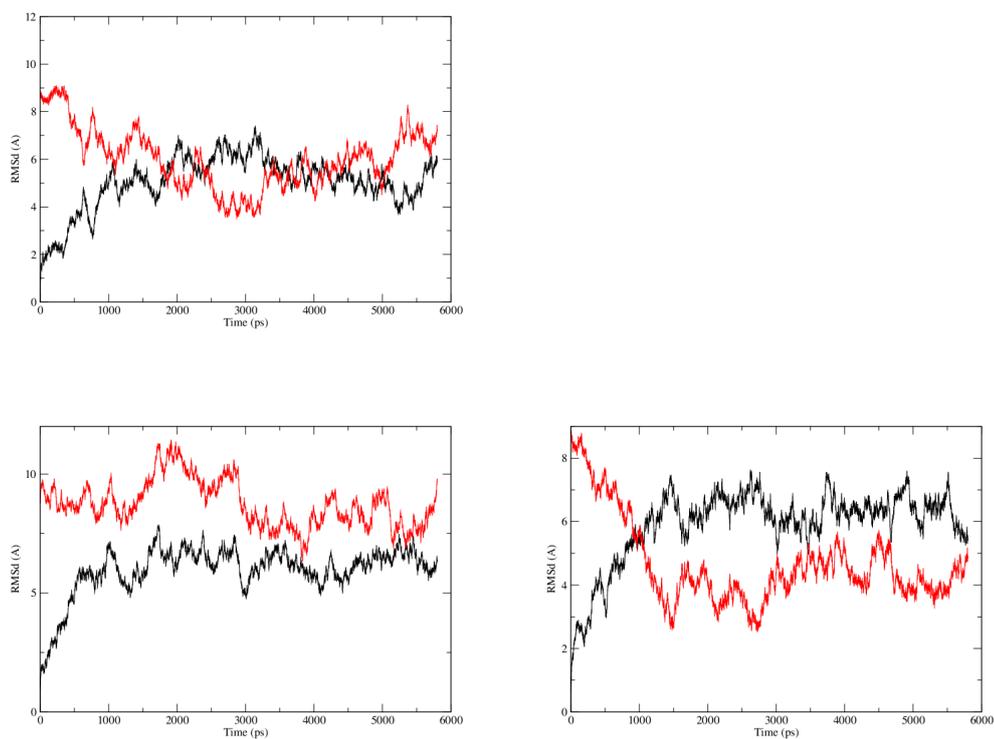
Figure 2.9: Root mean square deviations from A-form (black) and B-form (red) for the cases
Top: poly-d(AT)$_{20}$.poly-d(AT)$_{20}$
Bottom Left: poly-d(G)$_{20}$.poly-d(C)$_{20}$ and
Bottom Right: poly-d(GC)$_{20}$.poly-d(GC)$_{20}$
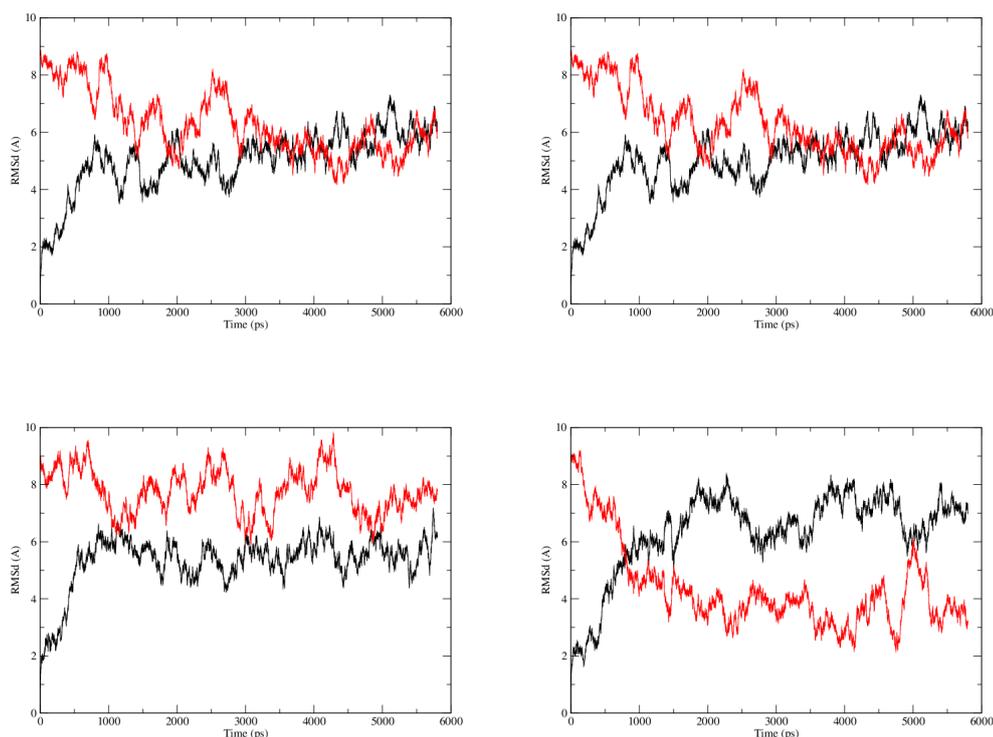
Figure 2.10: Root mean square displacement from the canonical A and B forms of the DNA backbone (black and red, respectively) for spermine diffusing into
a) poly-d(A)$_{20}$.poly-d(T)$_{20}$
b) poly-d(AT)$_{20}$.poly-d(AT)$_{20}$
c) poly-d(G)$_{20}$.poly-d(C)$_{20}$
d) poly-d(GC)$_{20}$.poly-d(GC)$_{20}$

2.4.3.2 Systems with Spermine

Again from an A-form start, in these simulations spermine was allowed to diffuse into the DNA from a short distance away. The graphs in Figure 2.10 represent the behaviour that has previously been seen. The adenine-rich simulations tend towards the B-form, with the poly-d(AT)$_{20}$.poly-d(AT)$_{20}$ oscillating as has been seen before. The poly-d(G)$_{20}$.poly-d(C)$_{20}$ remains closer to the A-form than the B-form, whilst a strand of alternating G and C bases becomes more B-like as time goes on.

2.4.4 The Long-Term Effects of Spermine

In order to better assess the precise changes that spermine - whether major groove-bound, or diffusing from a distance - had on the DNA, the DNA in each simulation was averaged using cpptraj. The average structure having been

System	Avg twist(°)	Avg Rise(Å)	Avg x-disp(Å)
poly-d(A)$_{20}$.poly-d(T)$_{20}$ (50ns)	29.7	2.90	-0.86
1 major groove spermine (run 1)	33.0	3.36	-1.32
1 major groove spermine (run 2)	33.1	3.36	-1.27
1 major groove spermine (run 3)	33.2	3.36	-1.28
2 major groove spermines (run 1)	30.7	3.34	-1.40
2 major groove spermines (run 2)	33.1	3.37	-1.22
2 major groove spermines (run 3)	34.0	3.26	-1.01
3 major groove spermines (run 1)	33.7	3.32	-0.87
3 major groove spermines (run 2)	33.3	3.37	-1.21
3 major groove spermines (run 3)	33.0	3.36	-1.27
poly-d(AT)$_{20}$	31.9	3.12	-2.06
poly-d(G)$_{20}$	29.4	3.11	-3.69
poly-d(GC)$_{20}$	33.4	3.25	-0.99
poly-d(A)$_{20}$ + spermine diffusing	33.5	3.35	-1.17
poly-d(A-T)$_{20}$ + spermine diffusing	31.8	3.11	-2.07
poly-d(G)$_{20}$ + spermine diffusing	30.0	3.14	-3.34
poly-d(G-C)$_{20}$ + spermine diffusing	33.6	3.28	-0.64

Table 2.2: Average conformational parameters for all simulations
To save space, the sequence of only one strand is reported in the first column.
It is to be understood that the second strand carried the complement of the
reported strand.

prepared as a pdb file, it could then be analysed using Curves+[49]. The value
of each parameter at each base pair was averaged, and so one number was
reached for the average value of the time-averaged values at each base. The
relevant parameters are given in Table 2.2.

It can be seen that there is little difference in the average values of parameters
in the majority of cases of spermine diffusing in from a distance. However, the
average values were significantly different for the cases where spermine was
initially to be found in the major groove, suggesting either that the placement
of spermine was a factor in how it could interact with the DNA on a long
term basis, or that the sequence was favourable to interactions. In the case of
the spermine diffusing towards the DNA from a distance, the poly-d(A)$_{20}$.poly-
d(T)$_{20}$ had the strongest response, suggesting that this sequence is particularly
susceptible to conformational change thanks to spermine compared to the other
sequences.

2.5 Analysis of AMBER Simulations

It is easily seen when looking at the Figures in Section 2.4 that the spermine has had a profound effect on the structure of the DNA as it changes conformation. Comparing various parameters as they change through time shows the spermine inhibiting the DNA's A-B transition, and this effect growing in strength as the number of spermines in the major groove rises.

These graphs also show the backbone structure deviating from both A- and B-form for some systems. Whilst a remarkable effect, and one that is seen on multiple graphs, only twice does the effect persist or lead to a noticeably large change from the no-spermine case. This indicates that the DNA response causing this behaviour comes about from a certain positioning of spermine - e.g. a specific binding site which is not always interacted with. Confirming this via more detailed analysis of the trajectories would be greatly of interest.

It is impossible from the graphs of average parameter values through time to be sure whether the spermine leads to a change in conformation along the whole of the DNA strand, or whether its effect is short-range and local - the root mean square displacement is a broad measure, which gives no information as to which atoms are contributing most to it. However, by studying the average structure, a small insight can be gleaned. The spermine diffusing in did not have a striking effect on the average conformational parameters in the simulations performed, except for the sequence poly-d(A)$_{20}$.poly-d(T)$_{20}$, indicating that the effect of the spermine is in some way dependent on the DNA sequence. In other cases, the effect is either negligible, or it is averaged out. If the effect has been averaged out so that it is barely noticable, that suggests that the interaction with spermine has changed the DNA only on a local level - had the induced conformational changes been wide-ranging, a difference would surely be seen. A more detailed account of the exact interactions between spermine and DNA atoms would help to resolve this question, whilst extending that survey to include the bend of the DNA axis (both on average and local to the spermine), sugar puckers, and y-displacement may give greater insight into how the interaction works.

3 ONETEP

3.1 A Brief History of ONETEP

Whilst density functional theory (DFT) enjoyed considerable success in its earliest formulation, it has always been limited by the amount of time taken to perform calculations, which traditionally increases with the cube of the number of atoms. Because of this, one of the most sought-after techniques in solid state physics is a DFT method which will scale more favourably - ideally linearly.

ONETEP (order-N electronic total energy package) is a recent development towards this goal. First introduced in its current form in 2005 [51], ONETEP came out of a group based at the University of Cambridge, who were also responsible for the traditional DFT code CASTEP. Motivated by a desire to approach biophysical problems from a quantum mechanical perspective, the group reformulated the DFT approach in order to take advantage of a key property of these systems: their band gap. In insulating systems, the electron localisation's effects on long-range interactions can be exploited. As the interaction between well-separated electrons rapidly drops to zero, an interaction cut-off can be imposed, greatly reducing the number of calculations which need to be performed. By careful choice of cut-off distances, a linear scheme can be reached, which - assuming the system is indeed an insulator - has the same accuracy as traditional DFT, opening the door to highly efficient *ab initio* calculations for systems of unprecedented size.

3.2 The Rough Guide to DFT

DFT is based entirely on one theorem: the Hohenberg-Kohn theorem. This great piece of quantum mechanical insight was that rather than describe a system using the electronic wavefunctions, the charge density could instead be used, and by finding the ground-state charge density, the ground-state energy would follow, as the mapping between them is unique. Simply minimising a functional of the density will give the ground state energy, a considerably more straightforward task than iteratively solving the time-independent Schrödinger equation.

Generally, we wish to minimise

$$E[n] = \int \psi^* \widehat{H} \psi d^{3N} r \tag{3.1}$$

with respect to the wavefunctions which give the charge density. For N electrons, the density must, of course, satisfy

$$\int n(r) d^3 \vec{r} = N \tag{3.2}$$

that is, integrating the charge density over all space must give the total number of electrons in the system.

The Hamiltonian may be split into two terms by the familiar application of the Born-Oppenheimer approximation, which, thanks to their size, allows treatment of the nuclei as classical potentials. This yields

$$E[n] = \int \psi^* \widehat{H_0} \psi d^{3N} r + \int V_{ext}(\vec{r}) n(\vec{r}) d^3 \vec{r} \tag{3.3}$$

i.e. a set of non-interacting electrons in a homogeneous electron gas, moving in an external potential thanks to the ions of the system. The first of the terms on the right hand side depends only on the electrons, and so can be rewritten as a functional of them alone. Sadly, the form of this functional is unknown, so as a first approximation is described by three terms:

$$\int \psi^* \widehat{H_0} \psi d^{3N} r = F[n] = T[n] + \frac{1}{2} \int \int n(\vec{r}') \frac{1}{|\vec{r} - \vec{r}'|} n(\vec{r}) d^3 r' d^3 r + E_{xc}[n] \tag{3.4}$$

which, left to right, represents the kinetic energy of the charge distribution, the Coulomb interactions between electrons, and the electron exchange and correlation effects. Once again, we are presented with unknown terms. The kinetic energy of a charge distribution is not well defined, and the form of the exchange-correlation functional cannot be analytically derived. Clearly, then, two further methods must be employed.

The first is to describe the density as a set of wavefunctions, where each wavefunction represents an electron, and they do not interact. These wavefunctions are chosen purely to describe the ground-state charge density, and do not have any other physical meaning. Mathematically,

$$n(\vec{r}) = \sum_{i=1}^{N} |\psi_i(\vec{r})|^2 \tag{3.5}$$

As is usual in quantum mechanics, the kinetic energy may be found via

$$T = -\frac{1}{2} \sum_{i=1}^{N} \nabla^2 \psi_i \tag{3.6}$$

where we have used a system of units in which

$$\frac{\hbar^2}{m} = 1$$

Substituting the expressions for the density and kinetic energy into that for the energy, we obtain

$$E_i \psi_i = \left\{ -\frac{1}{2} \nabla^2 + V_{eff} \right\} \psi_i \tag{3.7}$$

and V_{eff}

$$V_{eff} = \int n(\vec{r'}) \frac{1}{|\vec{r} - \vec{r'}|} n(\vec{r}) d^3 r' + V_{xc}(\vec{r}) + V_{ext}(\vec{r}) \tag{3.8}$$

All of the terms in V_{eff} are known except for the exchange-correlation. This must be approximated from other calculations, using other methods. Once the appropriate V_{xc} has been obtained, it is trivial to expand the wavefunctions as a sum of known basis functions with unknown coefficients. This turns the Schrödinger equation into a matrix problem, the solution to which may be found with standard methods.

3.2.1 DFT on a Budget: Linear Scaling

The treatment of DFT above is not, however, particularly amenable to scaling, and so an alternative to the sets of equations produced by the traditional theory must be found. In ONETEP, the solution is to use instead density matrices[51] and the localisation properties of an insulator to reduce the amount of data stored and operated upon (interactions beyond a certain distance need not be calculated), minimising the computationally expensive operations necessary to perform the calculation. The number of functions used to describe each atom - known as non-rthgonal generalised Wannier functions - is also minimised, and can be as low as the number of valence electrons on the atom. Unlike other linear-scaling DFT codes, these atomic functions are calculated at run-time rather than making use of pre-existing values. This increases the computational expense, but also the stability of the code. During the charge density optimisation, the locality of the NGWFs is enforced, and any effects further than the NGWF radius are meticulously weeded out.

3.3 Systems Studied and Methods

In order to compare the two codes as rigorously as possible, it was necessary to ensure that as far as possible each calculation was done under the same conditions. This included as a primary concern the solvation of the system: using explicit solvent in ONETEP and TIP3P water in AMBER would make the systems inherently different, affecting the results. Without data with which to take account of the effects of the different water models, the results from these calculations would then be incomparable - there would be no way to compare purely ff10 with ONETEP's *ab initio* method. In this section, the means of ensuring the calculations were as similar as possible - and therefore that the results were as reliable as possible - will be explained.

3.3.1 Systems Studied

The starting point for the ONETEP comparison with AMBER was the set of short (5.8ns) simulations of various base-pair sequences described in Section 2.3. Using cpptraj, ten equally spaced snapshots from each simulation were extracted in the pdb file format. These snapshots were then prepared as input files for both codes, and calculations performed to extract physical properties. However, each simulation was run in TIP3P water, which would not be possible to use in ONETEP. Therefore, the water was stripped from the simulation cell, leaving only the DNA, counterions, and spermine if present.

3.3.2 ONETEP Set-Up

3.3.2.1 Generating ONETEP Input

Input for ONETEP is done by means of a single file, which contains the atomic species, positions, models to use for each atomic type, and various parameters for the calculations. Creating this by hand from scratch would be virtually impossible - typing the co-ordinates of each atom would be prone to errors. Fortunately, a ONETEP utility called pdb2dat generates input files from a pdb automatically. It does, however, have some issues: the files produced have some parameters pre-set, and these must be changed in order to achieve a well-converged, reliable calculation. These parameters are discussed in depth in section 3.4.

3.3.2.2 Pseudopotentials

In a DFT calculation, the valence electrons are the only ones of interest: it is valence electrons which determine whether a system insulates or conducts electricity, form chemical bonds, and interact with each other. Thus, the inner electron shells and the nucleus are of lesser importance, and, given that they are relatively inert, can be combined into an ionic potential with which the valence electrons will interact. These are calculated quantum mechanically specifically for DFT purposes, and give the potential energy due to the ion at positions around the ion, contributing to the V_{ext} term in the Hamiltonian. In the calculations presented here, the pseudopotentials used were H_00.recpot, Na_00.recpot, N_00.recpot, P_00.recpot, O_00.recpot, and C_00.recpot, pseudopotentials often used in CASTEP calculations.

3.3.2.3 Solution of the Solvation Problem

Water is indisputably important in DNA interactions, but it is not always feasible to accurately simulate. As previously discussed, there is no way to solvate the systems in a way which is identical across both codes. TIP3P vs explicit water, or various implicit solvent techniques in AMBER vs the implicit solvent in ONETEP are both different techniques, and will lead to different results. As such, the only way to ensure fully comparable results will be to exclude solvation entirely. Whilst physiologically incorrect, as it is the relative forces and energies within the system that are of interest - and they are being calculated for single frames rather than being simulated over time - this situation is acceptable for the purposes of comparing codes. The forces on atoms will be higher without the screening effects of water, but both codes will be comparable, and as an additional benefit the ONETEP calculations will be less computationally expensive as there are fewer atoms and terms to take account of.

3.3.3 Performing AMBER Calculations

Each pdb snapshot given by cpptraj was used to create both co-ordinate and topology files using tleap. Molecular dynamics was then performed on the snapshots at 0K, that is, simply allowing the atoms to move slightly thanks to the electrostatic interaction between them. The molecular dynamics was performed over a single timestep of 0.1zs (10^{-22}s), which meant that in practice there was no difference between the starting and ending configuration of the system. The ouput file for the single MD step included the energies of the system, and the input file was modified so sander produced an extra ouput file

containing the forces on each of the atoms. The energies and forces could then be extracted from the files, and compared to those from ONETEP.

3.3.4 Making Results Comparable

Because each code was written by different groups, and aimed at different groups of scientists, different units are in use. In ONETEP, the standard units of energy and length are the hartree and bohr, respectively, whilst for AMBER, kilocalorie per mole and angstrom are used. This means that an extra stage must be gone through prior to graphs being drawn or conclusions made: unit conversion. In this thesis, both codes' output will be stated in ONETEP units, Ha, a_0 and Ha/a_0.

3.4 Convergence

3.4.1 The Importance of Convergence Testing

Several parameters in the ONTEP input file represent physical quantities which have a diminishing effect over time, for example cutoff energy. As the cutoff energy is increased, the basis set describing the system is increased. Eventually, we reach a value above which additional terms in the basis set no longer change the calculation, as there the system is already well-described. However, a further increase would still have a detrimental effect on the runtime of the program, so it is desirable to find the value which gives the most appropriate mix of accuracy and efficiency.

Generally, a convergence parameter can be thought of as representing a convergent series. As we increase the number of terms in the series, the total will tend towards an asymptote. There will come a time when increasing the terms in the series has no appreciable effect on the total, or the change in the total is so small that it may be ignored. When this is reached, the calculation is said to have converged, and when all the convergence parameters have been satisfactorily assigned a value, the calculation is trustworthy. Under-converged calculations, which do not take into account all the significant interactions between atoms, do not describe the system fully. The physical properties extracted from such a calculation are therefore unreliable and should not be thought of as a good model of the system.

In the ONETEP calculations performed, there were two major quantities which needed to be converged with respect to: cutoff energy and kernel cutoff. As was previously stated, the cutoff energy limits the size of the basis set used when describing the system, and the kernel cutoff limits the range of the charge

density kernel. The range of the density kernel describes a sphere around each ion within which the electrons will interact with it, and so effectively gives the longest-range interaction which will be calculated. As was discussed previously, the DNA/spermine complex is an insulator, and the band gap leads to rapidly declining long-range effects. Testing convergence with respect to kernel cutoff will give the distance beyond which all interactions have decayed suitably, such that calculating them is no longer worthwhile.

The procedure for convergence testing is very straightforward. Many calculations of the same system are performed, in which everything is the same bar the value of the parameter with respect to which you want to converge the calculation. A certain property of the system as calculated by the test run is then noted, and the values for each test plotted. It is then visible where the property's value stop changing, and so a suitable value for the chosen parameter can be decided upon. Either total energy of the system or a physical property may be used as a convergence criterion, or a mixture of the two. In these convergence tests, the total energy of the system was primarily looked at. This was then verified by examining forces in the final converged calculations.

It can be important to converge parameters in the correct order, if the parameters depend on each other in any way, or if the initial guess for a parameter may not be a good starting point - the starting guess for parameters may be well converged, or not converged enough. Were a parameter to be converged when another parameter was not fully converged, the resulting final calculation may not fully capture all of a system's properties. However, for some parameters, an initial guess can be used such that it is effectively infinite. This was the approach taken for the kernel cutoff: 1000 bohr was used as the initial guess, a value which was guaranteed to be far in advance of what was necessary, not least because 1000 bohr is larger than the test system's simulation cell. Having converged the cutoff energy whilst using this effectively infinite kernel cutoff, a value for cutoff energy could be set. This was then used whilst performing test calculations to decide on the correct value of the kernel cutoff.

Finally, a balance between convergence and runtime must be found. In an ideal world, every parameter would be set to be virtually infinite, but this is clearly infeasible as calculations would then run forever. It is therefore of great importance to decide before starting convergence tests. In these systems, typical total energy was of the order -8000 hartree. A suitable convergence tolerance in the total energy was around 1 hartree, giving an error of less than 0.0125% whilst keeping calculation times relatively short.
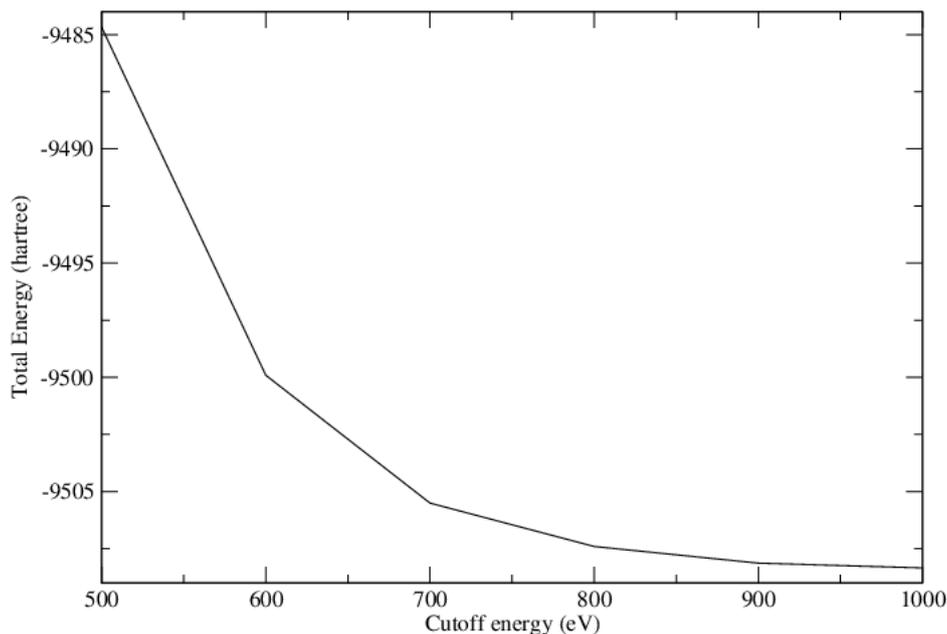
Figure 3.1: Total energy with increasing cutoff energy

3.4.2 Converged Parameters

3.4.2.1 Cutoff Energy

As previously said, the first parameter which was converged was the cutoff energy. Figure 3.1 shows the graph produced as the cutoff energy was raised from 500eV to 1000eV in steps of 100eV.

It can be seen that at 800eV, the total energy is within 1 hartree of its final value at 1keV. Moreover, further testing with a higher kernel cutoff indicated that there were severe memory implications for using 900 or 1000eV, so that the standard memory nodes of ARCHER could not perform the calculations on certain systems, and when run the calculations took an unacceptably long time. Considering the minimal extra precision found at a cost of considerable time and resources, 800eV was determined to be the best value to use over each system.

3.4.2.2 Kernel Cutoff

Having found a good value for the cutoff energy, the kernel cutoff was next to be examined. Once again, a variety of values were used, and the results graphed.

In this instance, less subtlety was needed when looking at the graph. It is

Figure 3.2: Total energy with increasing kernel cutoff

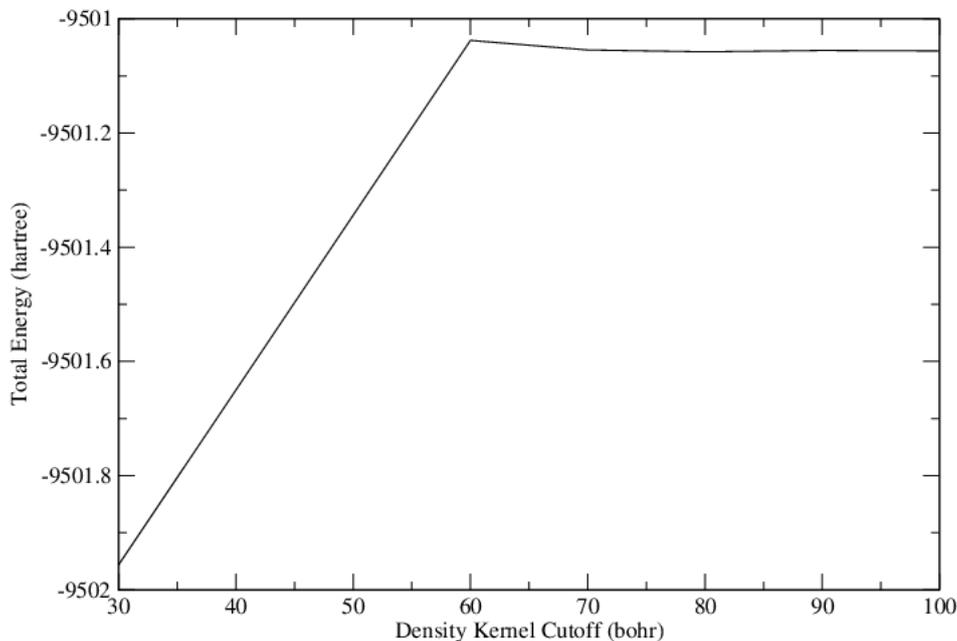immediately apparent that the system has converged at a kernel cutoff of 70 bohr, and has converged to an excellent level of accuracy. As the lowest value at which the converged value of energy had been reached, 70 bohr was the kernel cutoff used in all production calculations.

3.4.3 Checking Convergence with Forces

In all areas of computational physics, it is important never to become complacent in trusting results. An error in calculation may be glaringly apparent when analysing data through one method, but hidden by other factors in another. Total energy was not, therefore, the only means by which the calculations' convergence was tested. Observable quantities are of prime importance in physics, so an analysis based on these rather than the somewhat abstract notion of total energy is both a more meaningful test and a means of being more confident in the later analysis of the data. In these systems, the most relevant physical quantity which will be examined is force, so that is the observable which will be tested for convergence.

Once again, a convergence graph was produced, for the case of the cutoff energy increasing, using the kernel cutoff of 70 bohr. However, a straight mean or total of force is the wrong measure to use. In any system, the sum of the
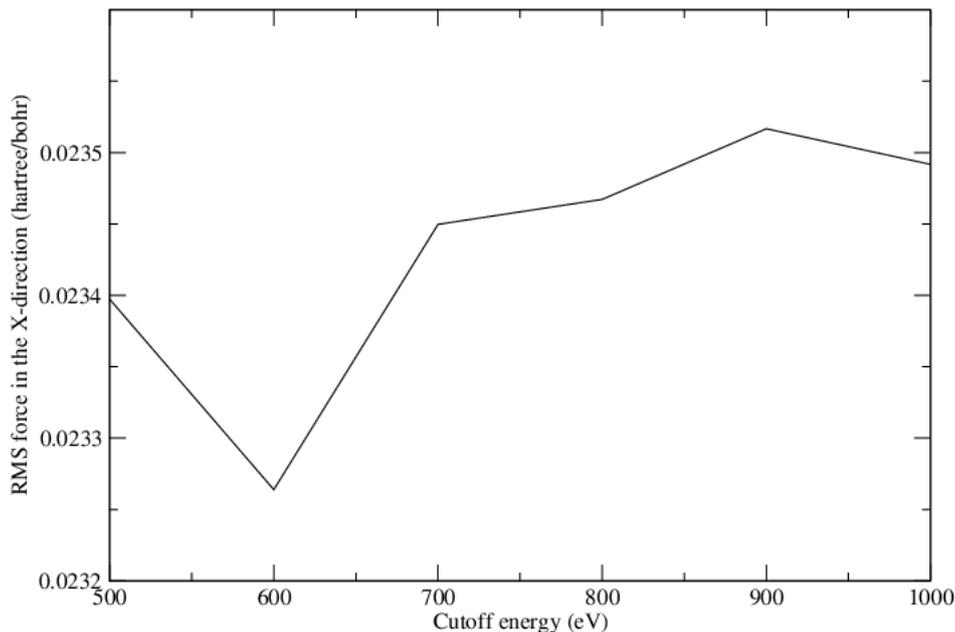
Figure 3.3: RMS force along x as cutoff energy increases

forces is necessarily zero due to Newton's laws, and so if total force were to be plotted, one would only see the line y=0. For that reason, root mean square force along the X-axis was calculated and plotted.

It can be seen in Figure 3.3 that at 800eV, the RMS force has reached a value very close to that at 1000eV - close enough in fact that the compromise between performance and accuracy described previously still holds true, and 800eV is the value to be used.

3.5 Results

3.5.1 Comparison of Energies Reported by AMBER and ONETEP

While spermine's effects on DNA have been well documented elsewhere[4] and in Chapter 2 of this thesis, previously the energetics of the systems used were not expounded. In this section, the shorter simulations from Chapter 2 are analysed for their changing total energy, and the effect of spermine on this is seen.

In order to begin comparing the performance of the two approaches, the

total energies calculated by each package were extracted. However, simply plotting the total energy given was not a good measure of the systems or of the programs' performance. This is because each code has a a different approach to calculating the energy, and as such a different baseline against which the energy is compared. An example which illustrates this may be a simulation of the hydrogen atom via two different fictitious methods. In one, the electronic and gravitational energies are calculated, while in the other both the electronic and gravitational energies are once again found, but this time along with the mass-energy. Clearly, both are physically valid descriptions of the system, but they are not directly comparable. In addition, both report their energies in different units, and so a graph of one evolving through time next to a graph of the other evolving through time would be more confusing than enlightening. Aside from converting the units to be the same, the data must also be manipulated so that the quantity graphed should be the same across both techniques. By far the simplest way of doing this is to define the initial state as being of energy zero, and taking this initial energy away from the energy reported as the system evolves. This scheme results in graphs once again of total energy - strictly the change in total energy of the system from its initial value. Assuming both codes to be perfect, this quantity should be calculated to be the same in both.

This is the methodology used in this section - the energy given by AMBER was converted from kilocalories per mole into hartree for easy comparison, and the total energy of each system redefined to give a meaningful quantity to compare amongst the codes. This was performed for simulations both with and without spermine, in order that spermine's effects may be observed.

3.5.1.1 A-Rich DNA

Figure 3.4 shows the total energies reported by AMBER and ONETEP for each of the snapshots taken from the molecular dynamics trajectory. It can be seen that in each case the trend of the line is the same: beginning from a higer-energy state, the DNA over time assumes a more stable - and hence lower energy - configuration. Uniformly, the energy reported by ONETEP is lower than that from AMBER, suggesting that whilst AMBER's treatment of the system results in a physically meaningful simulation,the details it fails to take into account preclude it from being able to report energy very accurately.

Looking at the profile of each line in detail, some interesting features are apparent. Whilst some similar features such as small peaks are visible in both lines, for example in Figure 3.4c, there are many others which are smoothed out in the line produced by ONETEP, suggesting that there are some conformations of the system which AMBER is unable to properly treat, leading to a significant

a)

b)

c)

d)
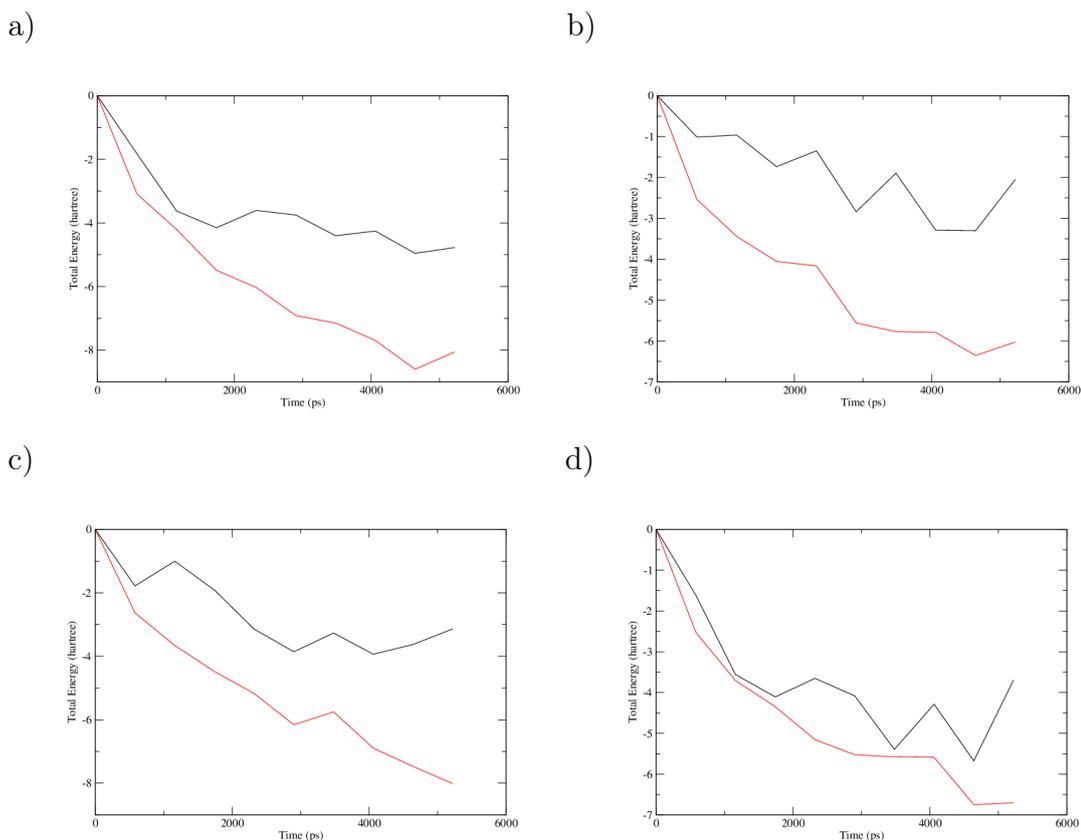
Figure 3.4: Comparison of total energies in AMBER (black) and ONETEP (red) for the systems
a) poly-d(A)$_{20}$.poly-d(T)$_{20}$
b) poly-d(A)$_{20}$.poly-d(T)$_{20}$ and 1 spermine
c) poly-d(A-T)$_{20}$.poly-d(A-T)$_{20}$
d) poly-d(A-T)$_{20}$.poly-d(A-T)$_{20}$ and 1 spermine

effect on the total energy.

3.5.1.2 G-Rich DNA

Once again, the trend for total energy reducing is apparent, which is unsurprising as all systems were begun in an energetically unfavourable state. ONETEP again reports lower energy than does AMBER, but in this case both the codes are more alike in their reportorting of similar features in the profile of the line - i.e. peaks and troughs. This indicates that there may have been more of a problem for AMBER in the A-rich systems, as it was less able to mimic ONETEP. This suggests that there is something in the bases adenine or thymine which is particularly troublesome for AMBER to simulate.

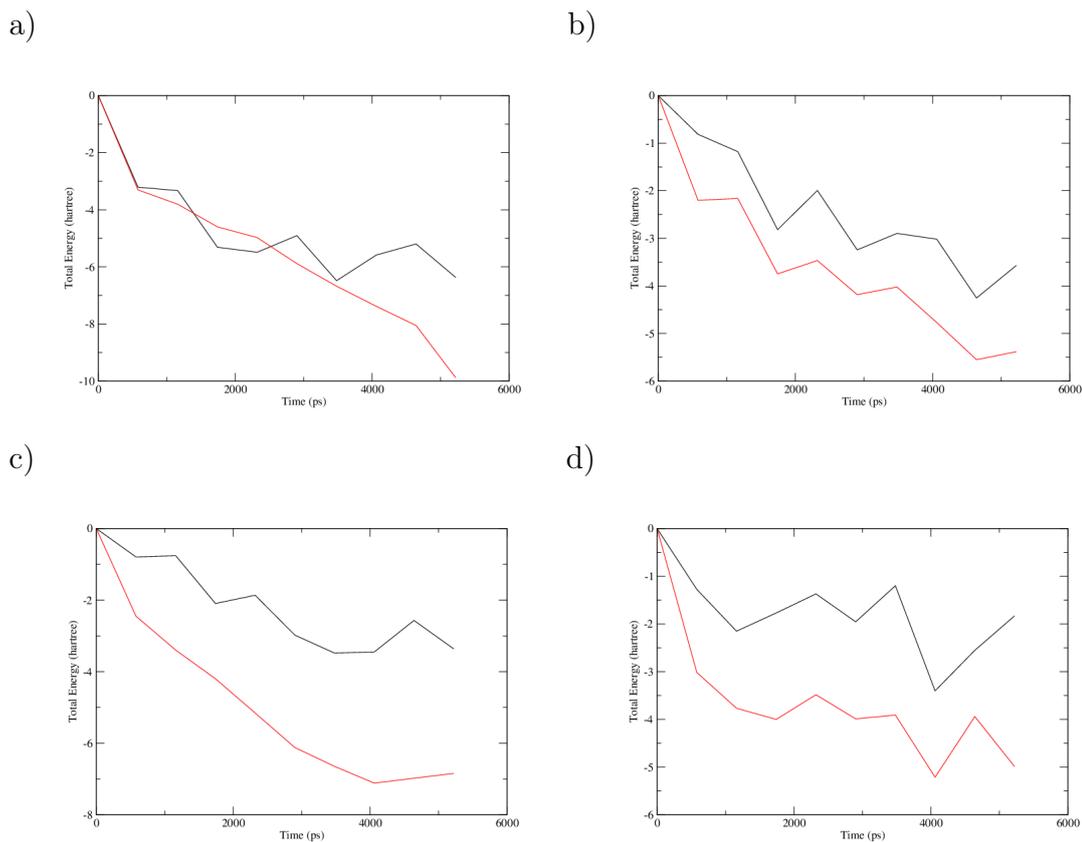Figure 3.5: Comparison of energies in AMBER (black) and ONETEP (red) for
the systems
a) poly-d(G)$_{20}$.poly-d(C)$_{20}$
b) poly-d(G)$_{20}$.poly-d(C)$_{20}$ and 1 spermine
c) poly-d(G-C)$_{20}$.poly-d(G-C)$_{20}$
d) poly-d(G-C)$_{20}$.poly-d(G-C)$_{20}$ and 1 spermine

3.5.2 Forces in AMBER and ONETEP

Having examined the total energy change as a function of time, it remained to find how the difference in energies led to differing values in the observables of the systems. Initially, the root-mean-square force in the x-direction was calculated for each of the frames simulated in both AMBER and ONETEP. These were then compared against each other, giving a measure of how accurate the AMBER dynamics are likely to be: a small difference in force over along time interval may lead to unexpected behaviour. However, RMS force is a very coarse measure of the accuracy of the calculations - some atoms may be disproportionately affecting it by being very poorly parameterised.

To assess the quality of the AMBER forcefield across all the atoms, the treatment of each of the initial frames from the short simulations by both codes were subjected to a more rigorous analysis. For each atom, the RMS total force was calculated according to both AMBER and ONETEP, and the ONETEP value subtracting from its AMBER equivalent, giving ΔF_{RMS}. The absolute value of this quantity ($|\Delta F_{RMS}|$) was then found, and the $|\Delta F_{RMS}|$ for each atom was plotted, to give an indication of any regions which were particularly different. Finally, the mean and standard deviation of the $|\Delta F_{RMS}|$ distribution was found, and a heat-map constructed. For this, each atom was placed at its co-ordinates, and coloured according to how far from the mean its $|\Delta F_{RMS}|$ lay. For any which were within three standard deviations, blue was used. Values between three and four standard deviations from the mean were coloured tan, and those further than four standard deviations from the mean were coloured red, ensuring that any regions which were treated very poorly by AMBER - or any patterns of poor agreement which arose - were easy to see.

3.5.2.1 A-Rich DNA

The RMS force in the x-direction for the simulations of adenine-rich DNA can be seen in Figure 3.6. Clearly, in each case the ONETEP forces are above those given by AMBER: in the case of spermine being included, the forces are approximately 1.5 times those from classical molecular dynamics. In general, it can be seen that each line follows approximately the same form, with peaks and troughs generally coinciding in location if not in magnitude.

Unfortunately, some data was recovered from ONETEP which was significantly different from the previous frames and was therefore unreliable. This data has not been plotted, leading to the shorter red lines in Figure 3.6a and c. The data which is there once again points to the general form of the line being similar in both. As with the spermine-included graphs, the RMS force
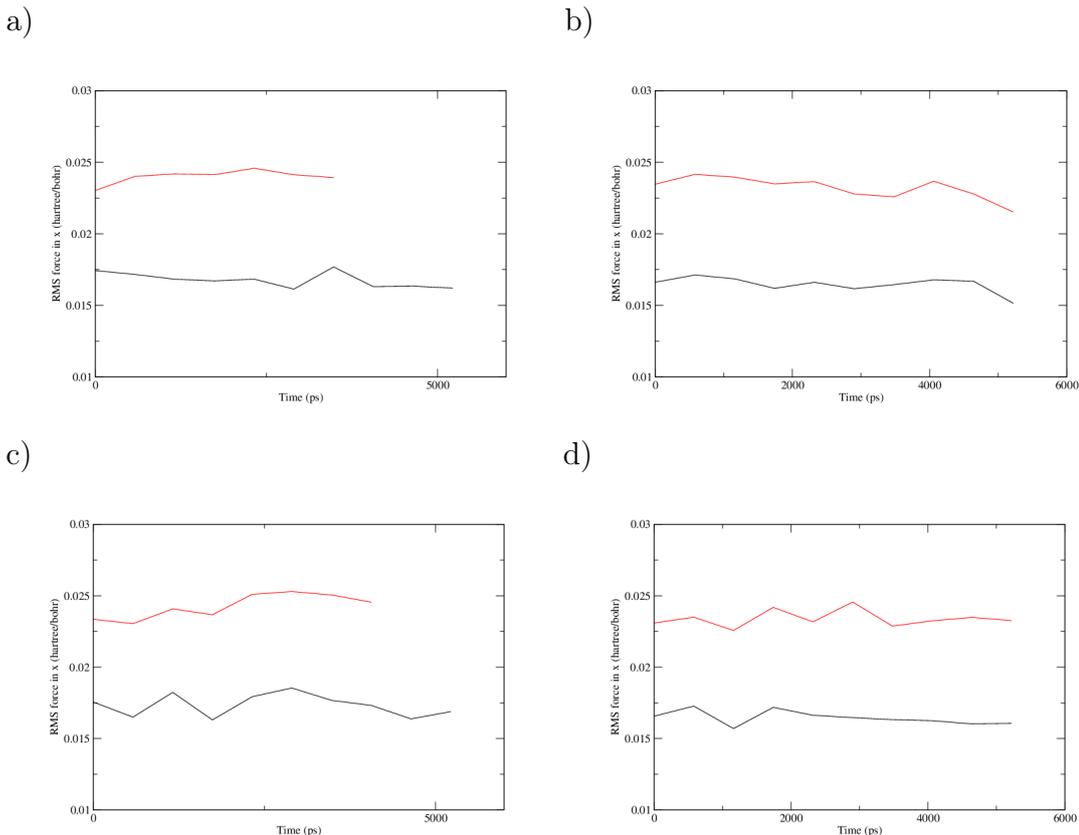
a)   b)

c)   d)

Figure 3.6: RMS force along the x-axis as calculated by AMBER (black) and
ONETEP (red) for the systems:
a) poly-d(A)$_{20}$.poly-d(T)$_{20}$
b) poly-d(A)$_{20}$.poly-d(T)$_{20}$ and 1 spermine
c) poly-d(A-T)$_{20}$.poly-d(A-T)$_{20}$
d) poly-d(A-T)$_{20}$.poly-d(A-T)$_{20}$ and 1 spermine

from AMBER is lower by about 1.5 times than that of ONETEP, suggetsing a
routine underestimation of the strength of forces in the AMBER regime.

Figure 3.7 shows by atom the value of $(|\Delta F_{RMS}|)$. The atom numbering
is done according to the standard AMBER/tleap identification scheme, which
assigns all of the atoms in one strand (from the 3´ to the 5´), followed by all
those in the second strand (again, 3´ to 5´), followed by any extra molecules
(such as spermine, which may be seen as additional fluctuations in Fig 3.7b at
atoms ~1290-1230), followed by counterions and solvation. This is an artefact
of the way in which the systems are produced: atoms are named as they are
added to the system, which in this case was the order Strand 1 - Strand 2 -
spermine (is applicable) - Na+ - TIP3P water. With this knowledge, we can
see that the left hand half - approximately - of the graphs corresponds to the
atoms in the first strand of DNA, and those in the right hand half to the second
strand.

71

In the graphs in Figure 3.7, several peaks considerably away from the mean can be seen at regular intervals. This suggests that it may be specific sites occuring with fixed displacement which are particularly poorly realised in AM-BER. Indeed, by counting the peaks which reach above 0.035, 20 are seen - a very suggestive number, as there are 20 base pairs in the simulations and twenty of each of the bases adenine and thymine. Further, the fact that all of the black spikes are on the right hand side in each graph, while the red spikes are equally distributed. The frequency of the black peaks is also twice that of the red peaks, and all this evidence combines to demonstrate that the regularly occuring sites of poor agreement are to be found in the thymine bases. To be sure of this, the heat-maps for each of the four systems must also be examined, however.

It can be seen from the heat-maps in Figure 3.8 that in the case of poly-d(A)$_{20}$.poly-d(T)$_{20}$, both with and without spermine present, the red (i.e. most different) atoms remain on one strand of the double helix as expected, and remain in the same place on each base on which they occur, which is consistent with the regular peaks seen in Figure 3.7.

Looking at the heat-maps for poly-d(A-T)$_{20}$.poly-d(A-T)$_{20}$, they are very different. The red atoms are no longer on one strand, but can clearly be seen to alternate sides. The bottom three bases in Fig 3.8d plainly show the sites of high discrepancy alternating strand with each base. This is further evidence for the worst described sites in AMBER to be on thymine. By looking at these heat-maps, it may also be seen that the atoms of interest are located in the base itself, close to the hydrogen bond formed with the complementary adenine base, indicating that the specific conditions under which the base remains intact and forming bonds is poorly understood and parameterised for, and may even mean that a separate representation for atoms in such situations needs to be created.

3.5.2.2 G-Rich DNA

As before for adenine-rich DNA, for the G-rich cases a three-stage analysis was carried out. The results of the initial test can be seen in Figure 3.9, the average RMS x-component of force throughout the simulations. As previously, the main features of the graph are similar in both, with the ONETEP forces again higher in each case, and a slightly smoother line profile overall. However, the ONETEP graph is in these instances slightly further away from in the case of A-rich DNA, indicating that G-rich DNA is slightly more poorly described in the ff10 force field.

Once again, some poor data for forces was calculated - again in the snapshots without spermine present. Again, only the reliable data is plotted in Figure 3.9.

a)
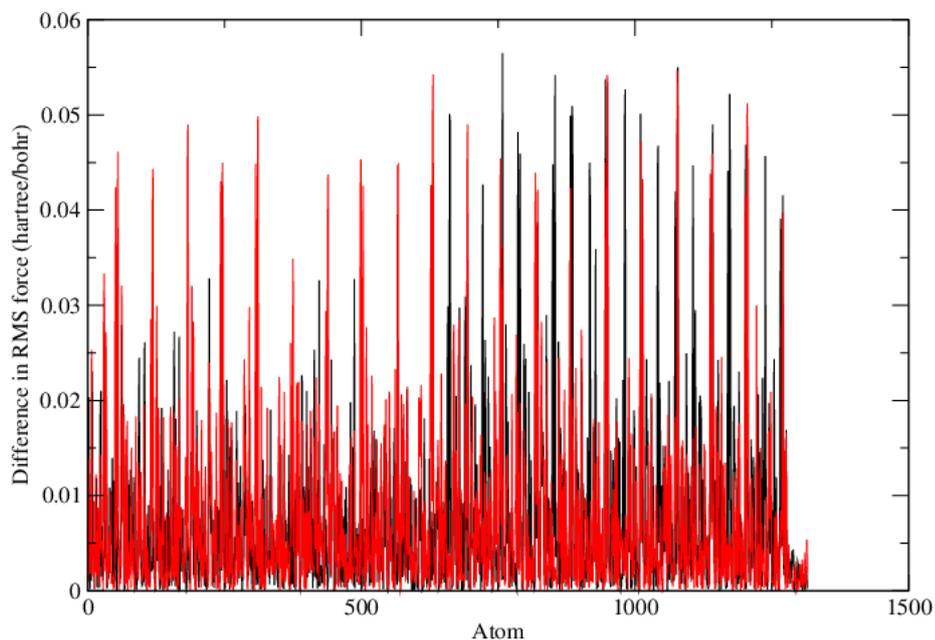
b)

Figure 3.7: Absolute difference between AMBER and ONETEP RMS force on each atom for the systems:
a) poly-d(A)$_{20}$.poly-d(T)$_{20}$ (black) and poly-d(A-T)$_{20}$.poly-d(A-T)$_{20}$ (red) (frame 1)
b) poly-d(A)$_{20}$.poly-d(T)$_{20}$ (black) and poly-d(A-T)$_{20}$.poly-d(A-T)$_{20}$ (red), each with 1 spermine (frame 1)
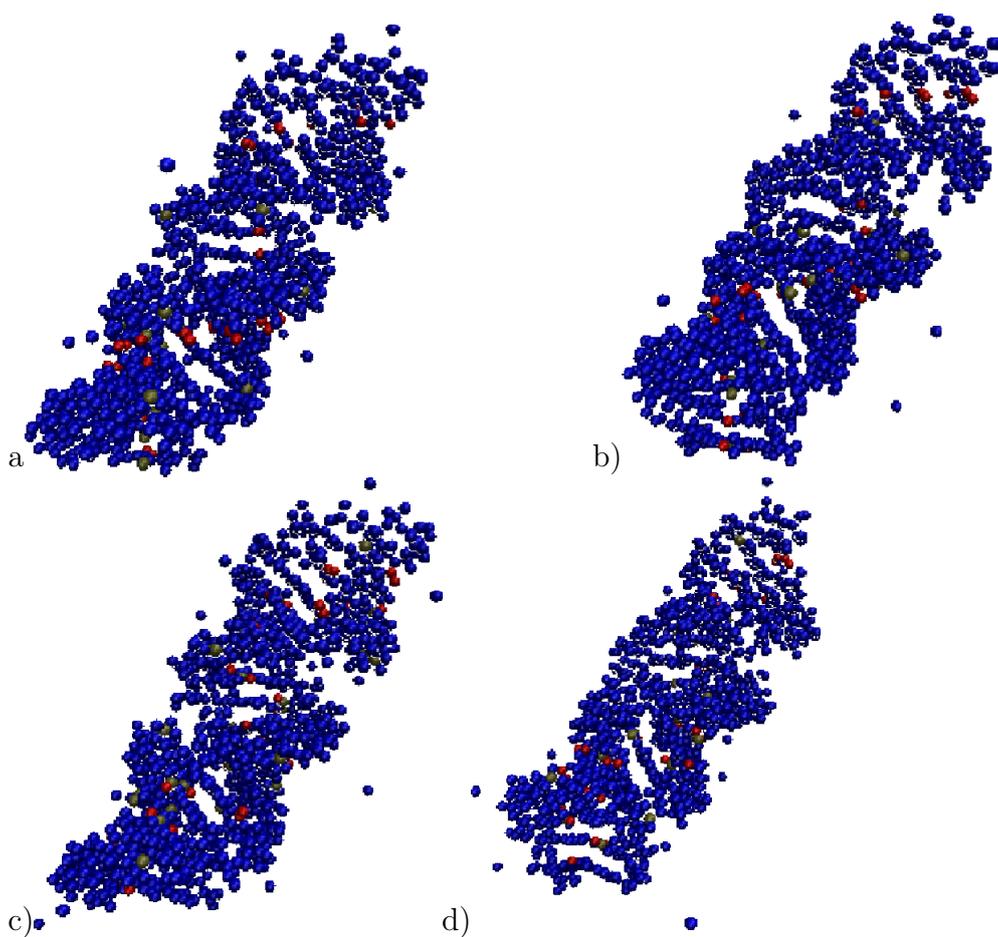
Figure 3.8: Heat-maps for the systems
a)poly-d(A)$_{20}$.poly-d(T)$_{20}$
b)poly-d(A)$_{20}$.poly-d(T)$_{20}$ with one spermine
c)poly-d(A-T)$_{20}$.poly-d(A-T)$_{20}$
d)poly-d(A)$_{20}$.poly-d(T)$_{20}$ with one spermine
Atoms within 3 standard deviations of the mean value are rendered
in blue, between 3 and 4 in tan, and greater than 4 in red.

a)
b)
c)
d)
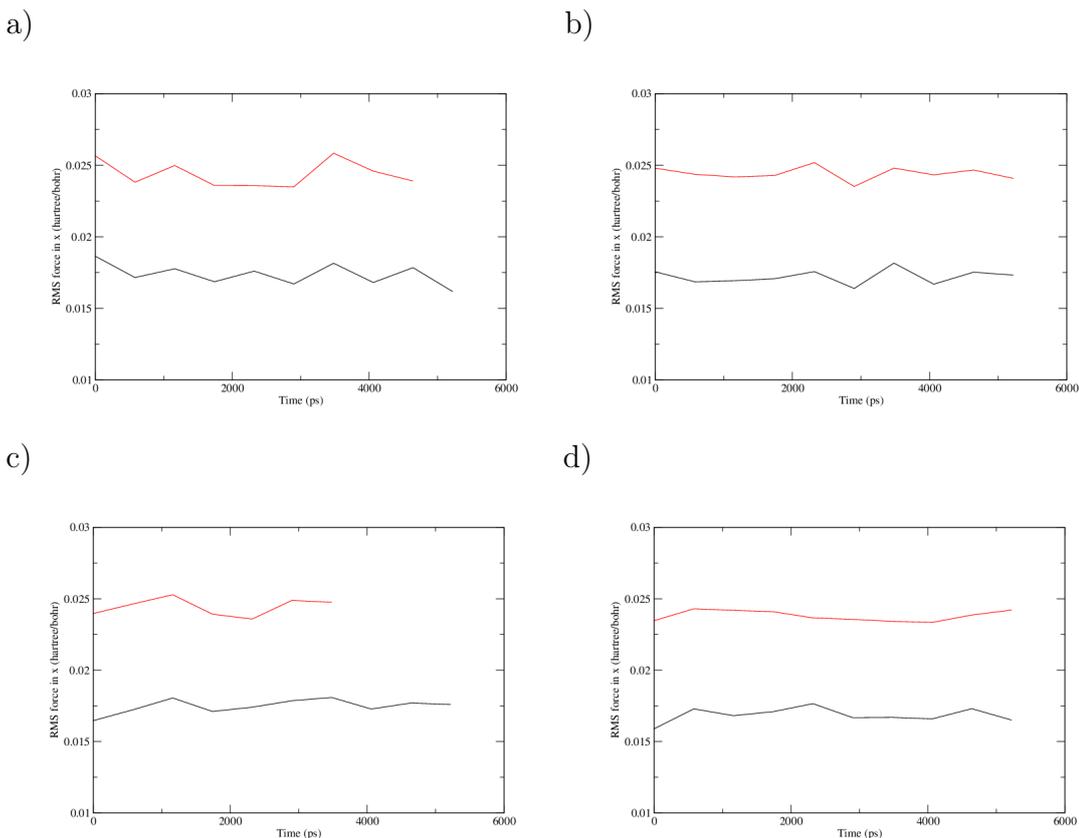
Figure 3.9: RMS force along the x-axis as calculated by AMBER (black) and
ONETEP (red) for the systems:
a) poly-d$(G)_{20}$.poly-d$(C)_{20}$
b) poly-d$(G)_{20}$.poly-d$(C)_{20}$ and 1 spermine
c) poly-d$(G$-$C)_{20}$.poly-d$(G$-$C)_{20}$
d) poly-d$(G$-$C)_{20}$.poly-d$(G$-$C)_{20}$ and 1 spermine

However, the line such as it is shows the same behaviour. Higher forces were
once again calculated by ONETEP, and a similar lineshap was given by both
codes - though sharper peaks are evident in the AMBER data.

Figure 3.10 shows the absolute difference in forces calculated by AMBER and
ONETEP for various systems. Once again, there are many tall peaks above a
noisy mean, but in this case there is more noise and more high peaks, which
are not at regular intervals. Unlike the case of A-rich DNA, no localisation of
peaks to one strand or base is observed - it is the system as a whole which
appears to be more poorly described. However, it is very difficult to gather
any information from the noise-stricken graphs of Figure 3.10 - a colour-coded
image of DNA will be much easier to understand.

Figure 3.11 demonstrates more obviously the information hinted at previ-
ously. In none of the simulated systems was there any ordering or pattern to be
found - just an overall worse representation, including higher numbers of atoms
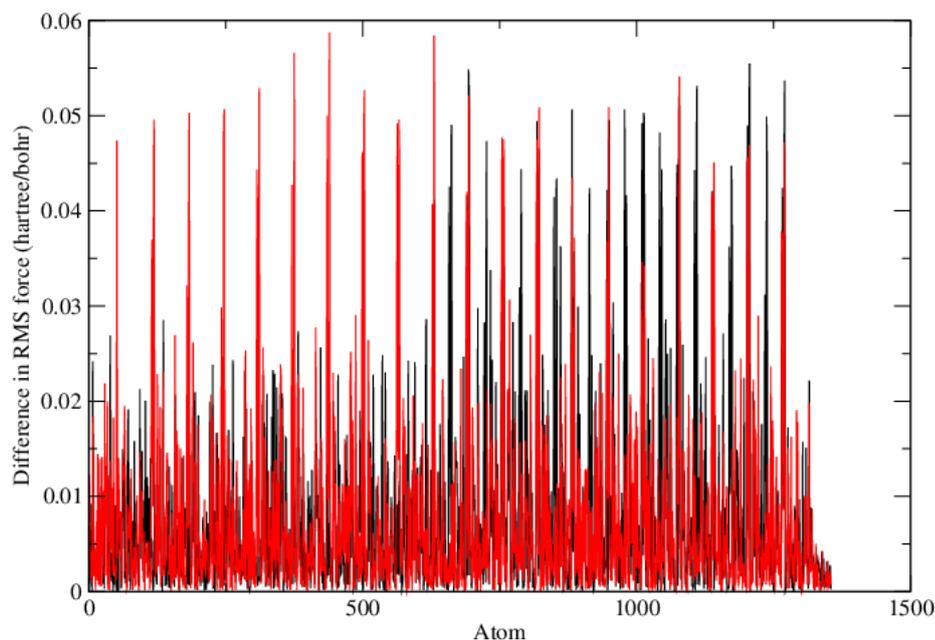
a)
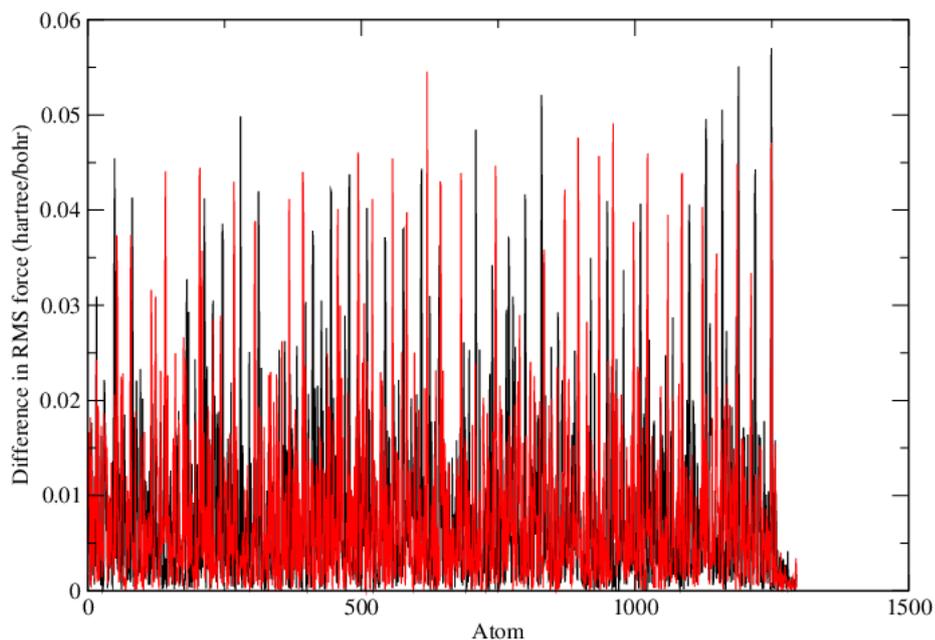
b)
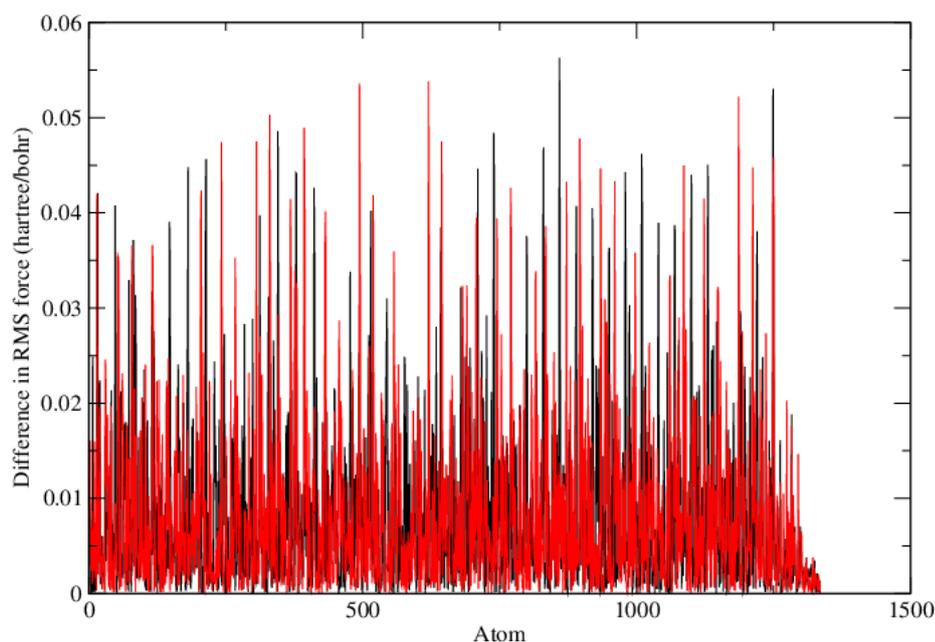
Figure 3.10: Absolute difference between AMBER and ONETEP RMS force on
each atom for the systems:
a) poly-d(G)$_{20}$.poly-d(C)$_{20}$ (black) and poly-d(G-C)$_{20}$.poly-d(G-
C)$_{20}$ (red) (frame 1)
b) poly-d(G)$_{20}$.poly-d(C)$_{20}$ (black) and poly-d(G-C)$_{20}$.poly-d(G-
C)$_{20}$ (red), each with 1 spermine (frame 1)
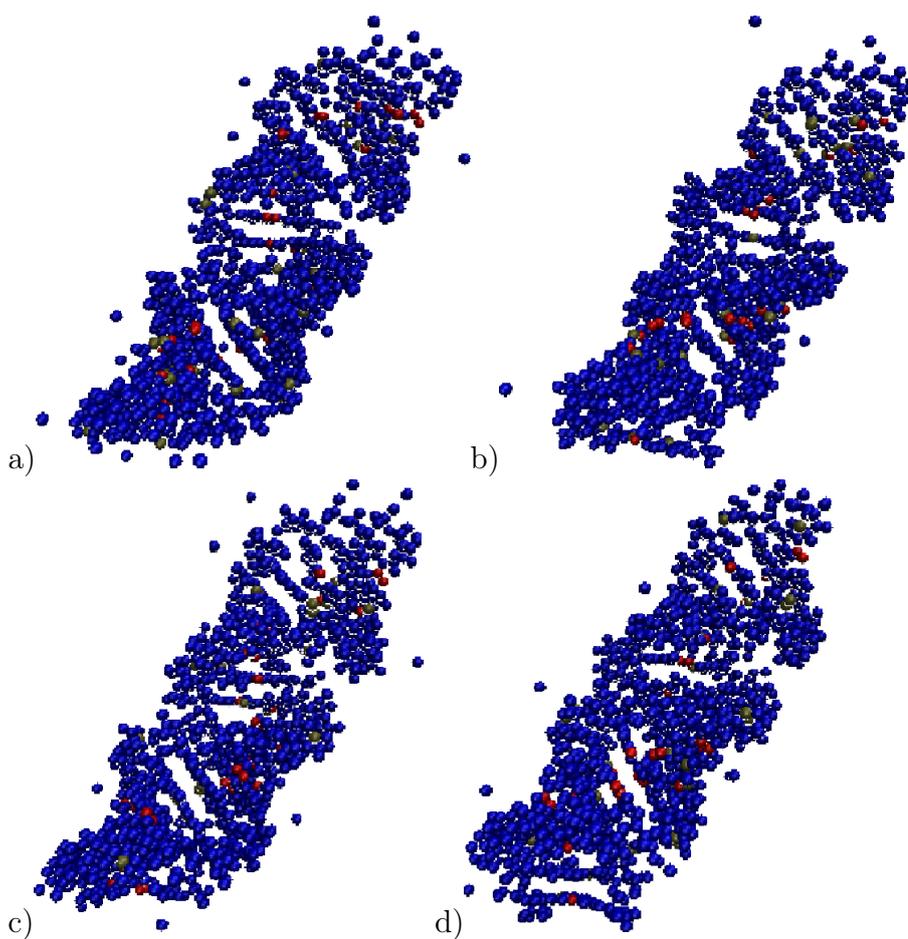
Figure 3.11: Heat-maps for the systems
    a)poly-d(G)$_{20}$.poly-d(C)$_{20}$
    b)poly-d(G)$_{20}$.poly-d(C)$_{20}$ with one spermine
    c)poly-d(G-C)$_{20}$.poly-d(G-C)$_{20}$
    d)poly-d(G-C)$_{20}$.poly-d(G-C)$_{20}$ with one spermine
    Atoms within 3 standard deviations of the mean value are rendered
    in blue, between 3 and 4 in tan, and greater than 4 in red.

with significant differences in their reported forces. However, once again the most atoms with the worst agreement appear to be close to the bond between bases on the two strands, and so further investigation which looked at particular atoms more thoroughly may turn up a greater insight.

3.6 Analysis

The calculations performed by AMBER and ONETEP on these systems has revealed a significant systematic difference between the two codes' treatment of DNA. Initially, it was seen that the total energy of the systems was reported as being lower in ONETEP than in AMBER - that is to say, its magnitude is greater. This undercalculation by AMBER was seen also in the forces: once again, the root mean square force in ONETEP easily outstripped that from AMBER. Generally, the performance of AMBER when looking at guanine-rich DNA was worse than when looking at adenine-rich DNA.

However, what was unexpected was the way in which the discrepancies were distributed throughout the DNA fragment. Rather than specific regions of the double helix being poorly treated (for instance thanks to edge effects), there were specific sites on at least one base - thymine - which stood out considerably from the rest however analysis was performed. In those bases which did not have an obvious single problem atom, it was seen generally that the forces were most likely to be reported exceptionally differently in the region immediately around the inter-strand hydrogen bonding. This suggests that, whilst the AMBER force field is a reasonable approximation, it lacks the subtlety to accurately deal with these small crucial areas, the dynamics of which are likely to be very different to those in - for example - the phosphate backbone, although the exact effects of this poor treatment remains unclear. A systematic review of bases' energies and forces through significant timescales and under varying conditions would need to be undertaken to fully understand if it presents a significant problem; if so, a more differentiated treatment of the atoms in each base may go some way to alleviating it for as long as *ab initio* molecular dynamics on the scales explored here remain prohibitively expensive. In the medium term, a mix of quantum and molecular dynamics techniques could be employed in the simulation of one system, though for that to be possible once again a more in-depth and robust study of the individual bases under a wide range of topological and solvation circumstances would need to be carried out.

Finally, it is heartening to note that the spermine model itself used for the simulations was not particularly poorly characterised by AMBER - by looking at the graphs in Figures 3.7 at around atoms 1290-1330, it can be seen that

the spermine atoms suffer a discrepancy wholly unremarkable when compared to the DNA atoms, making sure that the simulations performed suffered no particularly weak points outside of the areas mentioned.

4 Conclusion

4.1 Spermine

One of the aims of this work was to assess the effect of multiple spermine molecules on the A-B transition of poly-d(A)$_{20}$.poly-d(T)$_{20}$, and this has met with considerable success. It was clear that the presence of the spermine slowed the transition rate in all case, and in some cases novel behaviour was also observed - namely a movement of the DNA away from both the canonical A and B conformations which was only apparent with multiple spermine molecules present, and which decayed rapidly after some time, assuming the A-like structure expected of the DNA. This suggests that the interactions of the spermines with the DNA radically changed the stability of the system, and allowed the DNA to take on an energetically unfavourable conformation which, when the necessary conditions on the location of the spermines were no longer met, collapsed into its preferred state. This procedure was seen only when examining the backbone of the DNA, indicating that it was specifically the backbone atoms which were strongly interacted with. However, the rapid descent from the novel structure to the expected structure suggests that the specific interaction is a product not only of proximity to the backbone, but also the position of the spermine relative to the DNA. The average long-term structure of the DNA - after any unusual conformations had been moved away from - were not significantly affected by the presence of the spermine in the simulation, except for the case of poly-d(A)$_{20}$.poly-d(T)$_{20}$, which is evidence of a sequence-specific interaction. Two explanations for the relatively minor changes produced by spermine in the other sequences can be hypothesised: that the spermine-DNA interactions are effectively removed by the effects of averaging the structure, in which case it could either be that the interactions are very local, or that they are accompanied by a reaction in the DNA so a change in the twist at one base is compensated for by an opposite twist in another; or that the interactions with spermine of other sequences are negligible. One possible approach to determine which of these is correct is found in Section 4.3.

4.2 AMBER and ONETEP

The relative performance of AMBER and ONETEP was as may be expected, but had surprising features. The energies and forces given by ONETEP calculations naturally differed from those from the molecular dynamics approach, and both showed similar time-evolution of the system. However, unexpectedly there were specific sites within the DNA which had consistently large discrepancies between the codes, leading to the conclusion that AMBER forcefields are poorly equipped to treat all atoms in a base pair. Specifically, one site in the base thymine was observed to have consistently poor agreement between the codes, and further investigations into the reasons behind this atom or these atoms in particular being far from agreement would be of interest. In general, the magnitude of the forces were understated in AMBER, which can be seen not only in the graphs of root mean square force, but in the total energy also being of smaller magnitude.

Finding a specific problem site shows the importance of approaching a physical problem from multiple avenues. Checking the performance of AMBER using ONETEP is a powerful tool which has shown areas of specific weakness which would not have been able to be identified through other means. Having identified the site, it would be a clear target for AMBER developers, and would lead hopefully to more robust descriptions of the atoms in DNA. This methodology - refining models based on *ab initio* simulations of the same system - would be a good avenue to improved molecular dynamics force fields, which would benefit all future work on DNA using AMBER.

However, the differences between the codes' assessment of the systems is primarily a feature of their different applications. Whilst ONETEP can offer a more thorough and reliable description of atomic energy levels and electron density, it is ill-equipped to perform any long-term dynamics, and if it were, extracting the information seen in Chapter 2 would be a non-trivial task. Further, AMBER has repeatedly proven that it is capable of significant simulations which can be looked at with some confidence - the fact that it is not perfect and that some atoms have a more notable discrepancy does not reduce its worth. In all, though ONETEP's superiority as a method for calculating systems' properties has been made clear, both codes have produced graphs with similar features, demonstrating that while intricate calculations are not a suitable task for AMBER, long-term and large-scale simulations may be undertaken with some confidence.

4.3 Future Work

Creating more lifelike long-term methods of biomolecular simulation is a mammoth task, and ONETEP is uniquely capable - through favourable scaling *ab initio* techniques - not only to describe the features of biological systems but also illuminate the areas in which improvements should be targeted. As well as investigating binding sites and specific interactions, ONETEP would be well employed identifying troublesome sites as seen in Chapter 3, in order that the cheaper AMBER force fields may be further refined.

Polyamines, though abandoned as an effective means of combating illness, remain a subject of great interest. The sheer diversity of functions means that there are a great many investigations which would be both novel and worthwhile. Further systematic studies of DNA-polyamine interactions would certainly be a good starting point, while a more in-depth examination of the spermine-DNA backbone interaction would elucidate the mechanisms behind the information presented here. One such means of doing this would be to have spermine interact with a relatively short section of DNA - say five or ten bases - and to separate the bases into overlapping sets of three. As the simulation progresses, each set of three can be thoroughly analysed and compared to the others. The surrounding bases' response to the interaction of the spermine can then be correlated, and so an explanation of the average structure's insensitivity to the spermine may be reached. As computation gets ever cheaper, a dual *ab initio* and classical investigation into - for example - the free radical harvesting by spermine as seen in [15] could also prove fruitful.

Finally, a comparison of methods of simulation can only give information within certain bounds, and is limited in what it can achieve. Using the techniques of both packages (with knowledge of their performance) and combining this with experimentation in order to study specific processes would give a fascinating multi-scale insight into biological systems, and provide significant and deep understanding of the mechanics by which life as we encounter it is possible.

References

[1] N. Foundation. `http://www.nobelprize.org/nobel_prizes/medicine/laureates/1910/index.htm`. Accessed 15th September 2014. 1.1

[2] J. D. Watson and F. H. C. Crick, "A structure for deoxyribose nucleic acid," *Nature*, vol. 171, pp. 737–8, 1953. 1.1

[3] P. Varnai and K. Zakrzewska, "Dna and its counterions: a molecular dynamics study," *Nuclear Acids Research*, vol. 32, pp. 4269–80, 2004. 1.1

[4] A. Real and R. J. Greenall, "Influence of spermine on dna conformation in a molecular dynamics trajectory of d(cgcgaattcgcg)(2): major groove binding by one spermine molecule delays the a–>b transition.," *Journal of biomolecular structure and dynamics*, vol. 21, pp. 469–88, 2004. 1.1, 1.3, 1.5, 3.5.1

[5] D. Svozil, J. Kaline, M. Omelka, and B. Schneider, "Dna conformations and their sequence preferences," *Nucleic Acids Research*, vol. 36, pp. 3690–3706, 2008. 1.2

[6] C. Altona and M. Sundaralingam, "Conformational analysis of the sugar ring in nucleosides and nucleotides. new description using the concept of pseudorotation," *Journal of the American Chemical Society*, vol. 94, pp. 8205–12, 1972. 1.2.6.1

[7] H. W. Dudley, O. Rosenheim, and W. W. Starling, "The chemical constitution of spermine," *Journal of Biochemistry*, vol. 20, pp. 1082–94, 1926. 1.3

[8] C. W. Tabor and H. Tabor, "Polyamines," *Annual Review of Biochemistry*, vol. 53, pp. 749–90, 1984. 1.3

[9] K. Williams, "Interactions of polyamines with ion channels," *Journal of Biochemistry*, vol. 325, pp. 289–97, 1997. 1.3

[10] P. Bandyopadhyay, V. Janout, L. Zhang, J. A. Sawko, and S. L. Regen, "An ion conductor derived from spermine and cholic acid," *Journal of the American Chemical Society*, vol. 22, pp. 12888–9, 2000. 1.3

[11] R. B. Lottfield, E. A. Eigner, and A. Pastuszyn, "The role of spermine in preventing misacylation by phenylalanyl-trna synthetase," *The Journal of Biological Chemistry*, vol. 256, pp. 6729–35, 1981. 1.3

[12] C. Stefanelli, C. Pignatti, B. Tantini, M. Fattori, I. Stanic, C. A. MacKintosh, F. Flamigni, C. Guarnieri, C. M. Caldarera, and A. E. Pegg, "Effect of polyamine depletion on caspase activation: a study with spermine synthase-deficient cells," *Journal of Biochemistry*, vol. 355, pp. 199–206, 2001. 1.3

[13] C. Stefanelli, F. Bonavita, I. Stanic, M. Mignani, A. Facchini, C. Pignatti, F. Flamigni, and C. M. Caldarera, "Spermine causes caspase activation in leukaemia cells," *Federation of European Biochemical Societies Letters*, vol. 437, pp. 233–6, 1998. 1.3

[14] R. D. Allen and T. K. Roberts, "Role of spermine in the cytotoxic effects of seminal plasma," *American Journal of Reproductive Immunology and Microbiology*, vol. 13, pp. 4–8, 2013. 1.3

[15] H. C. Ha, N. A. Airisoma, P. Kuppusamy, J. L. Zweier, P. M. Woster, and R. A. Casero Jr, "The natural polyamine spermine functions directly as a free radical scavenger," *Proceedings of the National Academy of Science USA*, vol. 95, pp. 11140–5, 1998. 1.3, 4.3

[16] A. U. Khan, Y. Mei, and T. Wilson, "A proposed function for spermine and spermidine: protection of replicating dna against damage by singlet oxygen," *Proceedings of the National Academy of Science USA*, vol. 89, pp. 11426–7, 1992. 1.3

[17] K. C. Das and H. P. Misra, "Hydroxyl radical scavenging and singlet oxygen quenching properties of polyamines," *Molecular and Cellular Biochemistry*, vol. 262, pp. 127–33, 2004. 1.3

[18] E. W. Gerner and F. L. Meyskens Jr, "Polyamines and cancer: old molecules, new understanding," *Nature Reviews Cancer*, vol. 4, pp. 781–92, 2004. 1.3

[19] K. E. Cole and A. M. Valentine, "Spermidine and spermine catalyze the formation of nanostructured titanium oxide," *Biomacromolecules*, vol. 8, p. 1641, 2007. 1.3

[20] A. Peres, C. L. Cahu, and J. L. Z. Infante, "Dietary spermine supplementation induces intestinal maturation in sea bass (dicentrarchus labrax) larvae," *Fish Physiology and Biochemistry*, vol. 16, pp. 479–85, 1997. 1.3

[21] O. Peulen, P. Deloyer, C. Grandfils, S. Loret, and G. Dandrifosse, "Intestinal maturation induced by spermine in young animals," *Physiological, Genetic and Nutritional Aspects of Tissue Growth in Farm Animals*, vol. 66, pp. 109–20, 2000. 1.3

[22] B. G. Feuerstein, N. Pattabiraman, and L. J. Marton, "Molecular mechanics of the interactions of spermine with dna: Dna bending as a result of ligand binding," *Nucleic Acids Research*, vol. 18, p. 1271, 1990. 1.3

[23] B. G. Feuerstein, N. Pattabiraman, and L. J. Marton, "Molecular dynamics of spermine-dna interactions: sequence specificity and dna bending for a simple ligand," *Nucleic Acids Research*, vol. 17, pp. 6883–92, 1989. 1.3, 1.5

[24] A. Jain, G. Zon, and M. Sundaraligam, "Base only binding of spermine in the deep groove of the a-dna octamer d(gtgtacac)," *Biochemistry*, vol. 28, pp. 2360–4, 1989. 1.3

[25] B. G. Feuerstein, N. Pattabiraman, and L. J. Marton, "Spermine-dna interactions: A theoretical study," *Proceedings of the National Acadamy of Science USA*, vol. 83, pp. 5948–52, 1986. 1.3

[26] B. J. Alder and T. E. Wainwright, "Phase transitions for a hard sphere system," *Journal of Chemical Physics*, vol. 27, pp. 1208–9, 1957. 1.4

[27] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, pp. 585–90, 1977. 1.4

[28] C. Maffeo, B. Luan, and A. Aksimentiev, "End-to-end attraction of duplex dna," *Nucleic Acids Research*, vol. 40, pp. 3812–21, 2012. 1.4

[29] M. W. Konrad and J. I. Bolonick, "Molecular dynamics simulation of dna stretching is consistent with the tension observed for extension and strand separation and predicts a novel ladder structure," *Journal of the American Chemical Society*, vol. 118, pp. 10989–94, 1996. 1.4

[30] M. Levitt, "Protein folding by restrained energy minimization and molecular dynamics," *Journal of Molecular Biology*, vol. 170, pp. 723–64, 1983. 1.4

[31] D. Porschke, "Dynamics of dna condensation," *Biochemistry*, vol. 23, pp. 4821–8, 1984. 1.5

[32] I. S. Haworth, A. Rodger, and W. G. Richards, "A molecular dynamics simulation of a polyamine-induced conformational change of dna. a possible

mechanism for the b to z transition.," *Journal of Biomolecular Structure and Dynamics*, vol. 10, pp. 195–211, 1992. 1.5

[33] N. Korolev, A. P. Lyubartsev, A. Laaksonen, and L. Nordensklod, "On the competition betwen water, sodium ions, and spermine in binding to dna: A molecular dynamics computer simulation study," *Biophysical Journal*, vol. 82, pp. 2860–75, 2002. 1.5

[34] K. Brysont and R. J. Greenall, "Binding sites of the polyamines putrescine, cadaverine, spermidine and spermine on a- and b-dna located by simulated annealing," *Journal of Biomolecular Structure and Dynamics*, vol. 18, pp. 393–412, 2000. 1.5

[35] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, pp. 1668–88, 2005. 2.1

[36] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman, "Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules," *Computer Physics Communications*, vol. 91, pp. 159–89, 1995. 2.1

[37] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta Jr, and P. Weiner, "A new force field for molecular mechanical simulation of nucleic acids and proteins," *Journal of the American Chemical Society*, vol. 106, pp. 765–84, 1984. 2.1, 2.1.1

[38] U. C. Singh and P. A. Kollman, "A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the ch3cl + cl- exchange reaction and gas phase protonation of polyethers," *Journal of Computational Chemistry*, vol. 7, pp. 718–30, 1986. 2.1

[39] J. W. Ponder and D. A. Case, "Force fields for protein simulations," *Advances in Protein Chemistry*, vol. 66, pp. 27–85, 2003. 2.1.1

[40] A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham III, C. A. Laughton, and M. Orozco, "Refinement of the amber force field for nucleic acids: Improving the description of alpha/gamma conformers," *Biophysical Journal*, vol. 92, pp. 3817–29, 2007. 2.1.1

[41] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham III, and P. Jurecka, "Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles," *Journal of Chemical Theory and Computation*, vol. 7, pp. 2886–902, 2011. 2.2.1

[42] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *Journal of the American Chemical Society*, vol. 117, pp. 5179–97, 1995. 2.2.2

[43] A. Onufriev, D. A. Case, and D. Bashford, "Effective born radii in the generalized born approximation: the importance of being perfect," *Journal of Computational Chemistry*, vol. 23, 2002. 2.2.3.2

[44] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *Journal of the American Chemical Society*, vol. 112, 1990. 2.2.3.2

[45] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *Journal of Chemical Physics*, vol. 79, 1983. 2.2.3.3

[46] M. W. Mahoney and W. L. Jorgensen, "A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions," *Journal of Chemical Physics*, vol. 112, 2000. 2.2.3.3

[47] R. J. Loncharich, B. R. Brooks, and R. W. Pastor, "Langevin dynamics of peptides: The frictional dependence of isomerization rates of n-acetylalnyl-n'-methylamide," *Biopolymers*, vol. 32, pp. 523–535, 1992. 2.2.4, 2.2.4

[48] H. J. C. Berendsen, J. P. M. Postma, W. F. V. Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with couple to an external bath," *The Journal of Chemical Physics*, vol. 81, 1984. 2.2.5, 2.2.5, 2.2.5

[49] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, "Curves+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures," *Nucleic Acids Research*, vol. 37, pp. 5917–29, 2009. 2.3.6, 2.4.4

[50] R. Lavery. `http://bisi.ibcp.fr/tools/curves_plus/` `canal-user-guide.html`. Accessed 4th January 2015. 2.3.6

[51] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, "Introducing onetep: Linear-scaling density functional simulations on parallel computers," *Journal of Chemical Physics*, vol. 122, 2005. 3.1, 3.2.1