UNIVERSITY OF SOUTHAMPTON

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

SCHOOL OF CHEMISTRY

MOLECULAR DYNAMICS STUDIES OF DNA TRANSLOCATION THROUGH NANOPORES

Andrew Thomas Guy MBiochem

For the title of Doctor of Philosophy

October 2012

ABSTRACT

DNA sequencing as a technology has revolutionized the world of genetics, allowing much knowledge of genome organization and by extension genome manipulation. However, all sequencing undertaken today uses variants of the original method described by Sanger et al in 1977. While the efficiency of the Sanger method continues to improve, it is still relatively slow and expensive, requiring multiple reagents and amplification steps to achieve results. The relatively more recent nanopore sequencing method, based on the α-haemolysin pore forming toxin and first described in 1996, has the potential to offer fast and inexpensive single-molecule DNA sequencing. There are several barriers to the deployment of the technology. Two areas are of particular interest: Firstly, slowing the speed of strand translocation, as normal translocation speeds are too fast for current-generation sensors; and secondly, improving the resolution of base detection, as some bases show similar current blockage and can therefore be confused in the detection process.

Molecular dynamics simulations can provide molecular-level descriptions of these processes, and in conjunction with mutational studies and free energy calculations, can provide explanations and solutions to these phenomena. Presented in this work are studies of the α-haemolysin protein using such simulations. Using a simplified model pore system, translocation of both short and longer single-stranded DNAs are considered, and the effects of mutation on the conformations adopted by these DNAs are examined. Also considered are the effects of the available molecular mechanics parameter sets on the simulation of such molecules. Finally, free energy calculations using the umbrella sampling method are employed on small molecules representative of parts of a DNA base in a bid to uncover the effects of positional free energies on the translocation process.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

DECLARATION OF AUTHORSHIP

I, ……,

declare that the thesis entitled

……

……

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:

    1. Molecular dynamics simulations of DNA within a nanopore: arginine-phosphate tethering and a binding/sliding mechanism for translocation. 2011. *Biochemistry* 50(18) 3777-3783
    2. Single-stranded DNA within nanopores: conformational dynamics and implications for sequencing; a molecular dynamics simulation study. 2012. *Biophys J* 103(5) 1028-1036

Signed: ……

Date:……

Acknowledgements

There are many people to thank for their assistance during the years of working on this thesis. Particular acknowledgements must be given to Pete Bond for playing a major role in the creation of the model pore systems, and to Tom Piggot for valuable contributions in the work on DNA force field validation. Thanks also to Andrew Heron and Hagan Bayley for providing the experimental current data used in chapter 3. On the funding side, thanks to Oxford Nanopore Technologies for sponsoring this work, and specifically to John Milton and Jayne Wallace for useful scientific discussions. On a more general note, I'd like to thank all members past and present of the Systems and Synthetic Biology Modelling Group for their help and support over the years. Particular mention must, of course, be given to my supervisor, Dr. Syma Khalid, for much useful input both throughout the course of this Ph.D. and during the writing of this thesis. And finally, a special thank you to my parents, for putting up with me during the months of writing. Much love to you both.

List of Abbreviations

α-HL: α-hemolysin
DMPC: Dimyristoyl phosphatidyl choline
dsDNA: Double-stranded deoxyribonucleic acid
ssDNA: single-stranded deoxyribonucleic acid
PDB: Protein data bank
PFT: Pore forming toxin
PME: Particle mesh Ewald
PMF: Potential of mean force
WHAM: Weighted histogram analysis method
WT: Wildtype

Chapter 1. Introduction

1.1 DNA and DNA sequencing

The discovery of the double-helical structure of DNA in the 1950s, followed by the genetics revolution of the 1970s, accelerated the development of the field that became molecular biology. The establishment of DNA as the genetic material led to the question of how information was stored in the genome, and whether researchers could access this information. This would in theory lead to a wealth of knowledge about the molecular basis of life itself, with potential applications in genomics, genetic engineering and personalised medicine. Thus began the development of DNA sequencing methods.

Early methods were cumbersome and slow, with the first gene sequence to be reported coming not from DNA but from RNA from the bacteriophage MS2, and being derived from a nuclease digest rather than directly sequenced (*1*). The full genome of MS2 was published four years later (*2*), the first whole genome of any organism, and again derived from a nuclease digest of the genome. In 1977, a method based on chemical modification of individual bases followed by cleavage at these modified sites was reported by Maxam and Gilbert (*3*); the same year Sanger and coworkers reported the development of chain-terminator sequence (*4*). Maxam-Gilbert sequencing eventually fell out of use in favour of Sanger sequencing, mostly due to the greater relative ease of the latter.

Most current-generation sequencing technology is based on the chain termination method. Chain termination sequencing uses a DNA polymerase to replicate the sequence of interest in the presence of both normal nucleotides and labelled nucleotides that lack the 3' hydroxyl group necessary for chain extension. When these dideoxy nucleotides are incorporated into the growing strand, replication is halted. The incorporation of labelled nucleotides is a random process; as such, they may be incorporated at any appropriate point, resulting in terminated strands of varying lengths. These aborted strands can then be separated based on size using gel electrophoresis. Since each nucleotide carries a different marker, the sequence can then be determined from the gel.

Despite its success in sequencing even entire genomes, chain terminator methods remain somewhat impractical when deployed on a large scale due to cost and the requirements for amplification steps before sequencing can begin (*5*). This has lead to the search for replacement methods. While several have been proposed thus far, nanopore sequencing shows particular

promise as it is a simple, fast, single-molecule method that does not require expensive amplification or labelling steps that other methods often do. This has led to much interest in developing it as a next-generation sequencing technology (*6*).

1.2 Nanopore sequencing

Nanopores are nanometer-scale pores, either formed naturally by some membrane proteins or by boring a hole into a solid-state surface. The basic principle of a nanopore detection system is relatively simple. The detector consists of two compartments of ionic liquid separated by a nanopore. Application of an electric field or concentration gradient across the nanopore results in an ionic current, which is measured to give a baseline reading. Addition of an analyte to the solution results in a drop in the observed current reading when the analyte enters the pore, as the pore becomes partially blocked. This change in current is characteristic of the individual analyte, with different analytes resulting in the current decreasing by different magnitudes. The process is sensitive enough to distinguish even between stereoisomers of the same compound (*7*), and as such can in theory readily distinguish between DNA bases (*8*).

Nanopore sequencing theoretically offers fast, low-cost, single-molecule DNA sequencing, and is one of several candidate technologies to reach the National Institutes of Health's 1,000-dollar genome target. The goal of this target is to provide a method that will offer the ability to sequence an entire human genome at a consistent price of one thousand USD (*9*). This in theory would accelerate the development of personalized medicine by allowing an individual's genome to be sequenced rapidly, as well as allowing easy access to the genomes of uncharacterized species. Several potential variants of the technology are being explored, broadly speaking in two categories: solid-state and biological nanopores (*6*).

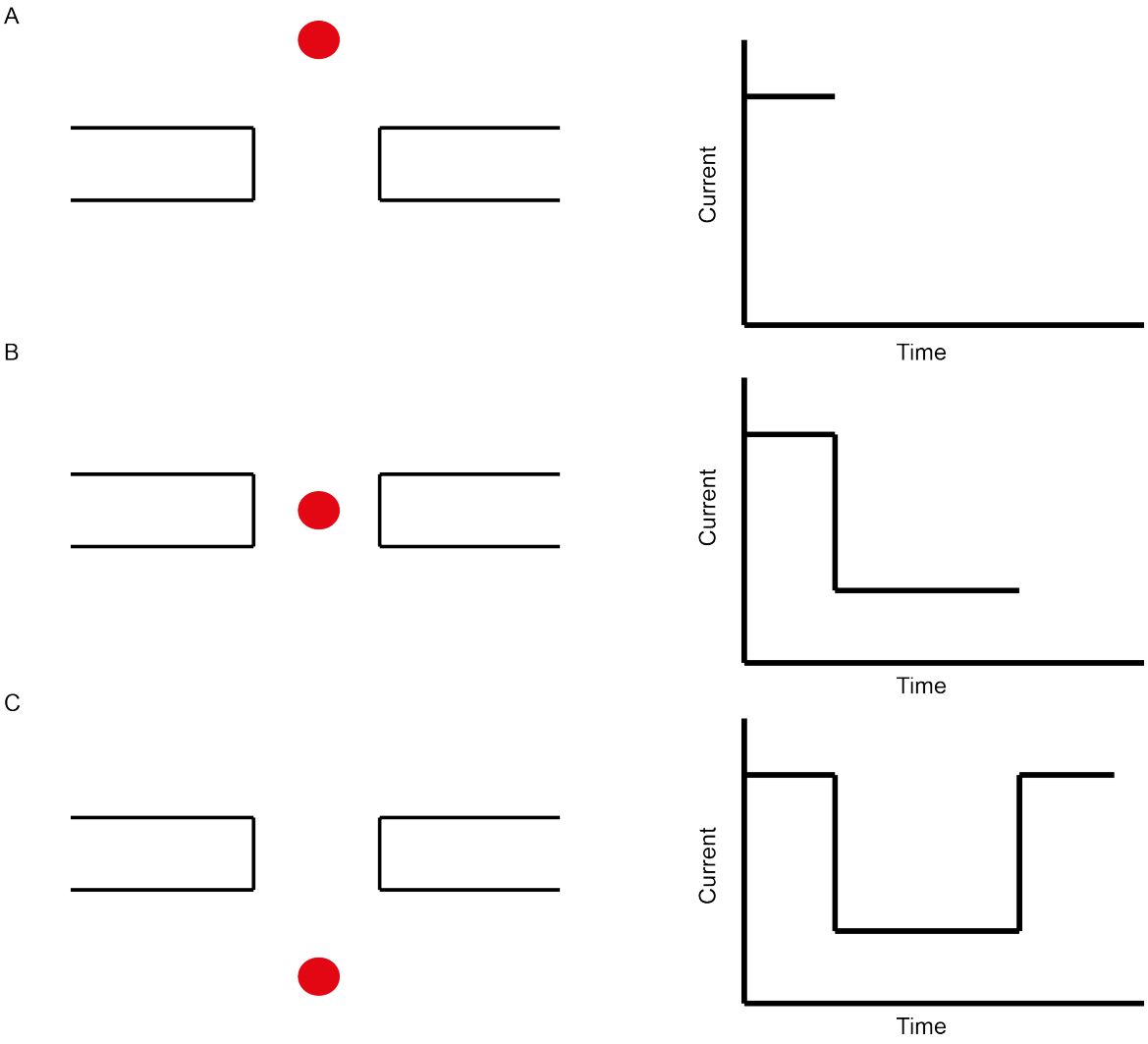

Figure 1. Schematic of nanopore sensing process. (A) the pore is in the open state with analyte present in solution (left); the open pore current is read (right). (B) an analyte enters the pore, resulting in a drop in observed current. This current block lasts as long as the analyte is present in the pore. (C) the analyte leaves the pore, resulting in a return to open pore current. Different analytes create different levels of current block.

1.2.1 Solid-state nanopores

Solid-state nanopores are simple nano-scale holes created in a solid surface, frequently silicon nitride, silicon dioxide, or, more recently, graphene (*10*). Pores in silicon-based surfaces are usually created by ion-beam sculpting (*11*): firstly, a relatively large hole is bored in the solid surface using a high-intensity beam, then lateral mass transport is used to close the hole to the desired width. This process involves stimulating the migration of surface atoms so that the large hole created originally begins to fill. Graphene-based nanopores are formed in a single step by drilling a hole using a particle beam as they lack available material to reduce the size of the pore, being only one atom in width. The graphene sheet is usually placed over a larger diameter hole on a silicon nitride support structure (*12*).

Solid-state nanopores have the advantages of being robust and very stable, and allow fine control of pore size, but lack the easy modifiability and potential for functionalisation of protein nanopores due to lack of potential reaction sites in the homogenous surface. As such, while they have great potential for use in future detection systems, protein nanopores are more commonly used in the present generation of nanopore sequencing devices (*13*).

1.2.2 Protein nanopores

Protein nanopores are found throughout nature, although not all are suitable for use in nanopore devices. While not generally as robust as solid-state nanopores, some proteins such as α-hemolysin (α-HL) are stable for long enough to make them useful in nanopore detectors. Protein nanopores offer easy modifiability of the chemical properties of the residues lining the pore, either by mutagenesis or direct chemical modification, as well as the ability to add adapter molecules to fine-tune the current and translocation properties. Several proteins have been investigated for use in nanopore devices, in particular OmpG, Phi29, MspA and, most commonly, α-HL. Further work into the use of small peptides has revealed the possibility of using rather minimal protein-based constructs instead of whole proteins as nanopore detectors (*14*).

1.2.2.1 α-Hemolysin

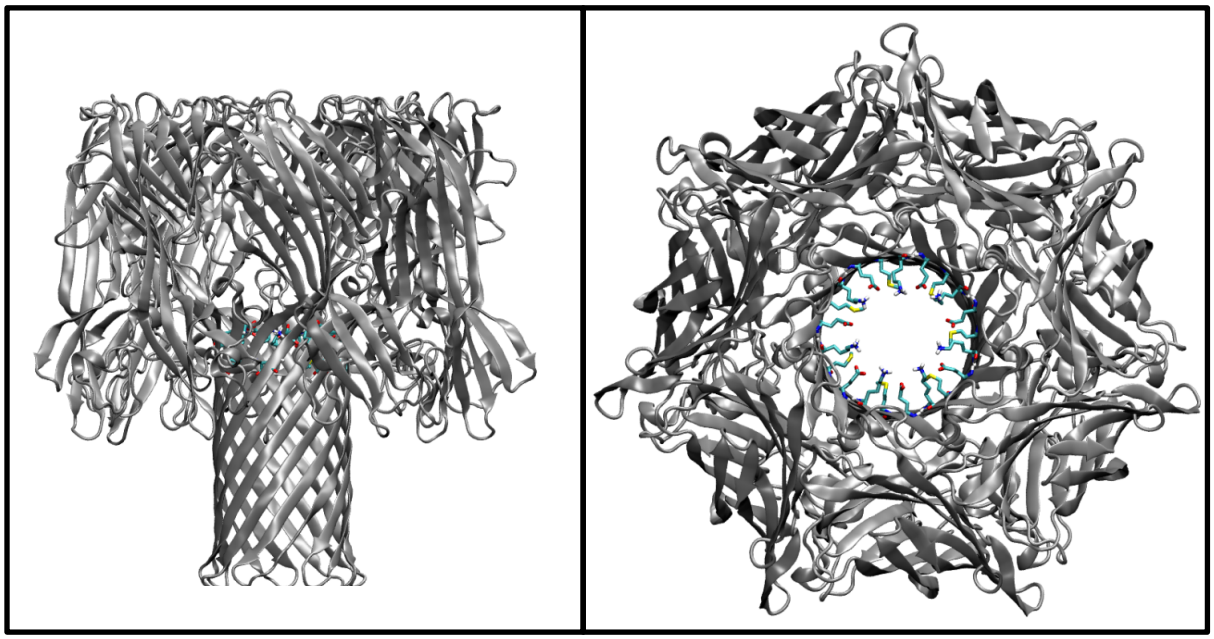

Figure 2. Structure of α-HL from side (left) and top (right). Highlighted are residues K147, E111 and M113 which form the major constriction in the middle of the pore. Structure taken from PDB ID 7AHL.

α-HL is perhaps the most widely studied of all protein nanopores, forming the basis of the first nucleotide detection experiments reported by Kasianowicz et al in 1996 (*15*). The protein itself is a staphylococcal pore-forming toxin (PFT), consisting of seven identical subunits that form a large heptameric complex of some 240 kDa (*16*). The complex itself is very stable and therefore an attractive target for modification to optimize the pore properties for sequencing devices. One particular problem faced in experiment is the translocation speed of nucleotides or nucleic acid strands: such molecules move too quickly through the pore for the present generation of current recording devices to measure. Attention therefore has been focused on slowing down translocation, often by adding extra positive charges to the pore by mutation to arginine or lysine. A common mutation is M113R, frequently used as the standard pore in sequencing experiments (*8*). Other reported modifications include M113F (*7*), which allows the addition of adapter molecules at different points of the pore to usual, and E111N/K147N (*17*), which substitutes a both a positively and negatively charged residue for a polar residue at the main constriction site, thus allowing an increased baseline pore current. A commonly reported mutation is RL2, which causes conservative mutations in four residues, and removes a positive charge from the upper pore entrance (*18-20*). While this mutation has some effect on translocation due to the substitution of lysine 8, it was originally designed as a means of introducing restriction sites into the gene, thus allowing easy manipulation of the protein by cassette mutagenesis.

Other modifications that can be applied to the pore include addition of a cyclodextrin adapter, which allows finer control over the pore diameter and therefore pore current, pore diameter being a property that is harder to modify in a protein pore compared to solid-state pores (*13*). Also proposed was the use of processive exonuclease enzymes attached to the pore to progressively degrade a DNA strand before it reaches the nanopore, thus allowing control the rate of strand translocation, with this modification allowing controlled release of individual bases at predictable, widely spaced intervals. Exonuclease sequencing using fluorescent markers as a detector was first proposed in 1989 (*21*); proposals exist to adapt it for use with a nanopore detector (*8*).

1.2.2.2 OmpG

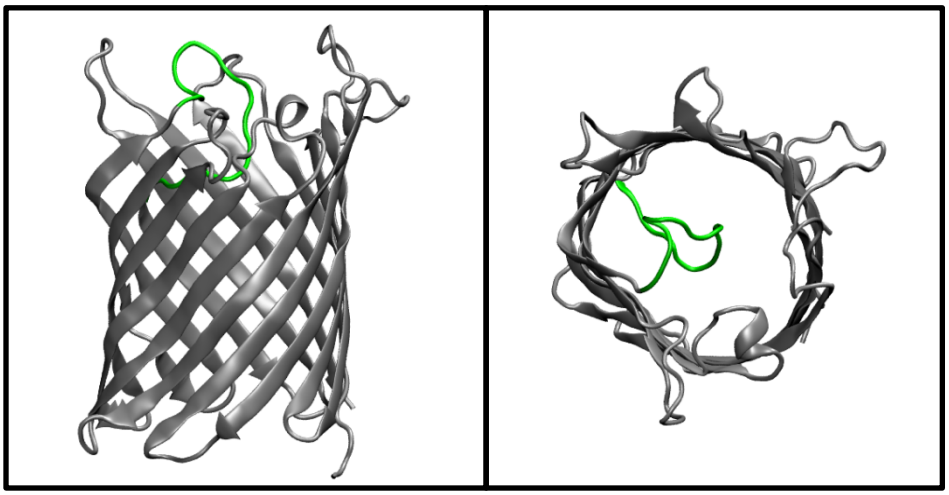

Figure 3. Structure of OmpG in closed conformation from side (left) and top (right). Gating loop highlighted in green. Taken from PDB ID 2IWW

OmpG is an outer membrane porin from *E. coli*. Like the α-HL transmembrane domain, it is a 14-stranded transmembrane β-barrel (*22*). Unlike α-HL, however, it is a monomeric protein and as such the interior of the pore lacks the characteristic sevenfold symmetry of α-HL. It has been shown to be usable in detection experiments, but suffers due to its intrinsic gating mechanisms, being both voltage- and pH-gated. The voltage gating is activated at ±100 mV, closing the pore; the pH gating closes the pore below pH 7. However, there is also spontaneous gating observed even when conditions favour the open conformation. Molecular modelling and simulations combined with protein engineering and electrophysiology experiments have thus focused on making it a 'quiet' pore (*23*) in an attempt to remove the gating, by enhancing the stability of the loop regions. This was achieved by deletion of D215, allowing a more optimised hydrogen bonding network between loops β-11 and β-12, and addition of a disulphide bond between loops β-12 and β-13 by introduction of a pair of cysteine residues. These optimisations reduced spontaneous gating by 90%.

1.2.2.3 Phi29 motor protein

The bacteriophage phi29 is a double-stranded DNA virus, and as part of the process of packaging and unpackaging its nuclear material in the viral capsid it utilizes the phi29 motor protein (*24*). It has been shown to be possible to reconstitute this protein into a vesicle and utilize it as a detector (*25*); this is of particular interest due to the pore diameter being sufficient to allow passage of double stranded DNA instead of being limited to single strands as with most other protein

nanopores. The pore diameter is between 3.6 and 6 nm wide, much wider than the ~2.2 nm of α-HL's transmembrane beta barrel.

1.2.2.4 MspA

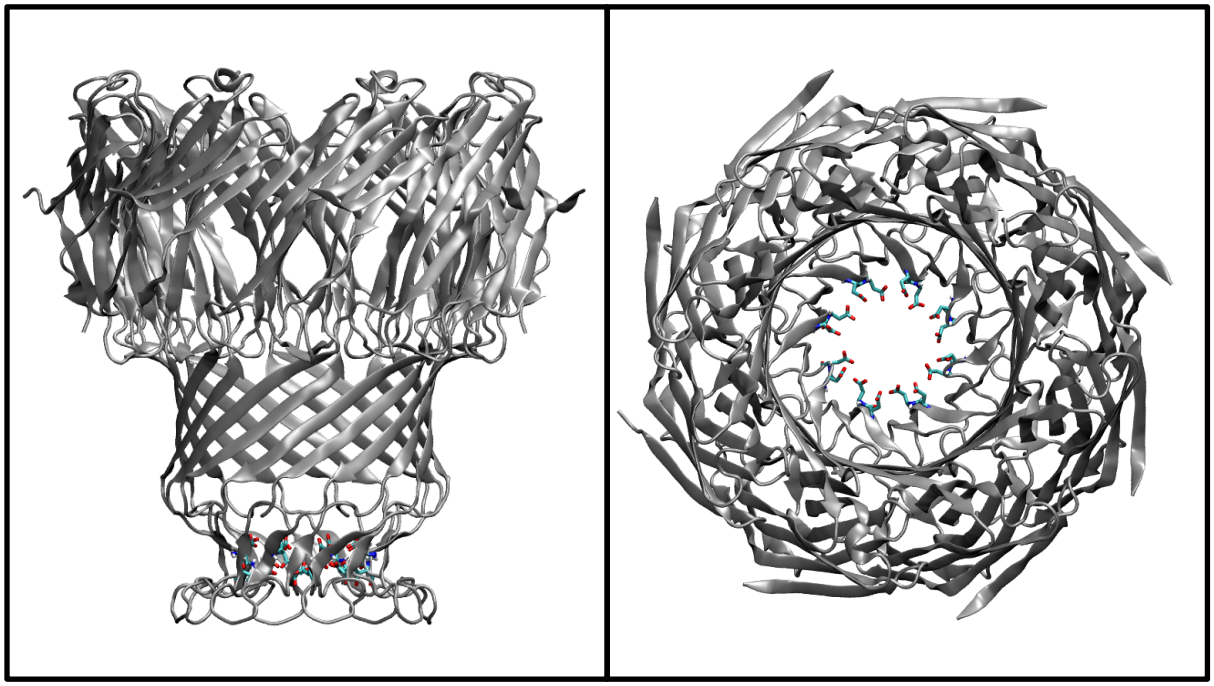

Figure 4. Structure of MspA from side (left) and top (right). Residues D90 and D91, which form the constriction site, are highlighted. Taken from PDB ID 1UUN.

MspA is a porin from M*ycobacterium smegmatis* (*26*). Like α-HL, it has proven to be stable and robust, and as such is another attractive candidate as a protein for nanopore detection (*27*). Unlike α-HL, however, the protein only contains a single, narrow constriction of ~1 nm, whereas α-HL contains two of diameter ~1.4 nm and 2 nm (*28*). This narrow constriction is theorized to allow greater resolution between the individual bases of a strand, as these constriction sites likely form the basis of detection of analytes, being narrower than the rest of the pore and thus being the limiting factor in the ionic current reading.

1.2.2.5 FhuA

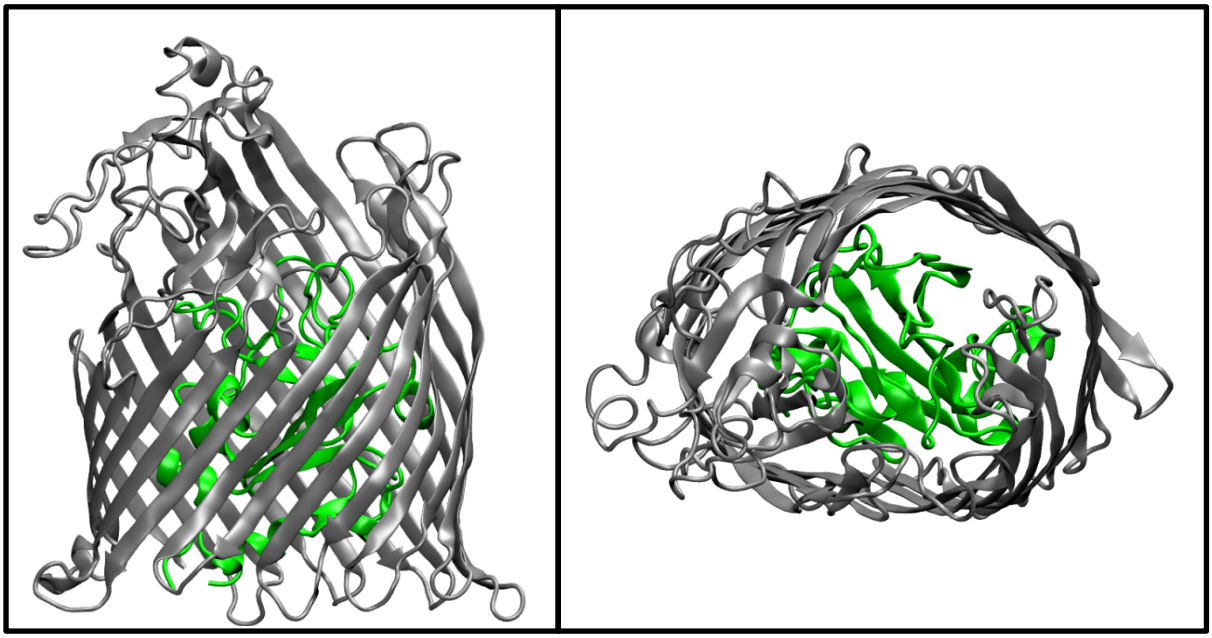

Figure 5. Structure of *e. coli* FhuA from side (left) and top (right). Plug domain highlighted in green. Taken from PDB ID 1FI1.

FhuA is a protein responsible for uptake of iron into bacterial cells (*29*). As iron in the free ferric state has low bioavailability, bacteria such as *e. coli* secrete the siderophore ferrichrome, a cyclic peptide that complexes with free iron. FhuA itself is a TonB dependent transporter that, when in complex with TonB, actively transports the ferrichrome-iron complex across the outer membrane. The protein is a 22-stranded β-barrel, which in theory would have an open pore diameter large enough to transport double stranded DNA; however, the protein contains a plug domain that blocks the majority of the pore. As such, use of FhuA as a nanopore has been proposed, but extensive protein editing has been required to allow DNA passage. As such, a mutated pore has been proposed that removes a third of the protein in order to delete the plug domain (*30*) and create a pore suitable for nanopore experiments.

1.3 Simulations and nanopore sequencing

While nanopores remain a promising field of research for single-molecule detection, there are potential issues in the construction and operation of such devices that must be solved to allow wide-scale deployment of detector devices based on the technology. In particular, issues remain such as nucleotide translocation speeds being too rapid, or the question of whether the configuration of the strand as it exits the nanopore can have an effect on the current readings observed. While such effects can be observed on a macroscopic scale, what is lacking is the fine

detail, particularly of interactions between the nucleic acids and the nanopore. Molecular dynamics simulations can allow observation of these interactions at the atomic level, providing insights that conventional experimental techniques cannot.

Accurate simulation requires the development of accurate simulation parameters, referred to as 'force fields'. Such force fields are generally well-tested in the case of proteins, but nucleic acids have proven harder to simulate due to a variety of factors. Firstly, DNA is a highly charged anionic polymer. Each additional nucleotide increases the formal charge of a strand by -1, through the phosphate moiety. The high charge density was known to cause problems in early simulations (*31*) due to the limitations of the methods used to calculate the electrostatic interaction energies. Such methods would only take into account short-range electrostatics and ignore long-range interactions for the sake of computational efficiency; however, these methods would lead to deformation of the double helix to the extent that separation of the two individual strands was often observed (*32*). To overcome this dissociation of the double helix the DNA electrostatics was significantly modified, for example by reducing the charge of the phosphate groups (*33*) or even removing the backbone charge entirely, to give charge neutral DNA molecules (*34*). More recently, the need to artificially modify the DNA electrostatics has for the most part been overcome by the development of better force-field parameters, coupled with new methods for evaluation of long-range electrostatics, such as particle mesh Ewald (PME) (*35*).

Secondly, double-stranded DNA relies largely upon inter-strand hydrogen bonding between the bases of individual strands, to prevent the two strands from separating. However, force fields commonly used for biological molecules do not contain an explicit hydrogen bonding term. Instead hydrogen bonds are defined purely in terms of the electrostatic interaction energy between two atoms, thus in practice the hydrogen bonding contribution is often underestimated. Clearly in a molecule such as DNA in which hydrogen bonding plays such a key role, it is inevitable that this will lead to inaccurate conformational dynamics. Indeed this was shown to be the case in simulations in which (11) strand separation was observed during simulations with timescales of the order of nanoseconds to tens of nanoseconds. There is also evidence that terminal nucleotide definitions are important (*36*). This is shown by the effect of using the Gromacs G53A6 force field with undefined terminal nucleotides; DNA without the correct terminal nucleotides will progressively lose its double-helical structure as the base pairs separate. However even when this is corrected, G53A6 DNA tends to favour a conformation somewhere between the canonical A and B structures.

Longer simulation timescales, made possible by steady advances in both hardware and software, have inevitably tested force fields to their limits, and uncovered previously unrecognised shortcomings. The realisation that to address these issues requires enormous computational resources, large datasets of problem types, and broad community support, led to the establishment in 2001 of the Ascona B-DNA Consortium (ABC), an informal network of DNA simulators around the world who pool their resources to address such issues. One key aim of the consortium has been to develop accurate and predictive models of the relationship between DNA sequence and DNA flexibility (*37-39*). For example, a breakdown in DNA helical structure for simulations performed with the AMBER parm99 force field only became apparent when simulations reached the tens to hundreds of nanoseconds in the early to mid-2000s, (*40*). This led to the reparameterisation of the force field to the parmBSC0 variant, to correct the description of the α/γ backbone torsion angles.

As a result of these parameterizations, it is now possible to draw a consensus view of aqueous B-DNA flexibility from the various available force fields. Pérez *et al* (*41*) use parmBSC0 (*40*), CHARMM27 (*42*) and two older AMBER variants - parm94 (*43*) and parm99 (*44*) - to examine the effect of force field on DNA flexibility in solution. This study systematically examines every base-pair type in a double strand compared to crystal structure through long simulations. It is found that although parmBSC0 and CHARMM27 differ in some respects, their representations of the flexibility of a double helix are very similar, and both are a suitable model for such a system over a long simulation timescale. The two older force fields, while having known issues, were also shown to provide a reasonable description despite their limitations.

In summary, the parameters used in nucleic acid simulation must be carefully chosen in order to accurately describe the systems of interest. Despite the limitations, it can be shown that when applied correctly, molecular simulation is a powerful tool for studying nucleic acids and their complexes. A number of the recent successes are discussed below.

1.4 Double-stranded DNA

1.4.1 Double-stranded DNA in solution

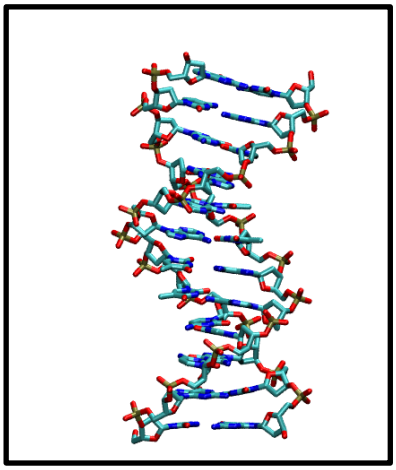

Figure 6. DNA in the double-stranded B-conformation. Taken from PDB ID 1BNA.

MD simulations of DNA are dominated by those of double-stranded B-DNA (dsDNA), unsurprisingly, as this is the most ubiquitous form of DNA in biology. The first MD simulation of this type in aqueous medium was reported by Seibel *et al* (*45*) using the force field developed by Weiner *et al* (*46*). The simulated system consisted of 5 base-pair double helix fragments with explicit water molecules and sodium ions. The simulation ran for 106 ps and little drift from the double-helical structure was observed. With the increased computational power now available, modern simulations of larger systems can now reach the microsecond timescale. This extended simulation time lead to the discovery of drift from the canonical structure, which was not observable on shorter timescales. Later force field revisions therefore often contain a suitable correction factor (*40, 47, 48*).

Modern computational resources have recently permitted the study of far larger DNA structures than the short oligomers that dominated the field for many years. For example, Harris and co-workers (*49, 50*) have analysed the response of DNA circles to positive and negative superhelical stress, and the conditions under which they will adopt supercoiled structures. The interaction between circle size, superhelical density and ionic strength is complex, and represented through phase diagrams. These interactions result in novel structures that require further testing in order to determine whether they are genuinely possible structures or artifacts of parameterisation.

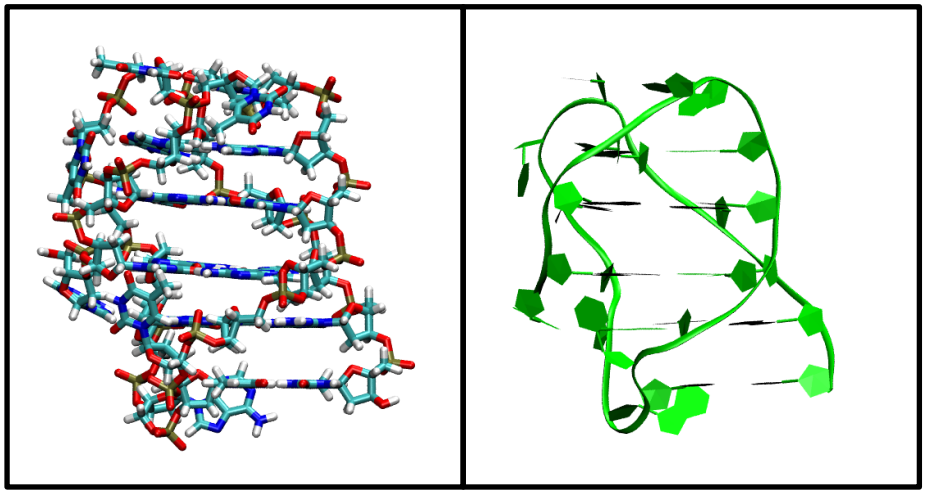

Figure 7. G-quadruplex structure in licorice (left) and ribbon trace (right) representation. Taken from PDB ID 139D.

In biology, DNA is not restricted to the double-stranded B-form discussed thus far. More complex structures such as G-quadruplexes are now well-characterised (*51*). G-quadruplexes are formed from four guanosines that are stabilized by Hoogsteen hydrogen bonds. G-quadruplexes can be formed by telomeric DNA sequences, and their formation in vivo is believed to a mechanism through which the activity of the telomere-extending enzyme telomerase is inhibited. Telomerase is up-regulated in many cancers, and as such is a potential drug target (*52*). Haider *et al* (*53*) have studied the structural properties of quadruplexes using molecular dynamics simulations and principal component analysis. They have demonstrated that over the timescale of the simulation, the structural integrity of the quadruplexes is retained. In particular, it was noted that the stability of the quadruplex was increased when more G tetrads were added. The authors described a methodology for modelling quadruplex multimers and also their interaction with ligands, and this methodology was used in later work (*54, 55*).

More recent work in this area has included studies of cation-quadruplex interactions. The ion-binding of G-quadruplexes using classical MD and hybrid quantum mechanics/molecular mechanics approaches was reported by Reshetnkiov *et al* and free-energy calculations of cation movement through G-quadruplex DNA channels was reported by Akhshi *et al* (*56*). More recent work has considered Holliday junctions, a structure important in DNA repair (*57*). It was shown that current force fields, despite not being parameterised against such structures, can provide a reasonable description over a relatively long timescale (100 ns).

1.4.2 DNA-protein interactions

In biology, manipulation of DNA is invariably achieved through its interaction with proteins that are highly specialised to carry out their function from histone proteins that act as a scaffold for wrapping and compacting DNA (*58*) to the transcription factors and polymerases that are the basis of gene expression. Thus, these interactions are a fundamental factor in the molecular basis of genetics. Many simulation studies of DNA-protein interactions have been reported in the literature. An early example is the simulations of the *lac* operator and associated protein reported by de Vlieg *et al* (*59*). MD simulation of DNA-protein complexes are reviewed in (*60*).

A key feature of sequence-specific DNA-protein interactions that molecular modelling studies have a unique ability to investigate is the degree to which the recognition process is direct – i.e. driven by the formation of specific protein-DNA contacts that are only possible when the DNA sequence is correct, or indirect – i.e. driven by the protein enforcing a distortion of the DNA double helix away from its canonical structure, which may be easier or harder depending on the local sequence. Zakrzewska *et al* have described a method that breaks down the interaction site of a DNA-protein complex into a series of fragments, and studies all possible DNA base sequences for each fragment. Using high performance computing, this allows examination of a wide range of possible protein and nucleic acid combinations (*61*).

A number of independent studies have suggested that water and/or ions are important in mediating the interaction between DNA and proteins, in particular at the interface between the two molecules (*62, 63*). Fuxreiter *et al* (*64*) have used grand canonical Monte Carlo simulations to identify the positions of water molecules at this interface. The results were in good agreement with crystallographic data, suggesting that the positions of the water molecules in the x-ray structure are non-random and therefore important to stabilizing the binding site. Other studies have focused on ions such as Zn^{2+} in the p53 tumour suppressor (*65*), in which it was revealed that the ion at the binding site, coordinates loop rearrangement for DNA binding. More recently, an elegant study of the recognition of methylated DNA by methyl-CpG binding domain (MBD) proteins using molecular dynamics and quantum mechanics calculations has been reported (*66*). The calculations revealed that methylated dinucleotides are recognized at the protein-DNA interface by two arginine residues, through an interplay of hydrogen bonding and cation-π interaction.

Molecular dynamics can be complemented by other techniques in the study of these systems, and in particular free energy calculations that allow some quantification of the strength of an

interaction (*67, 68*) Other techniques include steered molecular dynamics, in which an external force is applied to part of the system in order to pass a free energy barrier or accelerate a process that would otherwise not occur on a timescale accessible to simulation. Recently, steered MD was used to simulate the unwrapping of DNA from a histone (*69*), which allowed the identification of energy barriers to DNA unwinding.

1.4.3 DNA-ligand interactions

The interaction of DNA with natural and synthetic ligands is the key molecular recognition process for many therapeutic agents, such as cisplatin. Thus, identifying potential binding sites and modes of action of DNA-binding drugs is important for the future development of novel drugs that target DNA. MD simulations have played a key role in identifying both DNA-binding sites and also ligand-binding modes (*70, 71*).

DNA-binding modes are often divided into three categories for convenience; groove-binding, intercalation between base pairs and electrostatic association with the backbone. The latter two binding-modes have been reviewed in (*72*). DNA groove-binding offers more scope for specificity, through simultaneous interaction with a greater number of base pairs compared to intercalation, and thus is a desirable option for drug-design. Minor groove-binding drugs such as Hoechst 33258 have been shown to recognise AT-rich regions in the double helix (*73*), and this recognition and interaction has been shown to be in part dependent on induced fit to optimise the van der Waals contacts between ligand and DNA (*74*). Minor groove binding ligands with sequence specificity have been designed (*75*), potentially allowing design of drugs to target specific genes or regions of DNA, however it is clear that the accurate prediction of binding affinites by modelling methods remains a serious challenge (*76*). In common with DNA protein interactions, it has been suggested that DNA-ligand binding relies on water as a mediator (*77*). In these cases it has been shown that structures too rigid to fit dsDNA grooves use water as a bridging molecule.

The major groove offers greater scope for binding specificity, and is utilised by proteins for binding and manipulating DNA. Taking inspiration from nature's example, a synthetic, DNA major groove-binding molecule was designed by Hannon and co-workers (*78*). The interaction of the major groove with a supramolecular cylindrical ligand was shown to cause bending of the DNA to a degree that conforms well with experiment (*79*). The effect of modification of this ligand was further explored in (*80*), where the interaction is shown to have a high dependence on ligand shape.

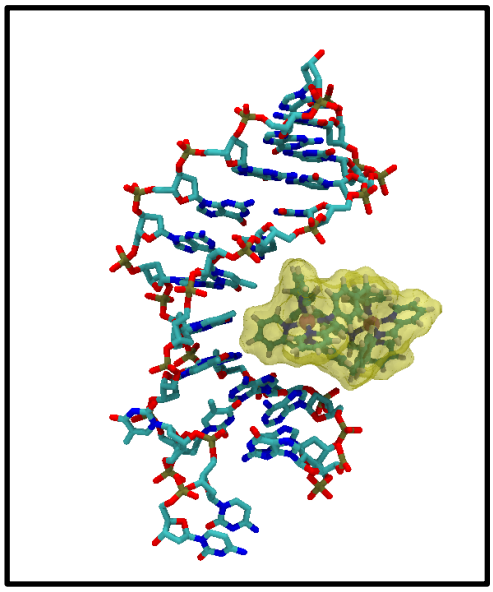

Figure 8. Interaction of a supramolecular cylinder with the major groove of DNA.

Determining the relative strength of interaction of two or more different molecules provides important details of their suitability as groove binding ligands, particularly in the context of sequence-dependent binding. Comparison of the relative free energy values potentially allows identification of the stronger binder, which has been demonstrated (*81*). Dai *et al* (*82*) show the mediation of interaction between two helices by multiply-charged ions; this has implications for the stabilization of DNA within a nucleus. Ricci et al (*83*) also show the dependence of interaction energy on dsDNA flexibility, with variations in DNA configuration also displaying variation in binding energies.

Other potential targets for ligands or drugs include G-quadruplexes, described earlier. Hou *et al* have shown a potential mechanism for ligand binding to a quadruplex, based on the displacement of the ions that normally bind between the bases. It was demonstrated that new telomerase inhibitors could be designed *de novo* using molecular dynamics and docking to examine the quadruplex (*84*).

1.5 Single-stranded DNA

1.5.1 Single-stranded DNA in water: simulations and limits to simulation

Molecular dynamics simulations of DNA reported in the literature are dominated by those focussed on ds DNA. In comparison, there are far fewer studies of ssDNA, reflecting the prevalence of the two within biological cells. In addition to its relative scarcity within the cell, practical difficulties in simulating ssDNA are also likely to be contributing factors to the low

number of reported simulations. These difficulties in simulating ssDNA include its much greater conformational flexibility relative to dsDNA (*85*). As dsDNA is relatively rigid biopolymer, it is reasonable to expect that a force field parameterized for double-stranded B-DNA may underestimate the flexibility of the single-stranded form. Indeed a recent comparison of GROMOS (*86*), CHARMM (*42*) and AMBER (*40, 44*) force fields has shown that the two AMBER force fields produce rigid ssDNA molecules, particularly if starting from a strand extracted from a B-DNA double helix (in press). Thus while the AMBER ParmBSC0 force field is generally regarded as the most reliable for dsDNA simulations, it is perhaps one of the least appropriate parameter sets for modelling ssDNA. These issues of force field reliability have somewhat hampered the development of ssDNA simulation.

One rare example of a single strand of DNA simulated in a simple water box is reported by Martínez *et al* (*87*). This study explores the effects of Na^+ ions on the dynamics and structure of poly-T ssDNA strands. The simulations show that the lifetime of Na^+ - ssDNA interactions is of the order of hundreds of picoseconds. Despite their relatively short lifetime, these interactions contribute to the stability of folded, base-stacked segments of DNA. Simulations of RNA and the synthetic nucleic acid analogue, PNA reported by Sen & Nilsson (*88*) demonstrated that base-stacking plays a key role in the stabilising the overall structure of single-stranded nucleic acids.

Exploration of the full conformational landscape accessible to a single strand of DNA is impractical on the timescale of current state-of-the-art equilibrium atomistic molecular dynamics simulations, simply because of the vast number of conformational states available to the DNA, and the energetic barriers that separate them. Attempts to improve sampling of the conformational landscape include the use of metadynamics. This is an accelerated MD simulation technique that forces transitions over an energy barrier by filling the appropriate energy well with an extra potential (*89*). It has been successfully used to simulate helix-coil transitions and calculate free energies of DNA hybridisation (*90*), as well as examining the suitability of two DNA force fields for non-helical structures.

1.5.2 ssDNA-protein interactions

The majority of ssDNA in eukaryotic cells is found in the cell nucleus. The interaction of ssDNA with proteins plays a key role in mediating nuclear activity in the cell such as transcription and DNA replication. In one simulation study (*91*), the movement of a helicase along an ssDNA strand was modelled using stochastic dynamics. Using calculated free energy profiles, a model of the

sliding of the protein along an ssDNA strand was developed, giving a potential mechanism for ATP to drive the motion of the protein.

Simulation studies of protein-DNA interactions reported in the literature include those focussed on the ssDNA-binding proteins (eg (*92*)) which are present in both prokaryotes and eukaryotes, as well as in viruses such as bacteriophages. The M13 phage (*93*) and the Pf3 phage (*94*) binding proteins have both been examined using restrained molecular dynamics. This technique uses NMR data both in the presence and absence of DNA to create a set of restraints, which are then used in the simulations. In both cases, a model of the DNA-protein complex was built based on the restraint data, allowing insight into the packing of DNA in the presence of these proteins. It was shown that both proteins, while not sharing any substantial sequence homology, form similar complexes with DNA.

1.5.3 ssDNA-nanotube interactions

Carbon nanotubes have many potential applications in the rapidly growing field of nanotechnology due to their unusual and unique properties. For example, they are being studied as a potential delivery vector for gene therapy or certain drugs. However, one potential obstacle to their application as a delivery vector is that due to their hydrophobic nature, nanotubes will prefer to partition into hydrophobic environment such as membranes rather than into aqueous environments (*95*). To overcome this potential problem, it has been suggested that ssDNA can be used to disperse nanotubes within aqueous phases (*96*). This has proven to be one of the most effective dispersants for this purpose. Interestingly, it seems that the efficiency of dispersion is in part, dependent upon the sequence of the ssDNA (*97*).

The structure of the nanotube-ssDNA complex was explored in (*98*). The results suggested that the ssDNA-nanotube interaction is stabilized by π-π stacking between the carbon nanotube surface and the DNA bases. The ideal structures of poly-GT dimers proposed in (*97*) were further examined, and it was concluded that the proposed structures were not energetically favourable. The effect of salt concentration was investigated on the interaction between the ssDNA and nanotubes, with simulations of low salt concentration displaying more wrapping of the strand around the nanotube due to electrostatic effects. In high salt concentrations, the electrostatics of the ssDNA backbone are effectively neutralized, and less wrapping to shield the charges occurs. The interaction of nucleotide monophosphates was explored in further detail in (*99*). All nucleotides were found to have different interaction energies in pure water, but similar free

energies in salt solutions, once again demonstrating that salt concentration is a key factor when considering interaction of DNA with other species.

As mentioned previously, accelerated dynamics techniques are often used enhance exploration of the conformational dynamics of biological molecules. One such method is replica exchange molecular dynamics. This uses multiple parallel simulations at different temperatures to drive the system over energy barriers, and periodic swapping of the obtained conformations back to lower temperatures (*100*). This approach was used to study a $(GT)_7$ nucleic acid strand (*101*) in complex with a nanotube. Multiple ssDNA conformations were obtained with all maintaining some degree of π-π stacking between the strand and the nanotube, further emphasising the role of base stacking in stabilising ssDNA-carbon nanotube complexes.

A single DNA strand can also be inserted into the cavity within a carbon nanotube, and the insertion process has been explored in (*102*). The simulation showed relatively rapid entry of an 8mer DNA strand within 500 ps of a 2 ns simulation. It was also shown that the diameter of the nanotube is key to allowing insertion, with an absolute minimum diameter of 1.08 nm suggested. The interaction was explored in greater detail in (*103*), and it has been suggested that the hydrophobic effect is not sufficient for the binding alone. Instead, van der Waals interactions are most likely to be the biggest contributor.

1.5.4 ssDNA-ligand interactions

Classical simulation methods have been employed to investigate the interactions of small ligand molecules with ssDNA, although reports of such simulation studies in the literature are less common than those describing dsDNA-ligand simulations. Wadkins *et al* (*104*) use Monte Carlo simulations to predict the lowest energy structures of actinomycin D, an antibacterial and antitumor compound, bound to an ssDNA hairpin. The binding was shown to have sequence dependence, with base mismatches of A-G and T-T being present in the strongest binders. Overall the results compared well with experimental data presented in the same paper. Another more recent simulation study, from Rahman *et al* (*105*) examined the interaction between ssDNA and a pyrrolobenzodiazepine adduct. It was shown that the ssDNA deforms from a helical-type conformation to one that folds to encompass the ligand, forming as many favourable interactions as possible. Again, these observations were shown to be consistent with experimental observations.

1.6 Simulations of nanopores

Molecular simulations have proven themselves useful in studies of nucleic acids, and of nucleic acids in complex with proteins or solid state systems such as nanotubes. Therefore, simulations appear a viable method for use in the study of nanopores, and have been successfully used.

1.6.1 Solid-state nanopores

Solid-state nanopores have been simulated both in the presence and absence of DNA. It has been shown that the pore shape can be controlled during the pore sculpting process, allowing featured pores such as conical or hourglass-shaped pores. Simulation of a proposed solid-state sequencing device was attempted in (*106-109*), which examined both the construction of such a device and the effects of chain length and gap width on translocation, also including the effects of shaped pores. These simulations, at least in their early stages, were shown to be lacking due to adequate parameters for solid metals in most commonly used force fields, at first relying on the Universal Force Field model (*110*), which models metals essentially as hydrophobic blocks. Switching to a more appropriate 'electrode charge dynamics' model removed the worst of the effects (*108*); however, questions still remain as to the validity of a model incorporating three distinct parameter sets in one simulation.

Confined geometry simulations have been used to examine the mechanical properties of DNA in a narrow silicon nitride pore (*111*), which allowed identification of the voltages required to force translocation of a double DNA strand through a narrow pore. Simulations have also been used to quantify translocation rates of dsDNA through silicon nitride nanopores in the presence of an applied electric field (*112*), and also to show the differences in voltage thresholds required for translocation between single and double stranded DNA. It has been proven possible to simulate the electric response to the presence of nucleotides in a pore, potentially allowing fine details of the interaction between DNA and a solid state nanopore to be observed and the sequencing devices adjusted accordingly (*113*). The effect of local charges on ion distribution within the nanopore was also studied, potentially allowing understanding of local effects of a DNA strand on ionic current.

The dynamics of ssDNA with solid-state nanopores has also been studied by simulation methods. It has been shown that ssDNA can translocate through a pore of 1 nm in size, compared to dsDNA which requires a pore width of at least 3 nm (*112*). This is in part due to the extra flexibility of ssDNA compared to dsDNA. Other studies of DNA and solid surfaces reported in the literature

include the peeling of ssDNA from a solid graphite surface (*114*). Characterisation of the microscopic processes that occur when DNA is confined with nanopores, in terms of interaction with the pore, and also the conformation of the DNA itself could play a key role in the future design of nanopores for DNA sequencing devices.

1.6.2 Protein nanopores

Protein nanopores are perhaps more common simulation targets due to the relative abundance of protein-nucleic acid force fields compared to solid-state/nucleic acid force fields. As with experimental work, much attention has focused on the α-HL protein, which is also tractable to simulation work, although it does provide a challenge purely in terms of required simulation system size.

It is known from electrophysiology experiments on immobilised DNA that the orientation of the ssDNA in the 5' direction within an α-HL pore will give a current reading approximately 30% higher than the same strand oriented in the 3' direction (*115*). MD simulations have provided a possible a mechanism for this orientational discrimination (*116*). It was shown that within the protein pore, the ssDNA adopts an extended conformation with its bases tilted towards the 5' end, assuming an asymmetric conformation.

The stretching of the DNA combined with tilting of bases results in an enhanced ability of ions to pass through the pore compared to when the DNA strand is oriented in the 3' direction. Thus simulations have been able to provide a molecular-level rationale for an experimentally measured phenomenon.

The translocation speed of a DNA strand is in the order of microseconds per base (*19*), which is impractical, if not impossible, to simulate using standard MD simulation approaches. Thus to study such processes accelerated MD methods are often used. For example Wells *et al* (*117*) have used grid-steered molecular dynamics to simulate the translocation of ssDNA through α-HL. Grid-steered molecular dynamics allows the simulation of this process by first calculating the average electrostatic potential through the lumen of α-HL, and then using this potential to steer large solutes.

1.7 Considerations

Nanopores have thus proven themselves tractable to simulation, and the parameters for DNA and protein are well established in the literature. This, therefore, allows exploration of nanopores in order to better understand their physical properties, and to use simulation to predict the effects of the nanopore on DNA translocation, as well as the effects of mutation on both the nanopore and any analytes within it.

Several properties of the α-HL pore remain relatively unexplored, particularly DNA translocation both in the presence and absence of mutated residues. This is because, while the macroscopic properties of translocation such as current effects are relatively easily observed, details of the interactions formed between DNA and the protein are not. Thus, molecular dynamics simulations are of use in examining the atomic detail of these interactions. Further insight into the translocation process could be gained by examining the energy landscape of the pore interior; this would allow identification of unfavourable and favourable interaction sites between the DNA and the nanopore.

1.8 Aims

- To explore the translocation of DNA through alpha-hemolysin using molecular dynamics simulations;
- To explore the effect of mutation on pore properties such as ionic current and DNA translocation;
- To use free energy calculations to explore the interior energy landscape of the alpha-hemolysin pore

Chapter 2. Methods

2.1 Introduction: computational chemistry

Computational chemistry is a branch of theoretical chemistry with two main goals: firstly, it seeks to validate or find a molecular basis for certain experimental observables, such as kinetic behaviour or energies. Secondly, it can provide hypotheses of what may happen in a system as yet untested in wet experiment.

Computational methods such as molecular dynamics are generally well-established as methods for study of biological systems, with parameters being available for all major groups of biomolecules. The question is how to relate macroscopic observables or properties of interest to the 'microscopic' systems studied by computational methods. Molecular simulation systems cover many scales, from the few atoms that can be studied with quantum methods, through to mesoscale-type methods that use large particles that represent dozens of atoms. All simulations that allow examination of specific details of molecular interactions, however, simulate quantities of atoms that are well below molar quantities. The question, therefore, is: how is it possible to relate the observed properties of such a small, tens-of-thousands-of-atoms system to the molar, 10^{23} quantities of atoms or molecules observed in the macroscopic world. Statistical mechanics is the branch of physics that seeks to bridge the gap between physical properties observed within a small group of atoms and the macroscopic properties observed in the whole.

Statistical mechanics as a branch of physics was developed by several people, but most credited with its formulation is Ludwig Boltzmann, who the Boltzmann distribution of molecular velocities is named for. Other major contributors were James Clerk Maxwell and Josiah Willard Gibbs.

An example of such a property is temperature: at the macroscopic scale, temperature is a measurable quantity of a medium, such as air or liquid, and can be measured by something as simple as expansion of a quantity of mercury. The microscopic equivalent, on the other hand, is something far more complex. Temperature is an *average* property of a group of atoms. Measurement of the temperature of a single, or even a small group, of atoms becomes much more complex; a small number of atoms will not provide a significantly significant result. In the example of molecular simulation, simulation temperature is determined based on kinetic energy of a group of atoms: therefore, all atoms within a group must fit into a distribution such that the average energy is the correct temperature.

2.2 Quantum and classical mechanics

The microscopic world is a quantum world. As such, one might expect quantum mechanics to provide the best description of the system of interest. The time-dependent Schrödinger equation (*118*) provides an exact description of the wavefunction of a quantum system over time; however, this level of calculation is impractical for anything but a handful of atoms; even with approximations, quantum methods scale poorly, often as n^3: simply put, the time required to compute scales to at least the cubic power of the number of atoms involved (*119*). Quantum mechanical methods are useful in small systems, and are required in simulations where explicit description of electrons is necessary – for instance, the breaking and forming of bonds (see, for example, (*120*)).

However, many simulations use classical mechanical descriptions of atomic systems. Classical mechanical simulations do not represent electrons explicitly, and consider a system in a Newtonian fashion. Bonds are represented as harmonic oscillators; atoms generally have their electrostatic nature represented as a partial charge. The adoption of classical mechanics as a simulation method allows much larger systems to be simulated, but of course sacrifices any description of electrons. The observed properties in a classical mechanical simulation can be related to macroscopic properties through statistical mechanics.

2.3 Statistical mechanical ensembles

In statistical mechanics, an *ensemble* is a group of *microstates* that can be used to derive properties of the *macrostate,* the macroscopic properties. A single microstate is a configuration of the system of interest that will occur with a finite probability. A collection of all possible microstates is essentially the macrostate. Therefore, enough microstates must be analysed for the computed macroscopic properties to be valid. Different ensembles are available, which represent different macroscopic states.

The *microcanonical* ensemble is the simplest ensemble to derive, and represents a thermodynamically isolated system. Also known as the NVE ensemble, where N is number of molecules, V is volume and E is system energy. As it is an isolated system the number of molecules, volume, and total system energy are kept constant.

The *canonical* ensemble represents a system that can exchange thermal energy with its environment, and is considered as a small system in contact with an infinitely large thermal

reservoir. Number of molecules and volume are constant; temperature is now a constant due to interaction with the large external reservoir. In simulation terms, number of molecules and volume are relatively simple to keep constant; constant temperature is slightly more computationally expensive, requiring constant rescaling of molecular velocities to keep overall temperature a 'constant'.

The *isothermal-isobaric* ensemble is a system both in contact with an infinitely large thermal reservoir and an infinitely large pressure 'reservoir'. The constancy of pressure therefore allows the simulation volume to fluctuate, as pressure is a function of force exerted on an area and is therefore dependent on individual molecular motion at a given time.

The *grand canonical* ensemble relies on constant chemical potential, volume and temperature. This constant chemical potential effectively allows exchange of molecules with an external reservoir as long as the chemical potential is a constant; in simulation terms, this usually means deletion or addition of atoms.

2.4 Classical simulation methods

There are two main classical simulation methods. Monte Carlo simulations use random sampling to generate a set of representative structures. A 'move' to another structure is randomly generated, and if it fulfils the required criteria it is accepted. Otherwise, the move is rejected. Monte Carlo simulations do not have a time dependency and as such cannot track dynamic processes. Molecular dynamics simulations, by contrast, track the time evolution of a system from a starting structure and a given set of atomic starting velocities.

2.4.1 Molecular dynamics

Classical mechanics are Newtonian in nature, and as such the central principle of MD is therefore related to Newton's second law of motion, ie

$$\vec{F} = m\vec{a} \tag{1}$$

Where F is the force acting on the atom, m is the mass of the atom involved and a is the acceleration acting on the atom. Calculation of the forces is therefore required. The forces acting on an atom can be derived from the potential energy of that atom at a given time, ie

$$F = -\frac{dU}{dr} \quad (2)$$

Where F is the force acting on the atom, dU is the derivative of the potential energy and dr is the derivative of the distance. The equations of motion are solved in a series of discrete timesteps. That is, the force calculation is performed and the atomic positions updated as if they had been allowed a short period of movement. The timestep is dependent on the model and parameters used, but in the case of a constrained all-atom MD simulation it is usually the short time of 2 fs (*121*). Timesteps are chosen such that solving the equations of motion in a single timestep does not allow sufficient movement for atomic radii to overlap on the next timestep; this usually causes a cascade effect whereby the large forces cause greater movements, which in turn cause more atomic overlap, leading to system instability and potentially causing failure of the simulation. Typical system size in a molecular dynamics simulation is in the thousands to hundreds of thousands of atoms range.

An integrator is the mathematical function that solves the equations of motion at each timestep. This can be done in several ways. A common numerical integrator in molecular dynamics simulations is the velocity-Verlet integrator (*122*), which uses positions *r* velocities *v* at time *t*. However, in the Gromacs software package used for this work (*123, 124*), the default integrator is the leapfrog integrator (*125*), which uses positions *r* at time *t* and velocities *v* at time $t - \frac{1}{2}\Delta t$; that is, the velocities are calculated at the half-timestep. A simplified integration schematic is given below.

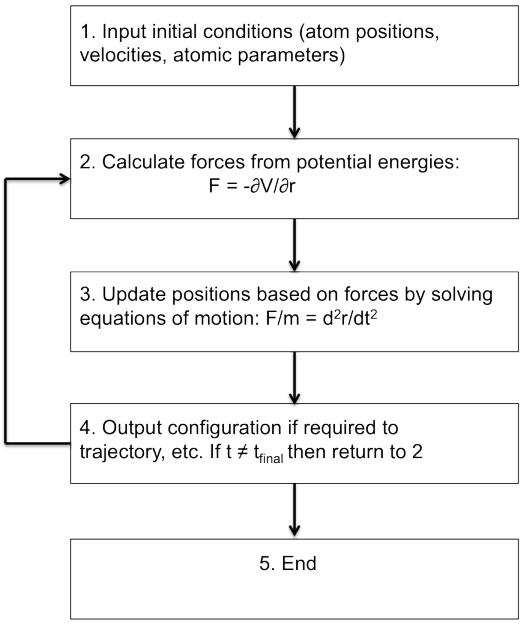

Figure 1. General molecular dynamics integration scheme

2.5 Force fields

It has been shown earlier that the forces acting on a particular atom can be related to the potential energy through derivation. Calculation of the forces for the equations of motion requires the calculation of the potential energy of the system at every timestep. A force field is both a data set that contains atomic information required for computation of forces and potential energies, and the equations that govern these interactions. There are many available force fields, most designed for specific tasks. Probably the most commonly used force field families for biological molecules are the AMBER (*126*), CHARMM (*127*) and GROMOS (*128*) families. For this work, the GROMOS G53a6 force field (*129*) was the most commonly utilised force field. Force fields as applied to the systems used will be discussed in more detail in a later chapter.

The force field is particularly important as it provides the details and the functions required to calculate the potential energies of the system, and by extension the forces based on equation (2). While the functional form in all cases differs slightly, the basic principle boils down to the calculation of two potential energy components: the bonded terms, and the nonbonded terms.

Thus, the basic potential energy equation is:

$$V = \sum bonded + \sum nonbonded \tag{3}$$

where V is the potential energy. The bonded term is the sum of all bonded potential energies in the system; the nonbonded term is the sum of all nonbonded pairwise potential energies. This can be further expanded into the individual components of each. Bonded terms include bond stretching potentials, angle potentials and torsion or dihedral terms. The nonbonded component encompasses electrostatic interactions, calculated using Coulomb's law, and Lennard-Jones or van der Waals type interactions. The expanded equation is:

$$V = \sum bonds + \sum angles + \sum dihedrals + \sum Coulomb + \sum van\ der\ Waals \tag{4}$$

2.5.1 Bonded interactions

Bonded interactions are simply interactions between two or more connected atoms. There are several types of bonded interaction: bond stretching, angle bending and dihedral or torsion interactions.

Bond stretching interactions represent the oscillation of a particular bond between two atoms. The simplest way of representing this is as a harmonic potential, such that

$$V = \frac{1}{2}k(r - b_0)^2 \tag{5}$$

Where V is the potential energy, k is the spring constant, b_0 is the equilibrium bond distance and r is displacement from equilibrium. Bond vibrations, however, occur on a timescale that is faster than the average molecular dynamics timestep; therefore, some of the details of these vibrations are lost. This can be avoided through the use of bond constraints, discussed in more detail later.

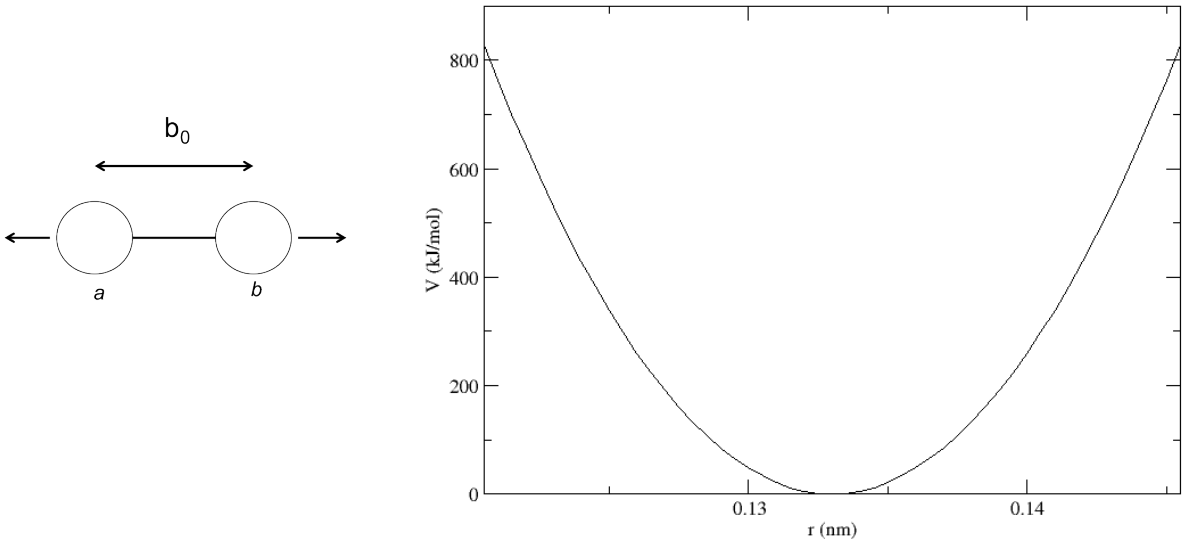

Figure 2. Principle of bond stretching (left) and potential form (right). b_0 is at the centre of the potential well.

Angle bending interactions represent the bending of the angle across a group of three atoms. Again, this is usually represented as a harmonic potential, in the form

$$V = \frac{1}{2}k(\theta - \theta_0)^2 \tag{6}$$

where V is the potential energy, k is the spring constant, θ is the angle displacement from equilibrium, and θ_0 is the equilibrium angle.

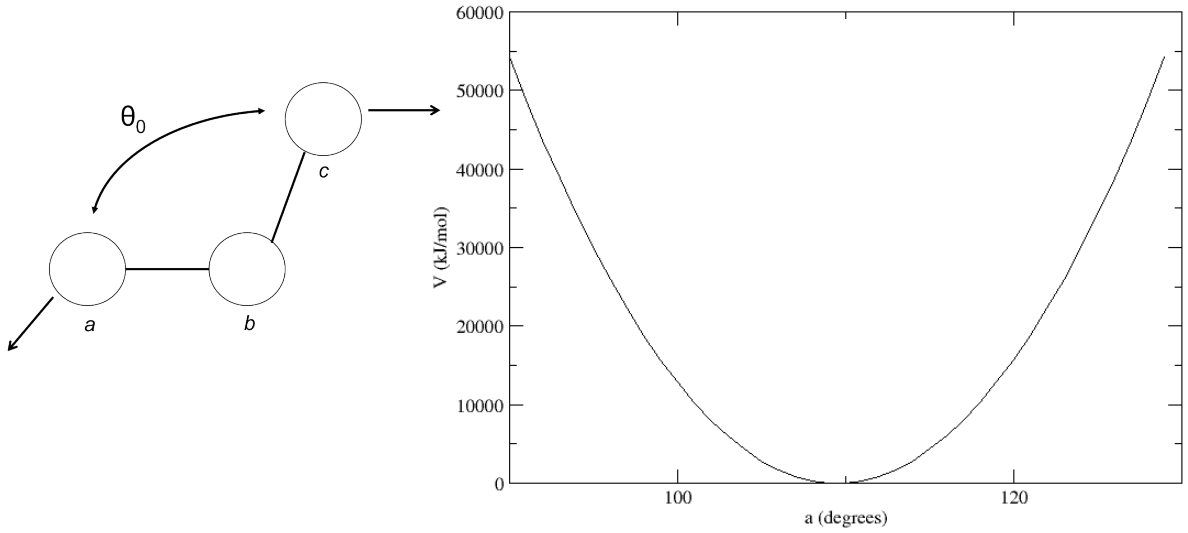

Figure 3. Principle of angle stretching (left) and functional form (right). θ_0 is at the centre of the potential well.

Dihedral angle potentials come in two categories, proper and improper. Proper dihedrals represent the rotation of a group of four atoms around a central bond, such that the arrangement is:

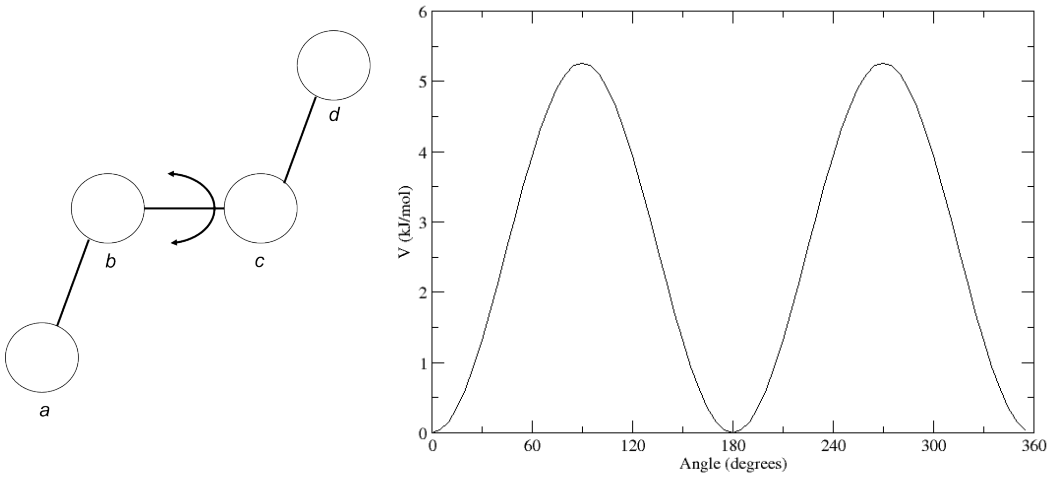

Figure 4. Principle of dihedral angle rotation (left) and functional form (right)

The governing function is instead a cosine potential instead of a harmonic potential due to the nature of the rotation. The function is:

$$V = k(1 + \cos(n\emptyset - \emptyset)) \tag{7}$$

Where V is the potential, k is the spring constant and Ø is the angle between the ijk and jkl planes, shown in the figure above. The cosine function is necessary as 360 degree rotation is possible, thus a simple harmonic function does not represent a dihedral term correctly. The *n* parameter essentially determines the number of periods within a 360 degree rotation; that is, the number of favourable potential wells.

Improper dihedrals are used, again for a group of four atoms, but under different circumstances. Proper dihedrals allow free rotation about the central bond; improper dihedrals, by contrast, are used to keep a set of four atoms in the same relative spatial configuration. This mostly covers aromatic ring groups – where the configuration must be kept planar – and tetrahedral centres, where the arrangement is kept fixed to prevent chiral flipping. The improper dihedral function is harmonic, such that

$$V = \frac{1}{2}k(\xi - \xi_0)^2 \tag{8}$$

where V is the potential, k is the harmonic constant, ξ is the displacement from equilibrium and ξ_0 is the equilibrium position. This therefore has the same functional form as the bond and angle potentials.

2.5.2 Non-bonded interactions

Non-bonded interactions are interactions between two atoms not connected by a chemical bond. This includes two types of interaction: Coulombic interactions, which are interactions between charged atoms, and Lennard-Jones interactions, which include repulsive and attractive interactions between any two atom types, not necessarily charged.

The Lennard-Jones potential (*130*) provides an approximation of both a longer-range dispersion (attractive) interaction and short range repulsion between two atoms, the general form of which is:

$$V = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \tag{9}$$

The parameters **ε** and σ are specific to the two atom types that are interacting. ε determines the depth of the potential well, and σ is the (finite) distance between the two atoms where the potential is exactly zero. The parameter r is the distance between the two atoms of interest.

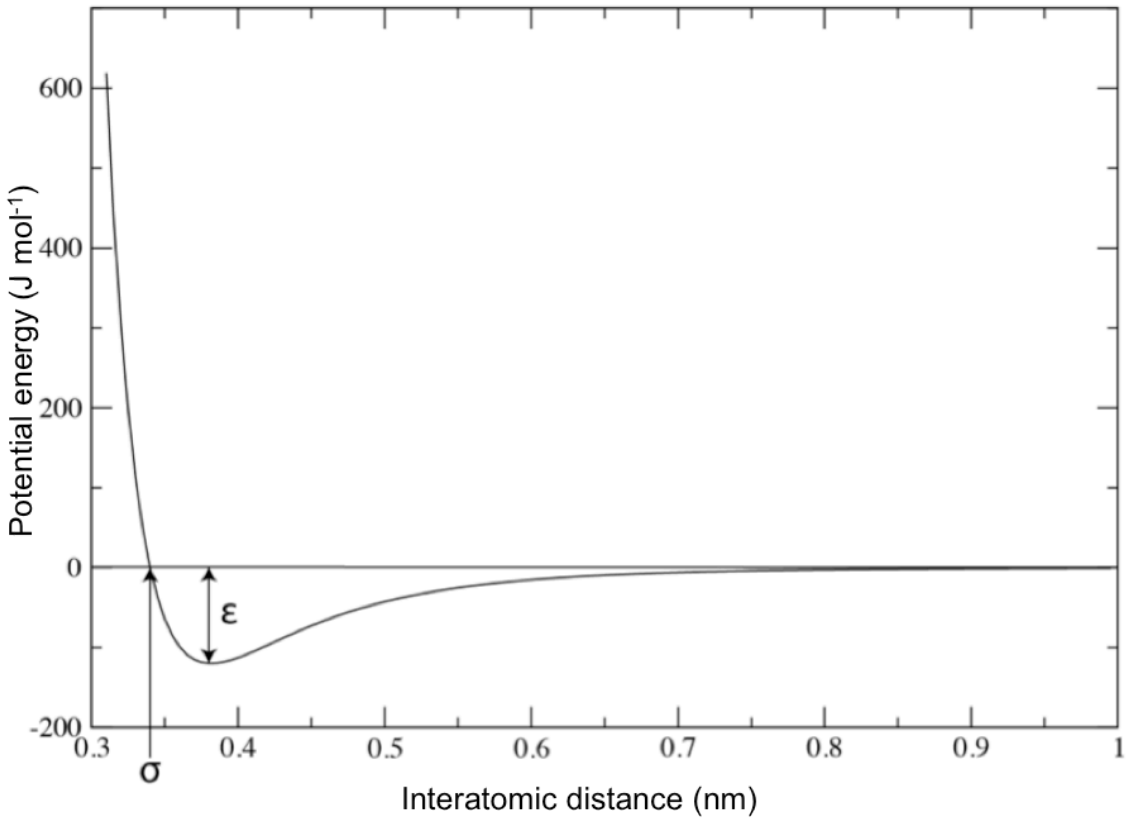

Figure 5. Functional form of the Lennard-Jones potential. At r of less than 0.3 nm, V rapidly approaches infinity. At r approaching infinity, V tends towards zero. Epsilon (maximum well depth) and sigma (finite zero potential distance) are marked.

The electrostatic potential energy between two atoms is based on Coulomb's law, the general form of which is:

$$V = \frac{1}{4\pi\varepsilon_0} \times \frac{q_1 q_2}{\varepsilon r} \tag{10}$$

Where $\boldsymbol{\varepsilon}_0$ is the permittivity of free space, q1 and q2 are the charges of each atom, **ε** is the dielectric constant of the medium, and r is the distance between the two atoms.

As can be seen from the above equations, long-range interactions are treated as pair-additive; that is, the total force on a single atom is simply the sum of all pairwise interactions between that and the surrounding atoms. As both interactions are effectively infinite-distance interactions, this

quickly leads to large computational expense in systems with many atoms. However, while these interactions are non-zero at long range, they are small enough that they can be ignored or modified to avoid their calculation.

2.5.3 Long-range cut-offs

Truncating long-range interactions by simply ignoring any beyond a given distance cut-off is possible. However, particularly in the case of Coulombic interactions, this can lead to substantial errors accumulating over time , such as abnormal structures forming just beyond the cut-off . There are several ways to avoid this.

Firstly, the interaction function (both Lennard-Jones and Coulombic) can be modified such that the potential is zero at the cut-off distance. This is treated in two ways: The interaction function is modified over the whole of its length, referred to as a shifted potential; or the interaction is normal up to a target distance then decayed smoothly to zero between this distance and the cutoff, referred to as a switched potential.

This still can be a problem in both cases, however, due to the lack of accounting for the longer-range components. This can be accounted for by dispersion correction terms, in the case of the Lennard-Jones-type interactions. Correcting for Coulombic terms is more difficult, however, and several methods are available.

Reaction field electrostatics (*131*) accounts for the short-range terms using Coulomb's law. Beyond the short range calculation, a uniform dielectric constant is assumed for the medium. A dipole in the short-range medium will induce a polarisation in the uniform medium, which is the reaction field. This interaction provides an approximation of the missing long range components.

Other long-range electrostatic calculation methods are based on the Ewald summation for periodic systems (*132*). The Ewald method is slow, however; instead, most long-range electrostatics are based on the particle mesh Ewald method (PME) (*133, 134*), which calculates the short range interactions by the Coulomb method and accounts for the long range interactions using fast Fourier transforms in reciprocal space.

2.6 Optimisation and control

2.6.1 Neighbour/Grid searching

Molecular dynamics calculations require the forces to be computed every timestep. In order to calculate using either Coulomb's law or the Lennard-Jones potential, then an atom must 'know' which atoms are its neighbours for the purposes of the calculation. Given that both types of interaction usually utilise a cutoff, the neighboursearch need only be carried out within the cutoff radius around an atom. This cuts the computational time required, as performing such a search for every atom within a search is computationally costly.

Furthermore, since an individual timestep and the associated movement is small, this allows the neighboursearch to only be carried out every *x* steps, usually 5 or 10, which increases computational efficiency.

Application of a cutoff to the neighboursearch can lead to some error, particularly in the case of polar molecules. Calculating only for part of a molecule can lead to artefacts such as a dipole that does not exist. To counteract this, charge groups are introduced. A charge group is a collection of several atoms with a total charge of zero; the centre of the charge group is taken as the point which is searched for in the neighboursearch step, instead of individual atoms.

2.6.2 Thermostats and barostats

The concept of a statistical mechanical ensemble requires that certain variables be constant in order to be valid. The most common ensembles in molecular simulation are NVT and NPT, both requiring a fixed temperature, and in the case of NPT a constant pressure as well. These variables by their very nature will not remain constant if not controlled in some fashion, due to both transfers of energies within the system and rounding errors caused by numerical integration. Because of this additional effort is required to fix these values.

Temperature, as a variable, is an average property of a group of molecules, and is a function of their kinetic energy, which in turn simply relates to velocities. A Boltzmann distribution is used to assign velocities to molecules at the start of a simulation, giving the system as a whole a given 'temperature'. This, as stated, will drift over time. Several methods of correcting this are available; two common methods are the Berendsen thermostat (*135*) and the Nose-Hoover thermostat (*136, 137*), both of which consider the system coupled to a large external thermal

reservoir. Simply speaking, the Berendsen thermostat exponentially relaxes a system quickly to a target temperature but does not generate a correct energy distribution across all molecules. The Nose-Hoover thermostat, while not particularly efficient for relaxing a system to a given temperature, gives a correct energy distribution once equilibrium has been reached.

Barostats follow a similar approach, through coupling to a large external pressure 'reservoir'. The Berendsen barostat (*135*) allows relaxation quickly to a target pressure but does not necessarily generate accurate fluctuations, and other more complex barostats such as the Parrinello-Rahman barostat (*138*) that are poor at large relaxations but generate a correct distribution of fluctuations at the target pressure.

2.6.3 Periodic boundary conditions

Simulation systems number in the thousands to low millions of atoms, which is small in relation to Avogadro's number. A simulation system of this size is therefore finite, and boundary conditions become a factor in the behaviour of the system – in particular, rounding errors created by the deflection of molecules from the box walls become a problem. Periodic boundary conditions avoid this by considering the simulation system as a periodic image within an infinitely repeating system. Simply put, a molecule exiting one side of the box will re-enter on the opposite side.

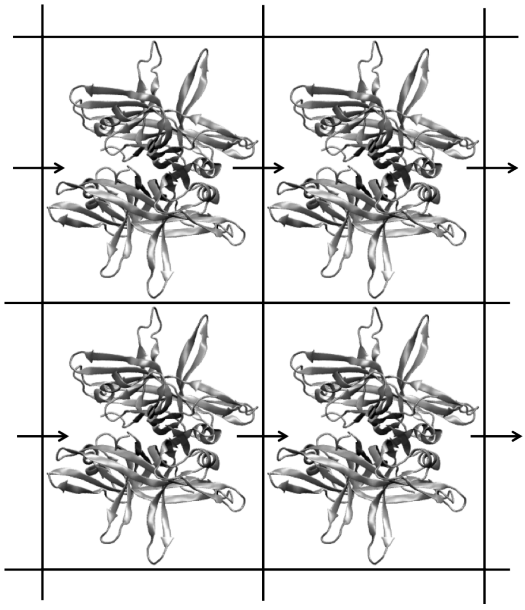

Figure 6. Basic principle of periodic boundary conditions. A movement in an individual image happens in all images in the same direction, and an object leaving on one side of an image will 'reappear' on the opposite side.

Most implementations of periodic boundary conditions utilise the minimum image convention; that is, that only one image of a periodic particle is considered for the purposes of calculations

(although this is not true in the case of long-range electrostatics calculations). Another consideration of a periodic system is that one image of particle must not be able to 'see' the periodic copy of itself in the adjacent unit cell; box size and force cut-offs must take this into consideration.

2.6.4 Atomic representations

Any model created must take into account what is required of it. There are multiple possible ways of simulating any given molecular system, and technically none are 'correct', in the sense that the method must fit the question being asked. This is particularly apparent when considering the possible representations of atomic detail that can be utilised.

All-atom models treat all atoms explicitly; this includes nonpolar hydrogen atoms. All atoms therefore are considered in the force calculations. All-atom models therefore provide the most detail of any model, but are the most costly in terms of computational time. The CHARMM (*127*) and AMBER (*139*) force field families use all-atom models.

The calculations on nonpolar hydrogens are particularly costly relative to the effects they have on a system; such interactions are thought to be a minor part of the overall interaction. To avoid this, united atom models typically remove the nonpolar hydrogens, accounting for them by increasing the radius and mass of the adjacent carbon atom, effectively creating methyl or methylene 'atoms'. This is particularly effective in the case of lipid tails, which contain large numbers of nonpolar hydrogens. If the model is correctly tuned, a united-atom representation can give results comparable to an all-atom bilayer. The GROMOS force field family uses the united-atom representation (*128*).

Further simplification can be made on a similar principle to the creation of united atoms. Coarse-grain models combine several heavy atoms into a single interaction site. The particulars vary depending on the exact model, but each particle commonly represents four atoms, such as in the MARTINI model of Marrink et al (*140*), used in the simulations of proteins and lipids.

2.6.5 Bond constraints

A chemical bond between two atoms can be treated as a harmonic oscillator. However, bond vibrations of this nature occur rapidly, with a typical timescale that is less than a standard molecular dynamics timestep. This therefore means that a bond vibration cannot be accurately

represented in the standard timestep. Two solutions are possible: the first is to simply reduce the timestep (this would normally be to 1 fs instead of 2 fs). However, this will effectively double the time required to simulate the same process. The second solution is a different treatment of the bond vibration.

Instead of a harmonic oscillator, bonds are instead treated as constraints in the equations of motion. In short, the equations of motion are carried out on, for example, a water molecule, with all atoms allowed to move in the directions dictated by the equations of motion. After the change in position, there is a slight drift between the atomic positions relative to the standard water structure. The next step updates the atomic positions based on the defined constraints to recreate the correct structure.

2.6.6 Parallelisation methods

Two basic methods are available for parallelising a molecular dynamics simulation: particle decomposition and domain decomposition.

Particle decomposition, previously the only implemented method in Gromacs version 3 (*124*), requires that each processor is assigned a set of atoms at the beginning of a simulation. Each atom must 'know' the position of at least half of the rest of the atoms within the system; this leads to large numbers of coordinates being required to be communicated between nodes with each timestep; at even relatively low numbers of processors this produces noticeable slowdown (eg 16 processors compared to 8).

Domain decomposition assigns instead a domain or region of the simulation box to each processor, and each processor is responsible for calculating the forces for that region. This reduces the number of node-node communications required, as internode communication is only performed after several timesteps. Assigning nodes to calculate only the electrostatics when using the particle mesh Ewald method can further increase efficiency.

Benchmarking results for Gromacs 4 and Gromacs 3.3 are given below in table 1, taken from (*123*). The results are given in nanoseconds per day for a simulation of a 23,000-atom protein in water system, performed on a 3 GHz Intel Core 2 Infiniband cluster. No results are reported for 64 and 128 cores in Gromacs 3.3 as no further speedup is achieved.

Code	4 cores	8 cores	16 cores	32 cores	64 cores	128 cores
Gromacs 4	3.1	6.0	11.6	21.6	38.0	60.1
Gromacs 3.3	2.7	4.8	7.0	8.4	n/a	n/a

Table 1. Scaling of the Gromacs code on multiple processors, using different versions of the code. Taken from ref. 122.

2.6.7 Restraints

Restraints are added to a simulation when an atom or group requires fixing in place, either as part of an equilibration or as part of an unstable system. Restraints can be added in several ways. Firstly, position restraints fix the position of an atom by applying a harmonic potential to the atom when it moves from the equilibrium position. The magnitude of the potential is proportional to the distance travelled from the equilibrium point.

Distance restraints take another atom as a reference rather than an absolute spatial positon, applying a potential once the distance exceeds the equilibrium value. This is often of use in simulations utilising NMR data, as NMR data is essentially a collection of distance restraints. In an early example (*141*), distance restraints were used in the simulation of the crambin protein, derived using NMR data.

The final 'restraint' option simply freezes atoms in the dynamics, ie they feel no interaction forces and do not move. While all restraint options lack some degree of realism, this is probably the least 'realistic' as the atoms under such a restraint do not move at all.

2.7 Other methods

2.7.1 Nonequilibrium molecular dynamics

Nonequilibrium methods involve applying external forces to a simulation system, or beginning with an internal structure that is not at equilibrium. External forces can be applied in the form of pulling forces, where a harmonic spring is attached to the atom(s) of interest. The other end of the spring is attached to a point in space. This point in space is moved at a given rate and along a given vector; the effect of the spring on the atom it is attached to moves the atoms based on the rate of movement and the spring constant. Nonequilibrium methods can also be used to derive free energies through the use of the Jarzynsky equality, such as in (*142*).

An electric field can also be applied to a simulation box. This appears to take the form of a constant force applied to atoms in proportion to their charge; thus, a completely neutral atom will feel no force whereas a charged molecule will tend to move in the appropriate direction to the field, either along or against depending on whether the molecular charge is positive or negative.

It is also possible to create a system that is far from equilibrium and observe the effects on, for example, ion transport. This can be done, for example, by adding a double bilayer to create two separate components, and allowing ion flow between the two assuming both compartments have radically different ion concentrations. This has been attempted in (*143*), where it was shown to be possible to stabilise a highly charged peptide in a membrane environment through use of a transmembrane ion gradient created by a double bilayer system.

2.7.2 Free energy methods

Free energy calculations are possible using molecular dynamics; they can also be performed using other techniques such as Monte Carlo simulations. Several types are possible; one of the most common is potential of mean force, or PMF, calculations. There are multiple ways of calculating a PMF; one of the most rigorous – and therefore most computationally expensive – is umbrella sampling (*144*). For an umbrella sampling calculation, multiple windows are created with the atom or molecule of interest spaced at small intervals, either half an angstrom or an angstrom. The molecule is restrained in the direction of the reaction coordinate of interest, say the z axis, and simulations are run for a sufficient sampling time. A sufficient time is one that allows for sampling of all the rotational degrees of freedom of the molecule at a particular point.

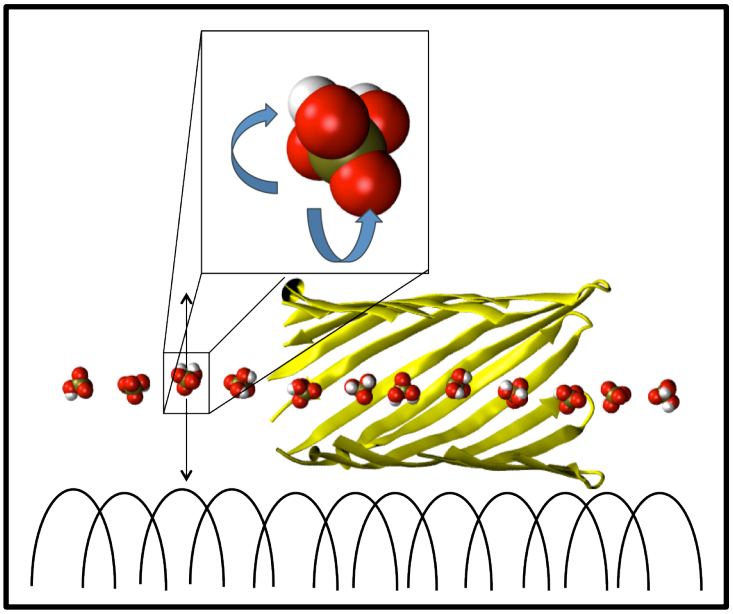

Figure 7. Illustration of umbrella sampling. A molecule in an individual window is allowed free rotation, and freedom of movement in two planes, but cannot move far in the third plane without the umbrella potential returning it to its original position.

If a molecule is in an energetically favourable position, then it will not act against the restraint. If it is unfavourable, it will attempt to change position. The forces acting on the restraint are recorded. The free energies of the system are calculated using the weighted histogram analysis method, or WHAM (*145*). This removes the bias added by the restraints and reconstructs the free energy profile.

The basic umbrella sampling method has been used in multiple contexts; for example, it has been shown to be possible to calculate free energies of drug molecules within a lipid bilayer (*146*), or to examine base pair opening in a DNA strand (*147*).

2.8 Simulation details

For the studies presented in this work, the Gromacs code (version 4.0.7 or 4.5.1 as appropriate) will be used, in conjunction with the Iridis3 computing cluster. Individual chapters contain the methods used that are specific to each study.

Chapter 3: Development and testing of a reduced model pore system

Abstract: While molecular simulation has been established as a useful tool in the study of biological systems, and has been used in the study of nanopores, it was thought necessary to test the simulated pore properties compared to their experimental values. In the process, however, it was discovered that the large simulation systems required in the study of a full α-HL heptamer in a bilayer is impractical with limited resources. As such, research then focused on designing simplified models of the wildtype α-HL pore and various mutants that preserved the essential properties of the pore while allowing much more rapid, high-throughput simulation. The models produced mimic well the properties of the full α-HL nanopore, both in observed current values and translocation properties. Multiple metastable conformations of the ssDNA are observed to form in the translocation process. This translocation occurs in a staggered fashion, with short but significant lengths of 'tethering' particularly between DNA backbone phosphates and positively charged residues within the pore, leading to a potential explanation of the mechanism of translocation.

3.1 Introduction

Molecular dynamics simulation has been established as a suitable tool for the study of biological systems (*148*) and has been used successfully in the simulation of nanopores in various contexts (*149, 150*). Initial simulation work focused on testing of an α-hemolysin (α-HL) homoheptamer within a lipid bilayer. The constructed simulation system consisted of the α-HL homoheptamer, 384 DMPC phospholipids, ~60,000 water molecules, and sodium chloride to a concentration of 1M.

Stability of the full protein within a bilayer was examined by simply performing a molecular dynamics simulation of the system for 20 ns and observing bilayer and protein stability. This was done in the absence and presence of an electric field. In both cases the protein showed little change from the starting structure. The bilayer, however, had a tendency to porate when exposed to a high electric field (equivalent to a membrane potential of > 500 mV), forming large holes that completely disrupted membrane integrity. This tendency was corrected for by addition of position restraints to the phospholipid head groups, leaving the lipid tails with some freedom of movement.

Satisfied with the stability of the system, the simulation setup was validated against experimental data. Of particular interest in the context of nanopore sensing experiments is ionic current. As ionic currents are the basis of nanopore sensing, in particular the partial current observed when an analyte is within the pore, it is therefore important to be able to recreate these in silico. A comparison was made with electrophysiology data, particularly taken from (*28*). As current data for mutants as well as the wildtype protein are available, models of the mutant pores were also generated for use in simulation. These mutants were generated using the MODELLER package (*151*).

Partial currents were also compared between experiment and simulation, and were created by addition of a poly-G 12mer to the pore. Poly-G was chosen as a test nucleotide as it is, by size, the largest base and thus served as a theoretical model of pore occlusion at its maximum. This strand was positioned so that it crossed transmembrane β-barrel, with two bases remaining in the vestibule above the constriction site formed by E111/K147. Position restraints were applied to these two bases, allowing the rest of the strand to move freely within the barrel.

3.1.1 Simulation setup

The partial current simulation systems consisted of 1 α-HL heptamer in a dimyristoyl-phosphatidylcholine (DMPC) lipid bilayer consisting of 384 individual lipid molecules. DMPC was chosen both for the reason that the lipid bilayer coordinates and parameters are readily available, and because α-HL is often reconstituted into DMPC bilayers in experiment. The system was solvated with ~70,000 SPC water molecules, and ions added to a concentration of 1 M for a total of ~1,400 Na^+ and Cl^- ions. For the partial current simulations, a 12mer poly-guanosine molecule was pulled into the α-HL pore using the GROMACS pull code, with a pulling speed of 1 x 10^{-6} nm ps^{-1} and a force constant of 350 kJ mol^{-1} nm^{-2}. The ssDNA was positioned such that the first 2 bases were above the constriction site formed by E111 and K147, with either the 3' or 5' end oriented towards the *trans* exit (the base of the beta barrel) depending on experiment. Position restraints were applied to these first two residues, while the rest of the strand was allowed to move freely. Simulations were run for a total of 20 ns simulation time.

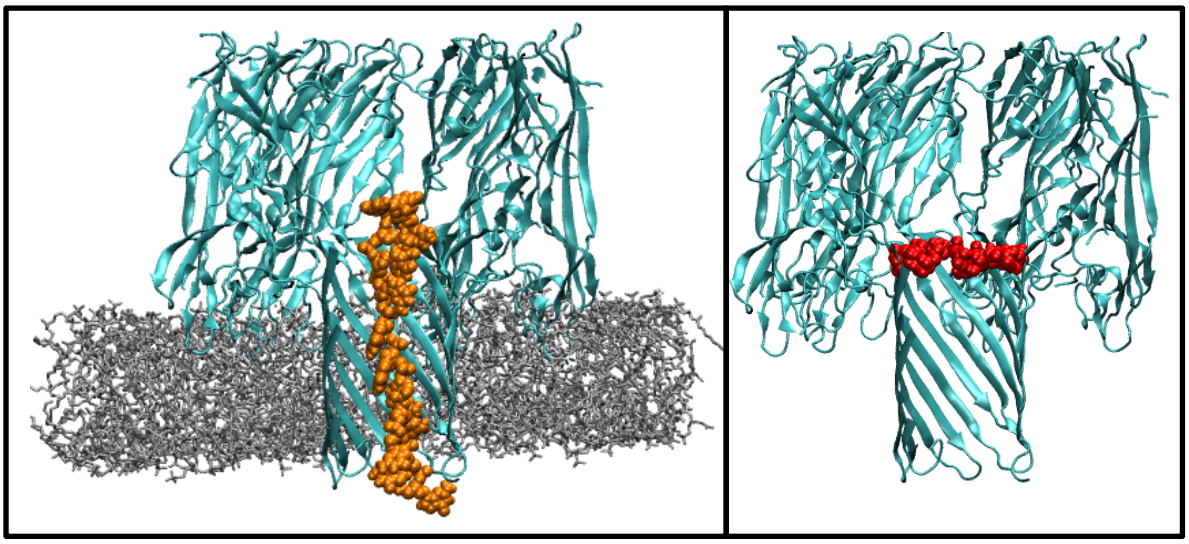

Figure 1. Full system setup for partial current testing. Left: Cutaway through protein. Bilayer in grey, DNA 12mer in orange, protein in cyan. Water and ions omitted for clarity. Right: location of E111/K147 constriction site within the pore.

3.1.2 Preliminary results

Initital testing of the membrane- α-HL system showed promise in the comparison between mutants and partial current values reported in Maglia *et al*. Both open and partial currents are given in table 1.

Table 1. Partial ion counts in full α-HL in the presence of ssDNA. Second and third columns are number of positive and negative ions respectively that fully cross the length of the transmembrane β-barrel in 20 ns of simulation time. Column 4 gives the cumulative total of ions. Column 5 ranks the simulated fractional currents from highest current to lowest current; the 14T4N1Q protein is discounted as it was simulated but not tested experimentally. Column 6 gives the rank order of partial currents in each mutant based on the experimental data given in (27).

System	No. +ve ions	No. –ve ions	Cumulative no. of ions	Rank order (simulated currents)	Rank order (experimental pore currents)
Wildtype	58	12	70	2	2
14T4N1Q	102	12	114	(1)	-
E111N-WT	39	56	95	1	1
M113R-RL2	24	22	46	4	5
M113R-WT	33	35	68	3	4
RL2	11	12	23	5	3

Given above are the ion counts and cumulative ion counts; observed open pore current; observed fractional current; and rank orders. The rank order in simulation compares well to rank order from experiment, with the exception of the RL2 mutant, although the partial current is in general considerably higher than the relative experimental partial current.

However, given available computational resources it became clear that this method of data gathering was impractical, with a single result on a 250,000-atom system taking two months to generate using 8 CPUs on the available computing resources. As such, attention turned to creating a simplified model pore system that could be used *in lieu* of the full protein-membrane system, while maintaining the essential properties of the pore.

3.2 Generation of the model pore

Experiments using α-HL often use mutated forms of the protein, tailored to provide specific translocation properties (*28*). These mutants almost without exception occur within the α-HL transmembrane β-barrel region. This region also contains the major constriction site – ie, the narrowest region of the pore – formed by glutamate 111 and lysine 147. Two more exist, at the upper region of the vestibule, formed mostly by lysine 8, and and the base of the pore, formed by aspartate 127. The constriction sites are thought to be the main determinant of current readings and therefore nanopore sensing – being the narrowest point, they are the limiting step in current flow.

With both this major constriction site and the surrounding mutatable residues being in the small transmembrane region of the protein, it was thought possible to design a pore based solely using the major constriction site and the mutatable residues. As the target is as simple a pore system as possible, the pore in question would use the E111/K147 constriction site as the upper mouth, and N123 as the residue at the pore exit, N123 being the last residue that is commonly mutated in translocation experiments.

Simplification of the protein was the initial goal; however, to reduce required computational time as much as possible, further reductions in system complexity were made. The membrane-facing sidechains were removed entirely and replaced with glycine residues. The protein backbone was restrained in position, with a restraining potential of 1000 kJ mol^{-1} acting on each backbone atom, with the sidechains allowed freedom of movement. The simplified protein chain is shown in figure 2A. As the membrane itself was a significant fraction of the system, it was replaced with a hydrophobic slab consisting of united atom methane 'atoms' restrained in position, initially with a restraining potential of 1000 kJ mol^{-1}. This slab consisted of a series of methane rings around the protein barrel itself, sitting within a rectangular methane surface. This is shown in figure 2B.

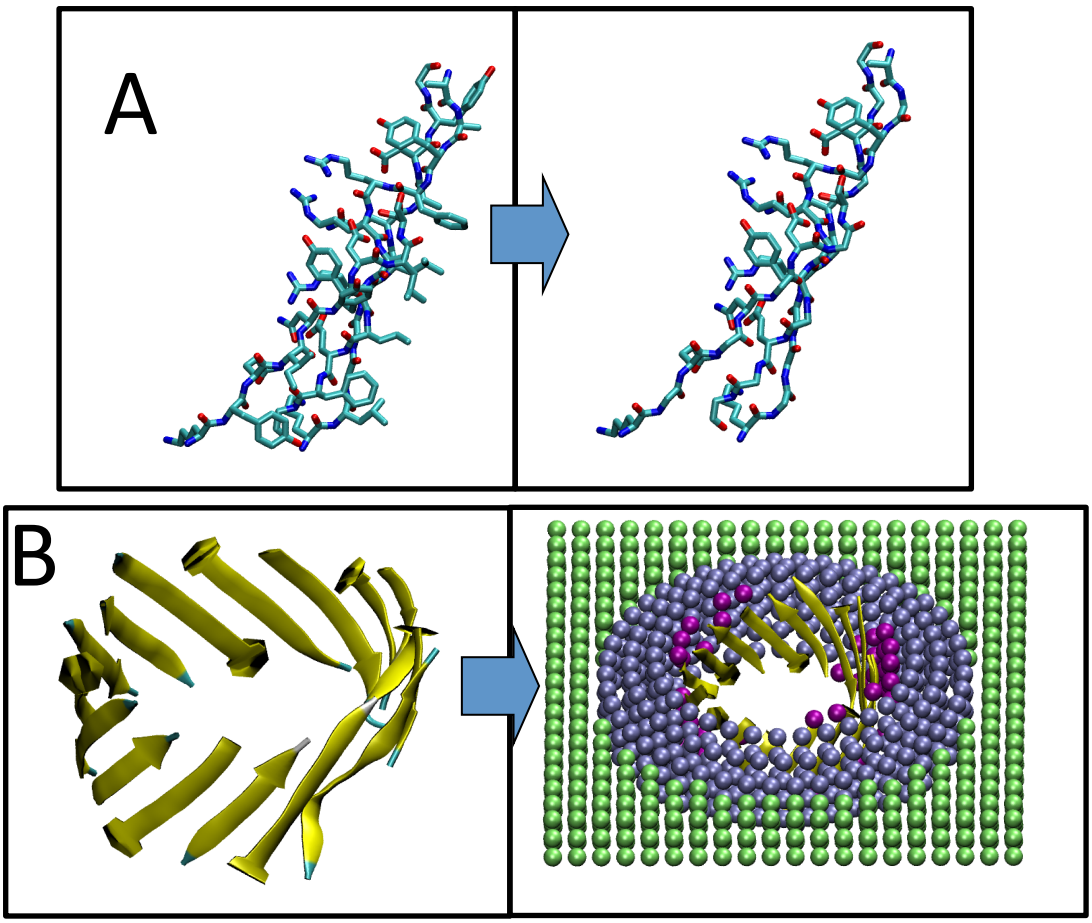

Figure 2. Schematic of model pore generation. Figure 2A: conversion of protein to model protein by removal of membrane-facing sidechains. Figure 2B: Placement of protein within methane slab; rings of methane surrounding protein are given in purple, rectangular slab given in green.

A further question remained as to whether NaCl was a suitable salt to use in these simulation experiments as most α-HL electrophysiology experiments are performed in the presence of KCl. As such, validation of both the available NaCl parameters and KCl parameters was deemed necessary. This required the conversion of GROMOS KCl parameters to a format compatible with the G53a6 force field, and required testing of current readings relative to experiment.

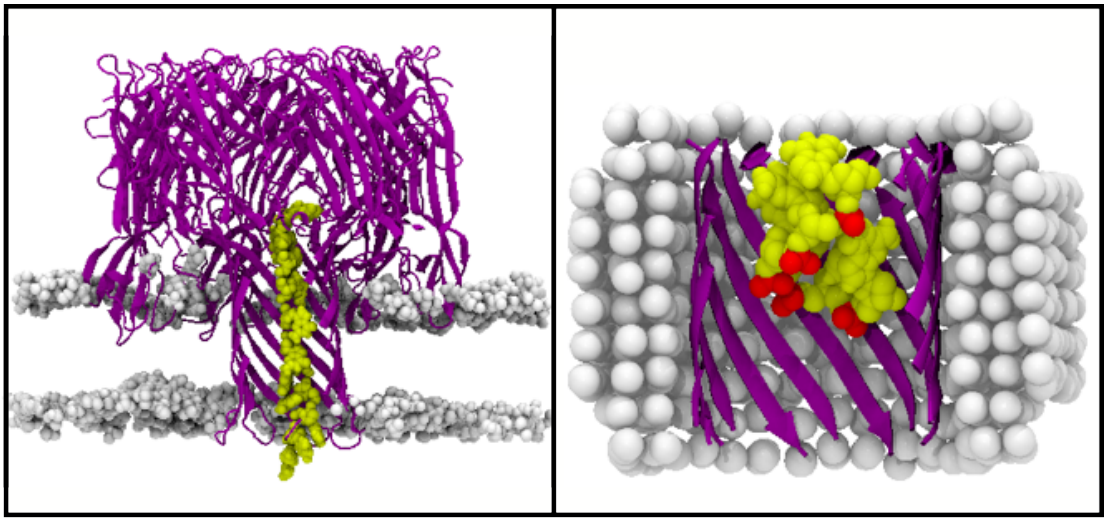

Figure 3. Comparison of the full α-HL protein system (left) and model pore system (right). Left: Protein in purple, DNA in yellow, lipid headgroups in silver. Right: model pore protein in purple, slab in silver, DNA in yellow, DNA phosphate groups in red.

3.3 Methods

3.3.1 Validation of simulated ion parameters

In order to determine how well the available ion parameters mimic actual experimental current data, simulations were carried out using NaCl and KCl solutions, in simulation boxes of 5 x 5 x 5 nm, at concentrations of 0.1 and 1 M. A uniform electric field was applied along a single box axis, at strengths of 0.008, 0.04, 0.2 and 1 V nm^{-1}. Currents were calculated using as in (*149*) using a purpose-written script. All systems were simulated for a total of 10 ns, after test simulations using 0.1 M NaCl showed quick convergence to a stable current value within 4 ns of simulation start (figure 4). The initial 2 ns were discarded as equilibration time, and the remaining 8 ns of simulation used for the final analysis.

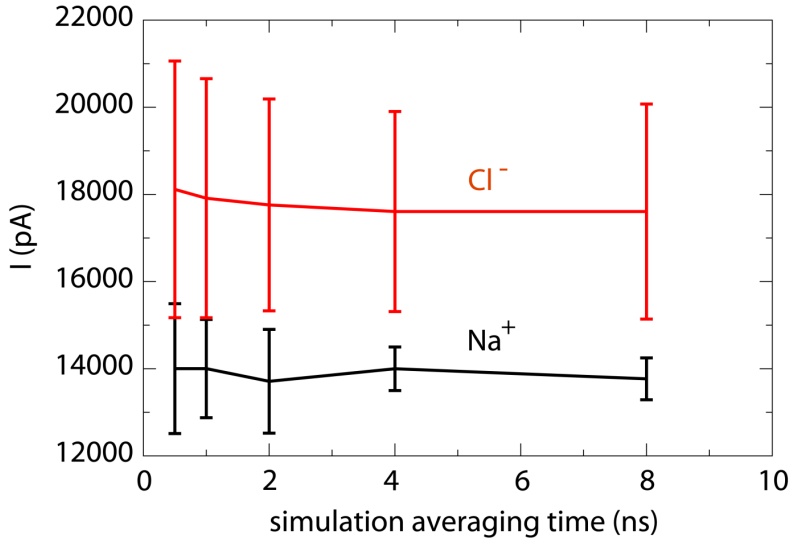

Figure 4. Convergence of current measurements. Convergence occurs within 4 ns.

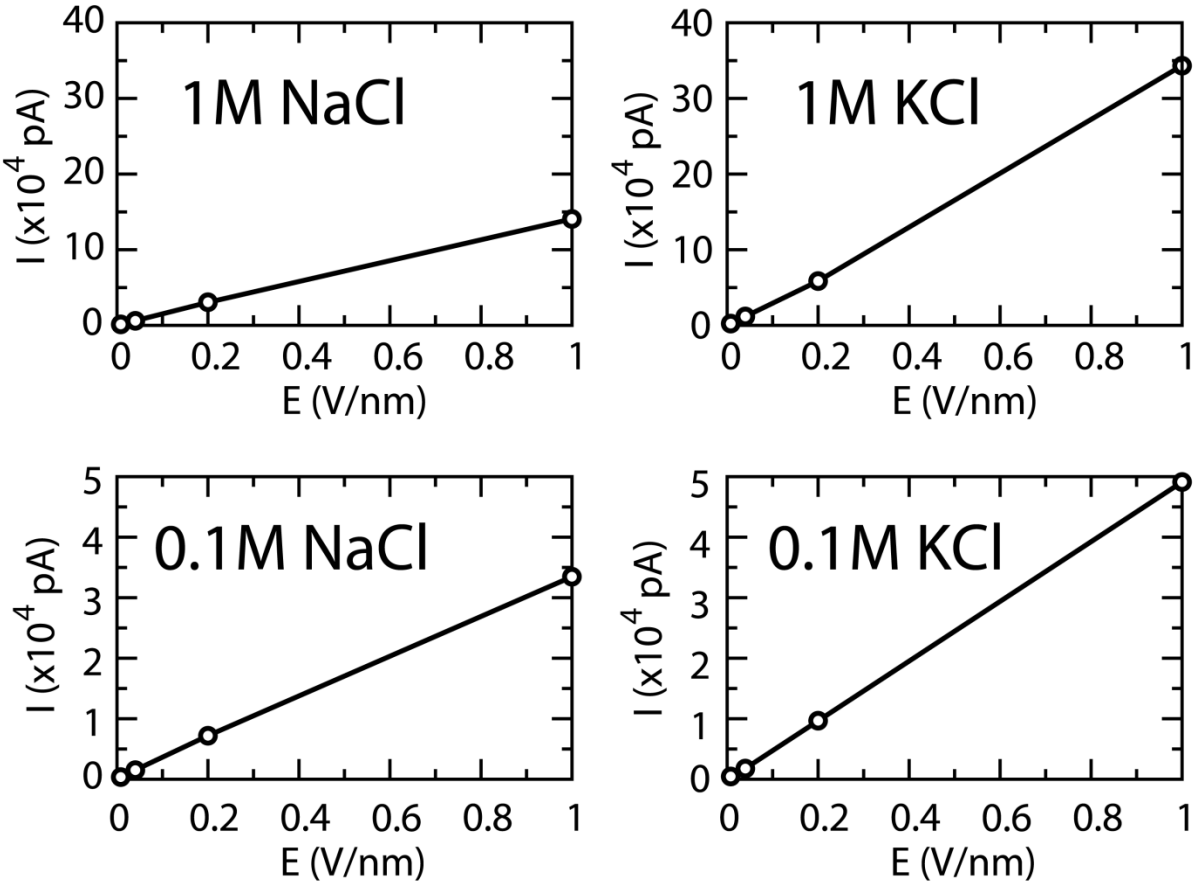

Figure 5. Relationship between simulated mean current and applied electric field for various electrolyte solutions.

The calculated current-electric field relationships are given in figure 5, and are, as expected from ions in solution, linear relationships. Observed currents for KCl were on average somewhat higher than those of NaCl. From this data, bulk conductivity values were extrapolated, given in table 2. These show reasonable agreement with the experimental data as given in (*152*).

Table 2: Conductivity measurements for each electrolyte solution.

Electrolyte	Simulated Conductivity (S/m)	Experimental Conductivity (S/m)
1M NaCl	5.59	8.16[1]
0.1M NaCl	1.33	1.38[1]
1M KCl	13.85	10.8[1]
0.1M KCl	1.97	1.28[1]

[1]From CRC Handbook of Chemistry and Physics, 86th Ed.

Partial current ratios based on bulk mobilities were also calculated; these are given in table 3 (KCl) and table 4 (NaCl). As is expected, there is no dependence between partial current ratios and applied electric fields, with the possible exception of at low voltages, which is likely a sampling issue. The observed KCl ratios are between 0.8 and 1.0, with an observed experimental value of 1.04. The NaCl ratios are between 1.6 and 1.7, compared to 1.52 experimentally. Further work using the KCl parameters in the model pore system also revealed some anomalous results, with observed currents being far higher than expected. Whether this was a general issue with the KCl paramters used, or an issue in the conversion to the correct format, remains unclear. With the NaCl ratios being satisfactory compared to experiment, and with fewer anomalies observed in testing, NaCl was chosen as the salt for the pore current experiments.

Table 3: Partial current ratios for 1M KCl electrolyte solutions at varying field strength.

E (V/nm):	0.008	0.04	0.2	1
$I[Cl^-]$ (nA)	1.1	5.8	27.4	153.9
$I[K^+]$ (nA)	1.2	5.9	31.1	189.6
$I[Cl^-]/I[K^+]$	0.9	1.0	0.9	0.8

Table 4: Partial current ratios for 1M NaCl electrolyte solutions at varying field strength.

E (V/nm):	0.008	0.04	0.2	1
$I[Cl^-]$ (nA)	0.8	3.9	19.3	86.5
$I[Na^+]$ (nA)	0.7	2.3	11.1	54.1
$I[Cl^-]/I[Na^+]$	1.2	1.7	1.7	1.6

3.3.2 Experimental IV curve measurements.

Experimental data for the IV curve measurements using NaCl was obtained through collaboration, using the protocol given in (*153*).

3.3.3 Simulation methods

Simulation setup consisted of the model pore placed in a 5 x 5 x 8 nm box, and solvated with ~4,500 simple point charge (SPC) water molecules. NaCl salt was added to the system using the GROMACS genion script, for a total of ~100 Na^+ and Cl^- ions. Including the pore and the ~1,000 methane slab atoms, total system size was approximately 20,000 atoms, less than one-tenth of the size of the full α-HL heptamer in a lipid bilayer. For the DNA translocation simulations, an ssDNA hexamer was added to the system above the mouth of the pore. This hexamer was polyG, chosen due to guanosine being the largest DNA base and therefore the most likely to show any effects on currents while within the pore. This hexamer was created using the Builder module of the Quanta program of Accelrys Inc. Pore mutants were created using the MODELLER program (*151*).

Initial tests, both with the full system and the model pore system, revealed that the likelihood of DNA entry into the constriction site was low, most likely due to the energy barrier created by the narrow constriction formed by K147/E111. While this entry may have eventually occurred, it was impractical to wait for a translocation event on simulation timescales. As such, for the remainder of the simulations the DNA was pre-threaded so that the terminal base was beyond this constriction, thus negating the barrier to translocation.

At each stage of setup – addition of protein to slab, addition of solvent, addition of DNA – energy minimisation was performed using the steepest descents algorithm. The system was equilibrated with position restraints of strength 1000 kJ mol^{-1} applied to non-solvent atoms, while the solvent and ions were allowed free movement. Further equilibration removed the restraints on DNA and protein sidechains, keeping the protein backbone and methane slab atoms restrained. Further testing revealed that the protein backbone did not necessarily require restraint, remaining within the slab even under an electric field; however, these restraints were kept as stability was not assured.

Simulations were typically 20 ns, with extension of up to 100 ns in the case of simulations where translocation of the ssDNA began but did not complete. Each simulation was repeated for a total

of at least 3 simulations per system. An electric field was applied across the simulation box, equivalent to a potential of 600 mV across the slab (approximately 0.18 V nm^{-1}). Electric field values used were typically higher than those used in experiment, which are usually potentials of 100 – 150 mV. The reason for this discrepancy is twofold: firstly, to ensure that ssDNA translocation completes on a satisfactory timescale, and secondly, to ensure that enough ion translocation events are observed to provide accurate current values. Temperature and pressure were controlled at 300 K and 1 bar respectively, using the Berendsen thermo- and barostats, with tau values of 0.5 ps. Non-bonded interactions used a short range cutoff of 1.2 nm, with long-range electrostatics treated using particle mesh Ewald (PME)(*133*) Simulations used the 53A6 variant of the GROMOS96 force field (*154, 155*).

3.4 Results

3.4.1 Ionic currents

Ionic currents are central to the nanopore detection process; as such, validating the model pore system as a useful mimic of the full α-HL pore first required comparison of the observed current values with those derived from experiment. The IV data obtained through collaboration was used as a comparison, and extrapolated to give an IV curve up to the higher potentials used in simulation.

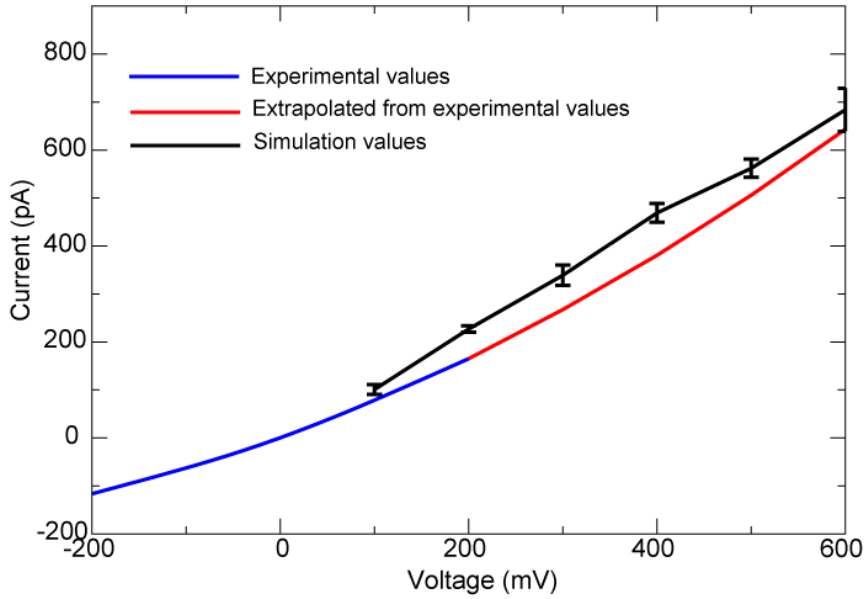

Figure 6. Simulated (black) and experimental (blue; extrapolated values in red) IV curves.

The obtained IV curve is given above in figure 6, with extrapolated data marked. Extrapolation was performed using regression analysis on the experimental data, and assumed the slight nonlinearity of the IV relationship would continue. The IV curve for the model pore was obtained

by running multiple simulations at different potentials. The observed currents in the model pore are in general slightly higher than the extrapolated IV curve for NaCl; however, they remain with the same order of magnitude.

With the open pore currents of the wildtype pore being satisfactory compared to experiment, a set of pore mutants was generated based on the work of Maglia *et al* (*28*) These mutants included the M113D and E111N mutants, both mutants at the constriction site, and a range of mutants that introduce arginine residues at different positions within the pore. The open pore currents were tested with reference to the data contained in (*28*). The open pore currents are extrapolated to a potential of 600 mV. The results are given in table 5 below.

Table 5. Simulated open pore currents in wildtype and mutant pores, compared to experimental data.

Mutant	Simulated open pore current (pA)	Extrapolated experimental open pore current (pA)	Rank orderings (sim/exp)
(Wildtype)	489±9	567±4	6/6
E111N	824±53	694±15	1/2
M113D	680±53	634±17	2/3
M113R	580±38	709±10	3/1
T115R	519±13	572±9	5/5
T117R	427±32	390±20	9/9
G119R	460±22	562±5	7/7
N121R	439±8	486±10	8/8
N123R	521±14	618±19	4/4

The results, while not preserving absolute current values, generally preserve the same ordering of current values, as seen to an extent in the full protein simulations. High current mutants such as the E111N mutants are seen to be high current in simulation; conversely, the generally lower current values of the range of arginine mutants are also preserved. The data sets were compared by rank ordering; that is, simply ranking the values with a number from highest to lowest. A comparison of the rank orderings of each data set shows good agreement through all simulations.

The next step in the testing of the model pore systems beyond open pore current values was to examine the effects of introduction of DNA into the pore. The data in (*28*) also includes partial current data in the presence of DNA, providing a basis for comparison between simulation and

experiment. Simulations were run in the presence of a DNA hexamer under an electric field, allowing for examination of both partial currents and DNA translocation properties. Table 6 gives the simulated ion counts compared to experimental partial current values for each mutant. It should be noted that ion counts are used in place of actual current values as some mutants showed an ion flux that was too low to calculate a reliable current value with the method used in calculating the open pore currents.

Table 6. Residual currents and translocation behaviour of the pore-ssDNA system. Residual ion count in column 2 is taken from the point at which two bases have crossed the E111/K147 constriction site, and is defined as an ion crossing the length of the beta-barrel between the E111/K147 constriction site and the N123 pore exit.. Currents in column 3 are extrapolated based on data in (27). The categories in column 4 are based on no entry into pore (none); at least 2 bases entering but translocation incomplete (partial); and translocation completing (full).

System	Simulated residual ion count	Extrapolated partial current (pA) (*28*)	DNA translocation in 20 ns
WT	80±16	69.74	Partial
E111N	114±12	76.34	Full
M113D	n/a	n/a	None
M113R	17±2	28.36	Partial
T115R	7±6	12.36	Partial
T117R	24±14	17.16	Partial
G119R	15±9	11.24	Full
N121R	14±20	14.58	Partial
N123R	23±6	11.7	Partial

Again, similar trends are observed between experimental data and simulation. The relatively high residual currents of the E111N mutant (and wildtype) are preserved in simulation. The low residual currents of the arginine mutants are also preserved. The low partial current values in the arginine mutants are likely a result of the presence of what effectively amounts to an extra constriction site, with arginine possessing a relatively large sidechain, and in the presence of DNA the pore is occluded, resulting in a lower overall current. It may also be that the presence of an extra charge in the pore in general results in a lower current due to sidechain-ion interactions, as observed in the open pore mutants. For example, removing the negative charge provided by E111 in the E111N mutant results in both a higher partial and open pore current relative to wildtype.

3.4.2 DNA translocation

While the model pores show good recreation of open pore and partial currents, such experiments are simply a recreation of experimental data. The next step was to use the model pore predictively, in examining translocation behaviours of ssDNA through the transmembrane beta-barrel. While macroscopic properties of translocation can be observed through experiment, the microscopic details of these interactions between protein and ssDNA cannot.

In terms of general translocation properties, all mutants show some level of translocation, with the exception of the M113D mutant. The E111N protein shows the fastest translocation of any mutant, frequently completing within the 20 ns of simulation time. This was perhaps somewhat unexpected, as observed translocation times in the full E111N α-HL are longer than wildtype and on a par with some of the arginine-type mutants, which are generally known for slower rates of translocation relative to wildtype (*20*). However, the experimental data also shows the frequency of translocation events in E111N are some 10 times higher than most other mutants. It may be more that this ease of translocation is being recreated; the observed translocation speeds across all mutants are generally at least an order of magnitude away from the real values anyway, and as such absolute velocities are hard to recreate. This ease likely stems from the removal of a negative residue at the constriction site, which would be otherwise repulsive to the negatively charged ssDNA backbone.

The general incompleteness of translocation in the arginine-type mutants, with the exception of G119R, is likely due to the attractive interactions between the ssDNA backbone and the positive charge of the guanidinium side chain, which is explored in more detail later. These mutants generally show both relatively slow translocation and a low frequency of translocation, about half as fast and half as frequently as wildtype α-HL. Extending some of these simulations up to 100 ns tended to show less completion of translocation compared to extended wildtype simulations, due in part to the backbone-guanidinium interaction mentioned above.

The result obtained in ssDNA translocation through the M113D mutant was also encouraging. This mutant is non-translocating in experiment, and shows the same behaviour in simulation, with none of the simulations showing any form of entry of ssDNA into the pore. This was further explored by pre-threading of the ssDNA further beyond the K147 constriction site; in these circumstances the ssDNA still does not translocate, and was instead ejected from the pore.

These interactions between the sidechains and the ssDNA play a major role in the translocation process. Charged residues in particular are key, shown by the effects of manipulating negative charges (E111N, M113D), and addition of positive residues (R mutants). A frequently observed interaction is the K147-backbone phosphate interaction. The K147 ring of positive residues is part of the constriction site at the mouth of the pore, and is preserved across all mutants. This leads to a 'tethering' of the DNA backbone to this residue, due to the interaction between the negatively charged phosphates of the backbone and the amine group of the lysine sidechain. Also observed are hydrogen bonding-type interactions of the E111 sidechain carboxylic acid with either bases or sugars of the ssDNA; these interactions are generally much shorter lived. Both types of interaction are shown in figure 7. These interactions are relatively long lasting on simulation timescales, existing for a nanosecond or more. This leads to a translocation process that is staggered, which is particularly exaggerated in the case of mutants involving the introduction of an arginine residue, providing another site for this tethering to occur.

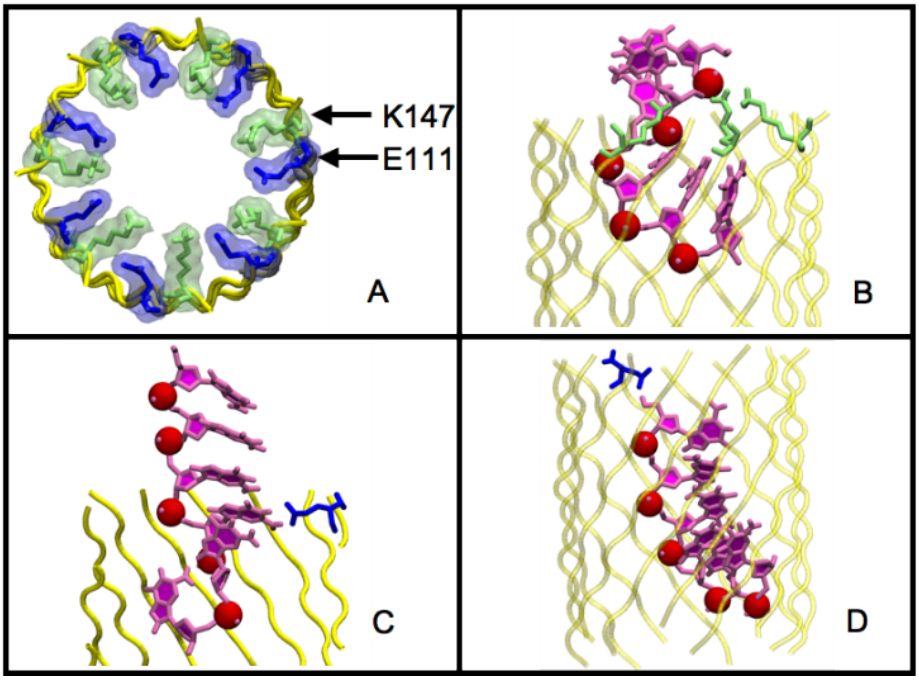

Figure 7. Interaction of the K147/E111 constriction site with DNA. (A) Location of the ring of K147 (green) and E111 (blue) residues. Pore backbone coloured yellow, in tube representation. (B) Interaction of three K147 sidechains with two DNA phosphates. The phosphorus atoms of the DNA backbone are shown as red spheres. Bases and sugars are coloured pink. (C) Inter. (D) The interaction of an E111 sidechain with a sugar hydroxyl group.

To examine the interaction of the DNA with the charged residues within the pore, the distances between backbone phosphates and sidechain centre of masses were calculated. Representative graphs for each type of interaction and are given in figures 8-10 below.

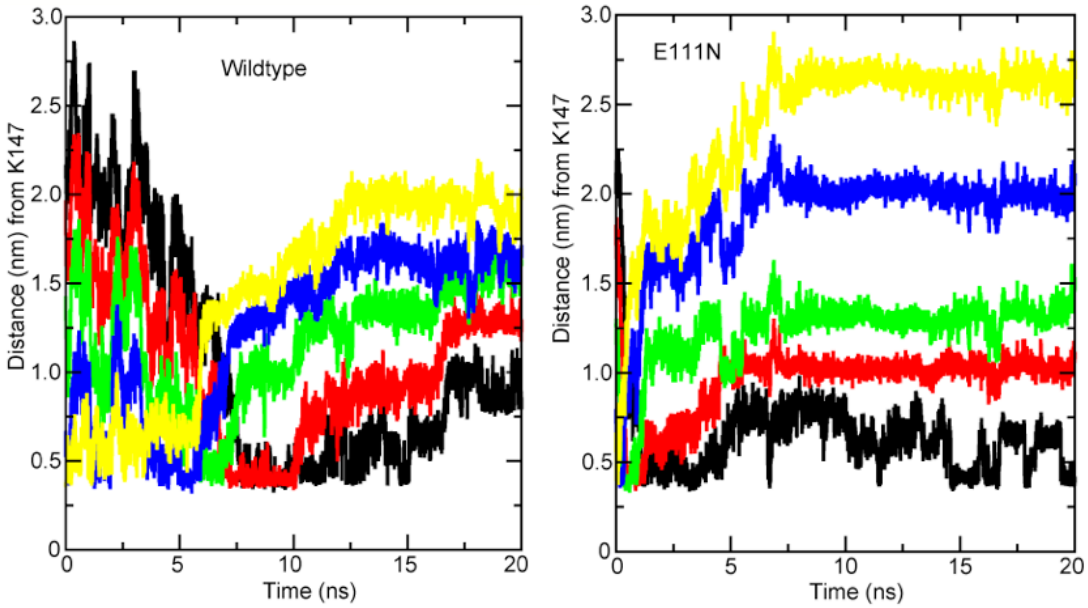

Figure 8. Distance as a function of simulation time between each phosphate group on the DNA molecule and the center of mass of the K147 sidechains for the wild-type (left) and E111N mutant (right) pore. The five phosphate groups are represented as differently colored curves, sequentially from yellow (first phosphate entering pore) through blue, green and red to black (last phosphate entering pore).

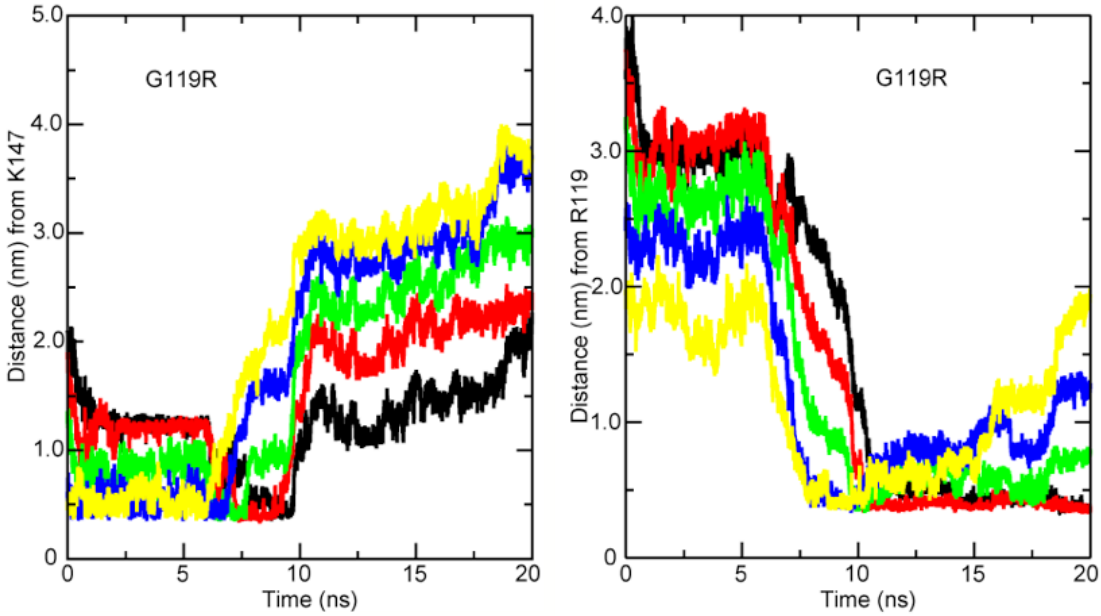

Figure 9. Distance as a function of simulation time between each phosphate group on the DNA molecule and the center of mass of the K147 (left) and R119 (right) sidechains for the G119R mutant pore. The five phosphate groups are represented as differently colored curves, sequentially from yellow (first phosphate entering pore) through blue, green and red to black (last phosphate entering pore).

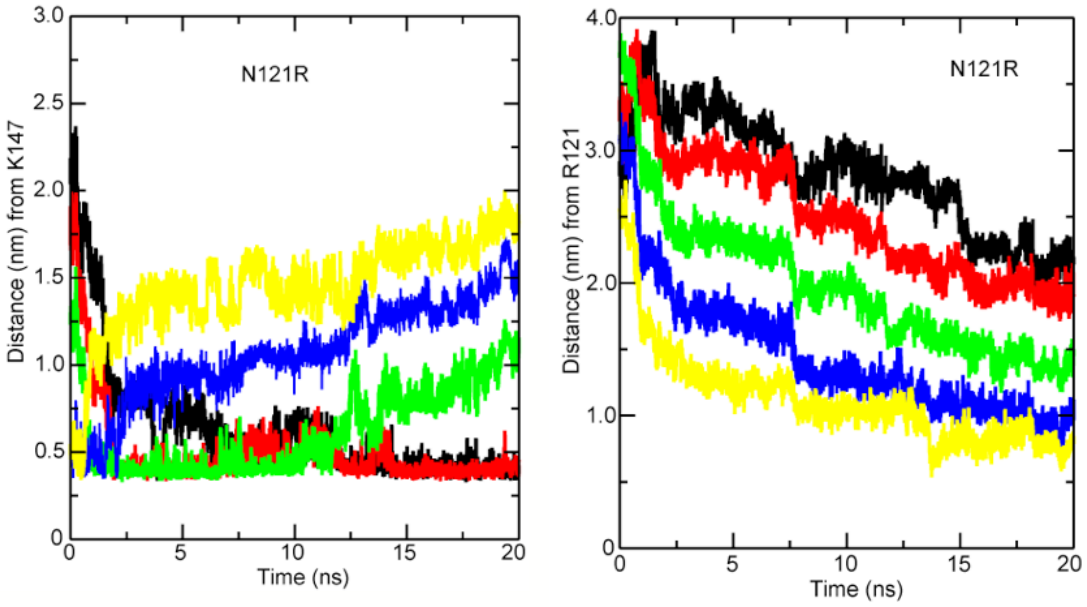

Figure 10. Distance as a function of simulation time between each phosphate group on the DNA molecule and the center of mass of the K147 (left) and R121 (right) sidechains for the N121R mutant pore. The five phosphate groups are represented as differently colored curves, sequentially from yellow (first phosphate entering pore) through blue, green and red to black (last phosphate entering pore).

The tethering is clearly visible, for example in E111N in figure 8, where some degree of interaction remains between the terminal phosphate (black) and the K147 ring for some 15 ns after the rest of the ssDNA has passed. The staggered translocation can be seen in figure 9 for the G119R mutant, where two periods of relatively little movement between 0-8 ns and 10-20 ns are divided by a period of rapid movement as the DNA passes between the K147 constriction site and the ring of positive residues formed by R119. This staggered, or 'binding and sliding', translocation mechanism is similar to the proposed mechanism of transport in other proteins (*156, 157*), and has also been observed in RNA transport through carbon nanotubes (*158*). Introduction of extra positive residues effectively provides a second binding site for the ssDNA, thus enhancing the staggered translocation mechanism. This was taken to an extreme in an engineered mutant created in a longer model pore, dubbed 3R6L, which introduced three extra rings of arginine residues at regular intervals throughout the pore separated by regions containing leucine residues. This mutant demonstrated a similar binding-and-sliding type mechanism, with binding occurring at each positive residue site for some length of time before breaking.

3.4.3 Conformational analysis

One particular feature of DNA transport through the pore was the adoption by the ssDNA of various folded states. These folded states were generally well-defined, and due to the limited conformational flexibility of such a short ssDNA could be characterised based on length. Once adopted, these new conformations could last for up to tens of nanoseconds, based on data from the extended nanopore simulations. A summary of the types of configurations observed is given in figure 11.

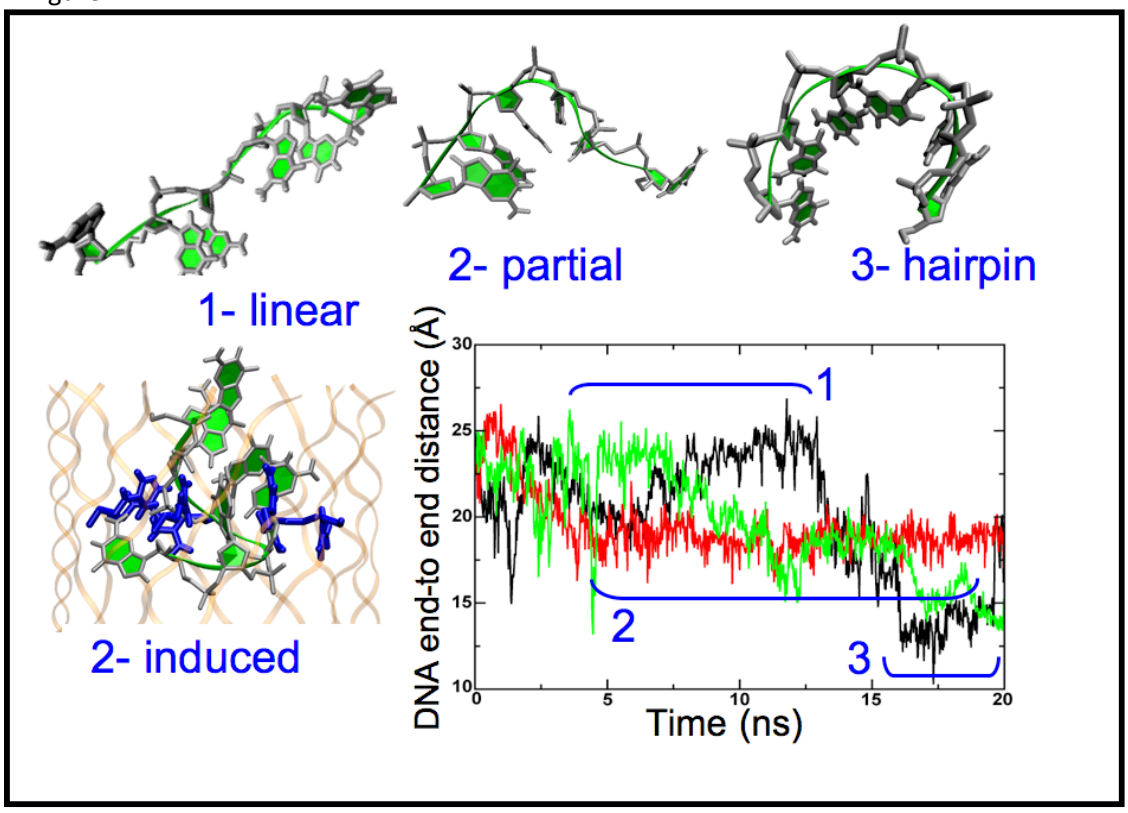

Figure 11. Summary of the observed DNA configurations, characterised by end-to end distance, with the corresponding lengths displayed in the graph. DNA is in grey and green; protein backbone is shown in orange and arginine residues are displayed in blue. Configuration 1 is observed both on the green and black lines on the graph. Configuration 2 corresponds to the region marked on the red line. Configuration 3 corresponds to the end region of the black line.

An extended ssDNA was counted as being approximately 2.4 nm in length, calculated as being the distance between terminal hydroxyl groups. This configuration was essentially linear. The partial folding was counted as having an end-to-end distance of approximately 1.9 nm. The exact configuration of this state varied slightly, with it being the result of several possible 'kinks' in the DNA. The third 'horseshoe' or 'hairpin' configuration was counted as an end-to-end distance of 1.4 nm. A second variation of the partial fold also exists, being caused by the interaction between the DNA strand and any positive residues within the pore, which appear to induce a conformational change. These end-to-end distance values were consistent across all simulations.

Different mutants exhibited different variations of this folding behaviour, shown in figures 12-15 below. The E111N mutant tended to show ssDNA translocation in the linear configuration (type 1 on figure 11), likely due to the rapid translocation observed. This rapid translocation prevents the formation of any interactions with the pore that may result in induced-type folds, and thus the ssDNA passes through virtually unchanged (see figure13, below). E111N is known to have a larger partial current with ssDNA in the pore relative to wildtype; this lack of folding coupled with the lack of an extra charged ring of residues at the pore mouth could be a potential mechanism.

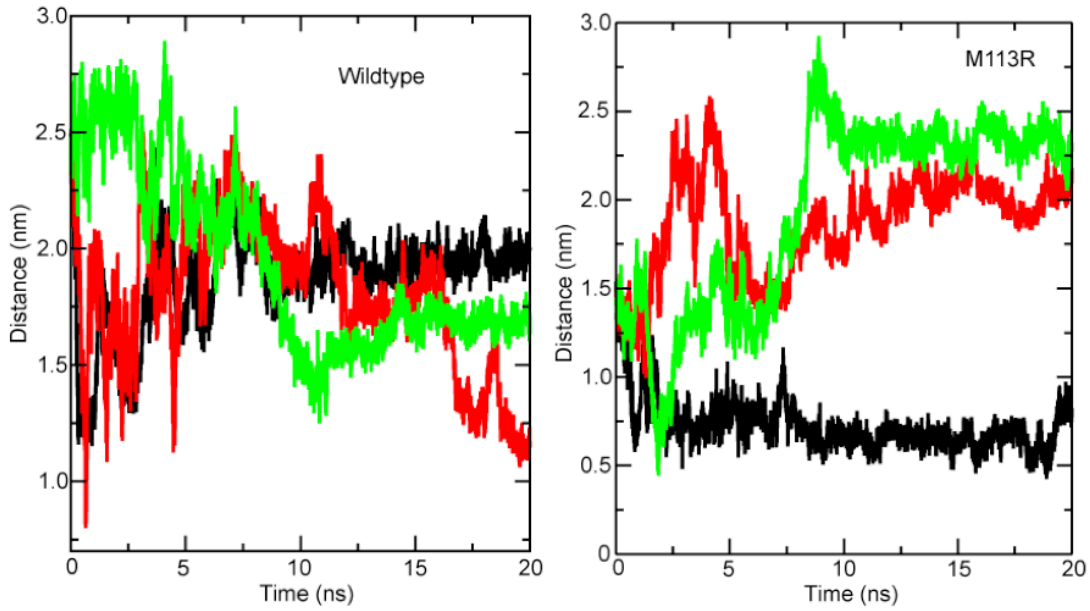

Figure 12. Distance between the centres of mass of the first and last residues of the ssDNA molecule relative to time for the wild-type pore (left) and the M113R mutant (right). Differently coloured curves represent repeat simulations.

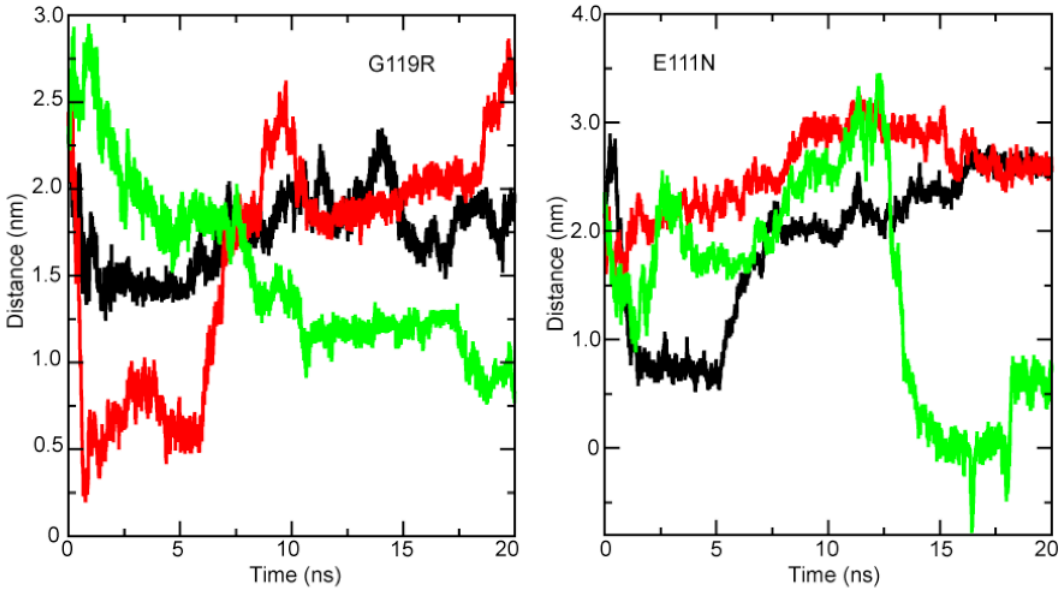

Figure 13. Distance between the centres of mass of the first and last residues of the ssDNA molecule relative to time for the G119R mutant (left) and theE111N mutant (right). Differently coloured curves represent repeat simulations.

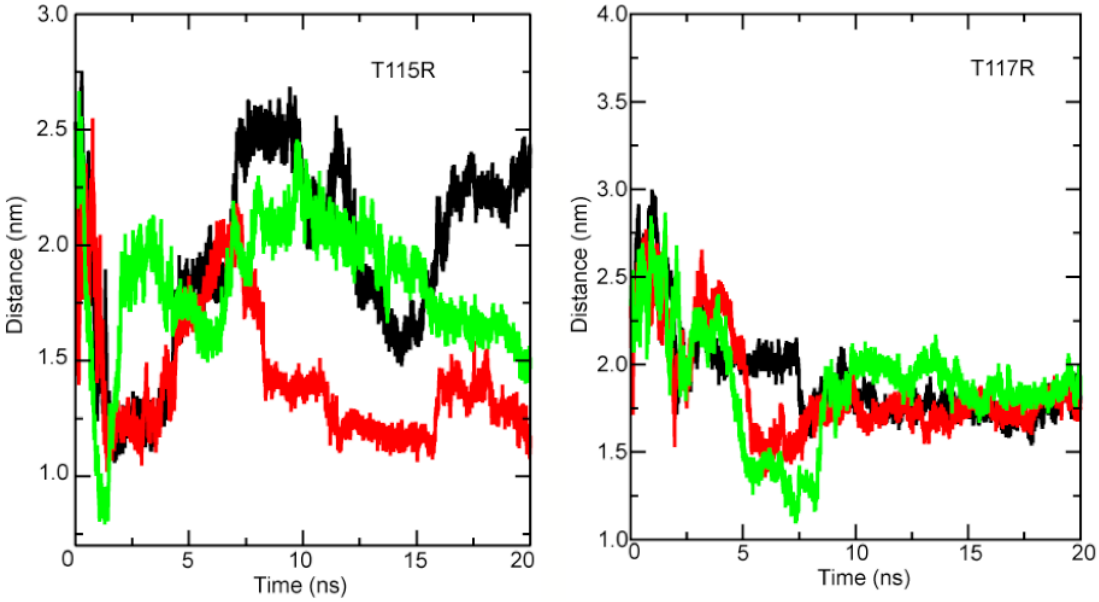

Figure 14. Distance between the centres of mass of the first and last residues of the ssDNA molecule relative to time for the T115R mutant (left) and the T117R mutant (right). Differently coloured curves represent repeat simulations.

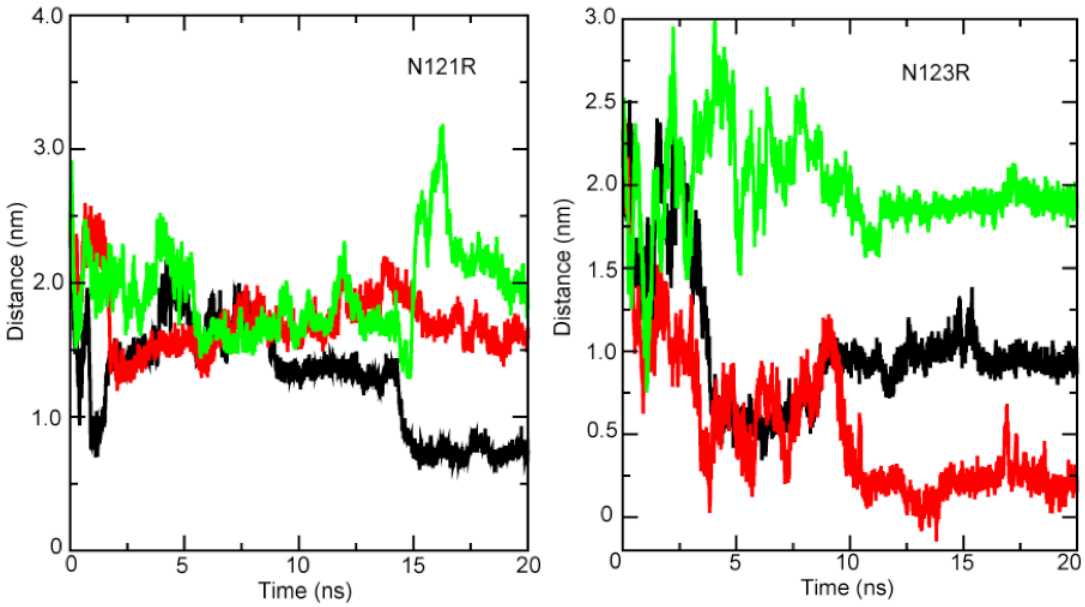

Figure 15. Distance between the centres of mass of the first and last residues of the ssDNA molecule relative to time for the N121R mutant (left) and the N123R mutant (right). Differently coloured curves represent repeat simulations.

Mutants that introduce an arginine residue into the pore tend to produce more folded ssDNAs, through the 'induced folding' mechanism displayed above in figure 12. In these mutants, the DNA forms multiple contacts with the arginine sidechains, maximising the interaction between the negatively charged phosphates of the ssDNA backbone and the positively charged guanidinium group of the ssDNA sidechain. Whether this interaction was particular to arginine was unclear, or whether introduction of the other positive residue, lysine, would have the same effect. Arginine is known to form a strong interaction with nucleic acid backbones, termed the 'arginine fork' (*159*), where the guanidinium sidechains can intersperse between backbone phosphates. The effect of

introducing lysine residues is discussed in a later chapter. These interactions, being long-lived, are in part an explanation for why arginine mutants have a longer translocation time. However, it is always noted that, while the ssDNA configurations may considerably change within the pore, the ssDNA always exits in the same orientation it enters. With short DNAs the first base to enter is always the first base to exit; the effect of longer DNAs is again discussed in a later chapter. A summary of the interactions experienced by an ssDNA translocating through an arginine mutant is given in figure 16 below.

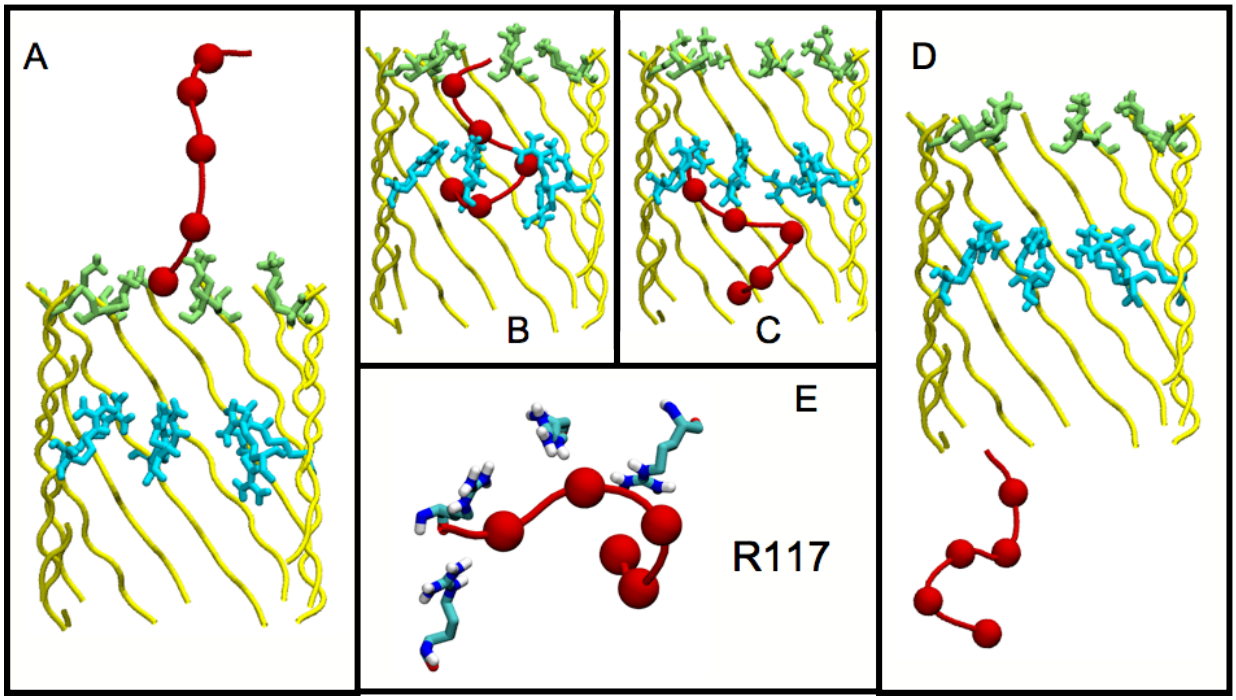

Figure 16. Interactions experienced by ssDNA translocating through the T117R mutant. Protein in yellow tube representation; R117 ring in cyan; K147 ring in green; ssDNA in red. (A) ssDNA in the starting configuration. (B) ssDNA has translocated to the point where the first bases are in contact with the R117 ring; last base remains tethered to K147 for some time. (C) Interaction has broken between K147 and ssDNA, allowing movement to a second tethering site, with terminal base tethered to R117. Once this interaction is broken, the ssDNA completes translocation (D). The interactions between ssDNA and R117 are shown in (E), with some phosphate-arginine interspersion shown to occur.

3.5 Discussion

Given the difficulty encountered in simulating the full α-HL protein, production of a simplified pore was a priority. The pores produced have been shown to recreate the trends observed in open pore currents particularly well, and to also preserve the general partial current trends observed in the presence of DNA, with consistent patterns observed across different groups of mutants. This recreation of currents on such a small system suggests that the rest of the protein is much less responsible for observed currents compared to the transmembrane beta barrel.

The translocation properties observed also mimic those of experiment, with non-translocating mutants being shown to not translocate *in silico,* and mutants that introduce positive charge providing extra interaction sites and generally slowing translocation. It has been shown experimentally that the introduction of extra rings of positive charges increases the translocation time in a linear fashion (*20*), and a general trend across all arginine mutants was a slower process of translocation (see, for example, the translocation pattern for N121R in figure 10, above). This translocation was shown to occur in a staggered fashion across all mutants, particularly due to the interaction between the DNA backbone and positively charged residues within the pore. These interactions, being relatively long-lasting, slow the translocation of DNA, providing an explanation for the observations in (*20*). The 'binding and sliding'-type mechanism of translocation has also been proposed for other forms of transport through other pores (*156, 157*).

The final observation of interest was the production of relatively stable conformations within the ssDNA, which could be described simply by means of comparing end-to-end distance. This simplicity of description is probably a function of these ssDNAs being relatively short. A longer ssDNA has much more conformational flexibility, described later. However, the adoption of these conformations – particularly in the context of interaction with positively charged residues – raises questions as to the behaviour of ssDNA within a nanopore, and the effects this may have on sequencing experiments. Although all of the small ssDNAs studied complete translocation in the same orientation to which they begin, some of the intermediates observed such as in figure 16 (B) may have an effect on the observed current readings in a nanopore sensing experiment.

Chapter 4 – Force field testing and validation

Abstract: The previous chapter has examined the behaviour of short ssDNAs within a nanopore environment. These short DNAs are limited in their conformational flexibility; however, longer strands are not. Parameter sets for DNA are generally parameterised and tested with double-stranded DNA, and as such any further work on longer ssDNAs would also require validation of the parameters used. This chapter focuses on the testing of commonly available parameter sets with regard to single-stranded DNA, both in solution and in contact with protein. The results tend to show two distinct patterns of behaviour between force field families, with varying levels of flexibility and tendencies to form certain structures.

4.1 Introduction

Core to most molecular mechanics methods are the force fields. A force field is both the atomic bonded and non-bonded parameters required for simulation, and the equations that govern these interactions. Many force fields exist for molecular dynamics simulatons, and with the exceptions of a few general force fields such as the Universal Force Field (*110*), they are tailored to specific tasks – for example, the AMBER force field family is suited to biological molecules (*126*), while the CLAYFF force field is designed for clay phases (*160*).

The diversity of available parameters presents something of a problem as the large number available for a single task can make the choice of suitable parameters difficult. This is compounded by the fact that many parameters remain untested, even for biomolecules for which they apparently are able to model.

A prime example of a system for which simulation paramters remain relatively untested is single-stranded DNA. While many force fields provide nucleic acid parameters, these parameters are intended for the simulation of double helix structures, and are used primarily for this purpose; furthermore, much of the parameterisation that occurs in order to create force fields for nucleic acids is based on work with double-stranded DNA (see, for example, the parameterisations reported in (*126, 161*)). Unbound single-stranded DNA in solution remains essentially untested, with only short, 1.5 ns simulations being reported in the literature (*162*) for atomistic DNA. Therefore, adequate testing of simulation parameters remains a priority.

The focus of this chapter is the testing and validation of nucleic acid force fields as applied to single-stranded DNA, both in solution and complexed with single-stranded binding proteins. While our previous work has focused only on short ssDNAs (*153*) with limited conformational flexibility, further studies within this thesis will focus on the use of longer strands of DNA. Since the possible limitations of the various available parameter sets have not been verified, it was deemed necessary to validate these parameters before proceeding with further studies by cross-comparison between the various available force fields.

4.2 Methods

4.2.1 ssDNA in solution

In order to test the response of a simulated ssDNA system to particular force fields, four of the more common and more recent nucleic acid force fields were chosen: the AMBER99 parameter set (*163*); the ParmBSC0 parameters, a revision of AMBER99 tailored to nucleic acids (*161*); CHARMM27 (*164*); and the then-most recent GROMOS96 parameter set, 53A6 (*155*).

The simulated system was comprised of one ssDNA strand extracted directly from the Dickerson dodecamer (*165*). This ssDNA strand thus began the simulation in the equivalent of the B-helical form from the crystal structure. The simulation box contained approximately 11,000 water molecules, with box dimensions of 6 x 6 x 6 nm. The water models used were force field dependent, and matched to those used in the original parameterisations of each. The AMBER99 and ParmBSC0 simulations used TIP3P, CHARMM27 used TIPS3P (the GROMACS implementation of which includes Lennard-Jones parameters for water hydrogens), and the SPC model was used for GROMOS96 53A6. Salt was added first to neutralise system charge, then to concentrations of 0, 0.1, 0.2, 0.5 and 1 M, which gave a total of approximately 200 Na+ and Cl- ions. Simulations were performed using GROMACS version 4.5.3 (*123*). The extended ssDNA structures used in some simulations were created using the GROMACS pull code. A summary of all simulations is given below in table 1.

Use of the ParmBSC0 parameters required conversion from AMBER to GROMACS format; the resulting parameters were tested through simulations of the Dickerson dodecamer *in vacuo* with infinite cutoffs in both AMBER11 (*166*) and GROMACS 4.5.3 (*123*). The differences in the energies for the dihedrals were within 1.4×10^{-3} kJ mol^{-1} (0.0001%).

Further conversion of GROMOS parameters was required, as the GROMACS implementation of the GROMOS96 force field does not by default include terminal nucleotide definitions for DNA. The missing 3' hydroxyl parameters were created using the standard GROMOS RNA parameters as a reference. Terminal nucleotide definitions for the GROMACs implementation of GROMOS96 have been reported previously (*167*), but are inconsistent with the standard GROMOS96 parameters.

Table 1. Summary of simulations by forcefield, number, salt concentration and length.

Forcefield	No. simulations	Salt concentrations (M)	Simulation lengths (ns)
CHARMM27	15	0, 0.1, 0.2, 0.5, 1	100
AMBER99	15	0, 0.1, 0.2, 0.5, 1	100
AMBER99 with ParmBSC0 torsions	15	0, 0.1, 0.2, 0.5, 1	100
GROMOS96 53A6	15	0, 0.1, 0.2, 0.5, 1	100
CHARMM27 (extended ssDNA)	3	1	100
AMBER99 (extended ssDNA)	3	1	100
AMBER99 with ParmBSC0 torsions (extended ssDNA)	3	1	100
GROMOS96 53A6 (extended ssDNA)	3	1	100

Each system was simulated using the methods described in the original literature for each force field, with the exception of the 53A6 force field. This force field was originally parameterised for use with reaction-field electrostatics, but has at least in some instances been found to perform better in conjunction with particle mesh Ewald electrostatics (*168*). All simulations were run for 100 ns, with a timestep of 2 fs and a total of 50,000,000 steps, and repeated 3 times for each force field at each salt concentration. Each system was also controlled at a temperature of 300 K and a pressure of 1 bar using a Berendsen thermo- and barostat respectively. Individual methods details are given below:

AMBER99 and ParmBSC0: van der Waals interactions were truncated with a cutoff at 1 nm. PME electrostatics used a short range cutoff of 1 nm, with fourth order interpolation.

CHARMM27: van der Waals interactions were treated using a switching function. The switch was applied at 1 nm, with the potential decaying to zero at 1.2 nm. PME electrostatics used a short range cutoff of 1.2 nm and sixth order interpolation.

GROMOS96 53A6: van der Waals interactions were truncated with a cutoff at 1.2 nm. PME electrostatics used a short range cutoff at 1.2 nm, with fourth order interpolation.

4.2.2 ssDNA-protein complex

The first challenge encountered was finding a suitable single-stranded binding protein for simulation. As a general rule, crystal structures of proteins complexed with ssDNA do not resolve well due to the inherent flexibility of the DNA strand (*169*), and many which are crystallized are crystallized in the presence of a model construct DNA, such as polyC or polyT. As such, the final structure will often consist of only a few bases of a non-standard DNA.

The model chosen was the *Plasmodium falciparum* ssDNA binding protein tetramer. Single-stranded DNA binding proteins are ubiquitous in biology and are responsible for covering exposed single-stranded DNA, which may form as the result of replication, repair or other processes (*170*). This prevents the double helix from reforming while the DNA is in use, and also protects it from degradation. *Plasmodium falciparum* itself is the parasite carried by mosquitos that is responsible for malaria. Its binding protein is present in the apicoplast of the cell, an organelle unique to the Apicomplexa phylum of eukaryotic organisms. This organelle, like chloroplasts and mitochondria, is thought to be of bacterial origin and contains its own genome, and the binding protein is thought to be involved in its maintenance (*171*). The apicoplast is essential to the survival of the host organism, and is therefore an attractive drug target; however, its function is uncertain beyond being involved in metabolic pathways such as fatty acid, isoprenoid, haeme or iron-sulphur cluster synthesis. While the crystal structure (PDB ID: 3ULP, (*169*); see figure 1 below) consists of polyT ssDNA, the length of the individual pieces is between ten and thirteen bases long. It was felt that longer pieces of non-standard DNA would perform better from a simulation perspective than short fragments of heterogeneous DNA. Total system size was 50,000 atoms, in a 8 * 8 * 8 nm box.

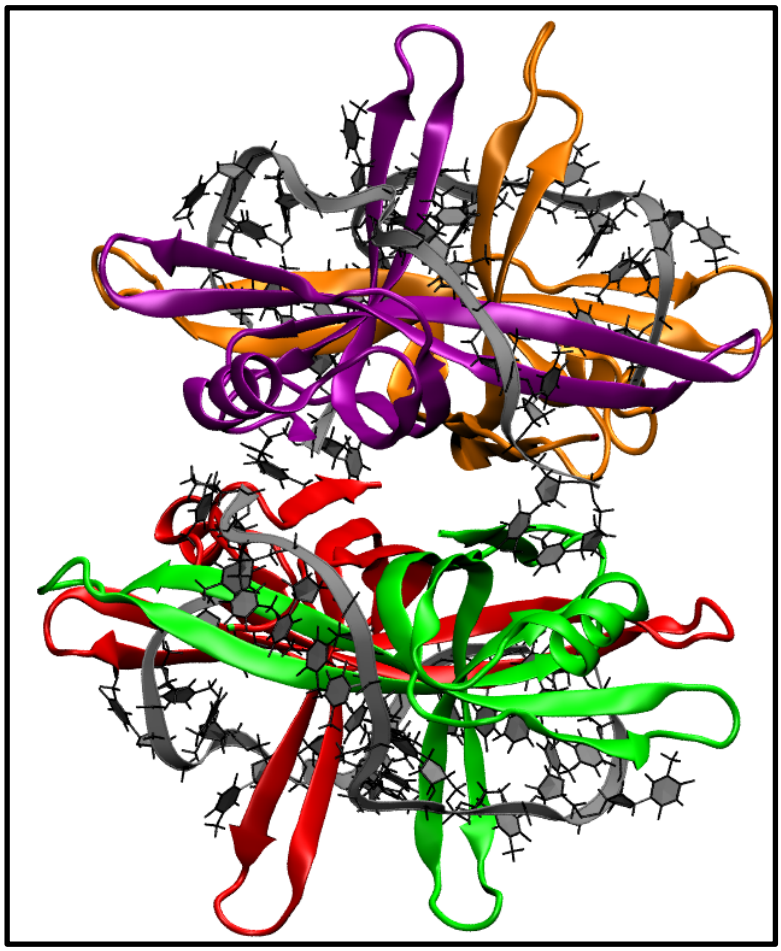

Figure 1. PDB structure 3ULP, the *plasmodium falciparum* SSBP tetramer. Protein chains are labelled in individual colours – green, red, purple and orange. DNA backbone is highlighted in silver; bonds and atoms are in black.

Time constraints did not allow for as comprehensive a set of simulations as with the solution ssDNAs, and as such three force fields were tested – AMBER99, GROMO53a6 and ParmBSC0 – at a single salt concentration. CHARMM27 was ignored due to the excessive time required for simulation. This longer time is a consequence of the CHARMM water model as implemented in the GROMACS software package: unlike other water models, the CHARMM TIPS3P model includes Lennard-Jones parameters on its hydrogen atoms.

Simulations were performed using the same parameters as ssDNA in solution (see section 4.1.1). Some difficulty was encountered in using the AMBER force fields due to excessive salt crystallization encountered at the 1M level. This is a known issue with the AMBER force field family (*172*). This is likely compounded by the common method of adding salt in the GROMACS package: the genion script replaces ions based on simulation volume, not number of water molecules, and with a large protein such as 3ULP the effective salt concentration is exaggerated; this was avoided by calculating the concentrations manually, and thus reducing the number of sodium/chloride ions present to 260 from 330. This force field issue was found not to be a

problem in the higher concentration simulations of section 4.1.1, likely due to bulk solvent being a proportionally greater component of box volume.

4.3 Results

4.3.1 ssDNA in solution: conformational behaviour

Two distinct patterns of behaviour were noted in all ssDNA simulations, split between the AMBER-type force fields and the non-AMBER force fields; this can be shown in a simple manner by measuring the end-to-end distance – ie, the distance between 3' and 5' hydroxyl groups – of the ssDNA throughout each simulation. Data from representative simulations are given in figure 2 below. The data is discussed in more detail below; however, generally speaking the AMBER-type force fields show a much greater retention of the initial helical structure of the ssDNA, leading to two distinct behaviour patters between the AMBER and non-AMBER force fields. Validation of the 'correctness' of either behaviour is difficult due to the relative scarcity of experimental data for ssDNA, but contrasting observations can be made.

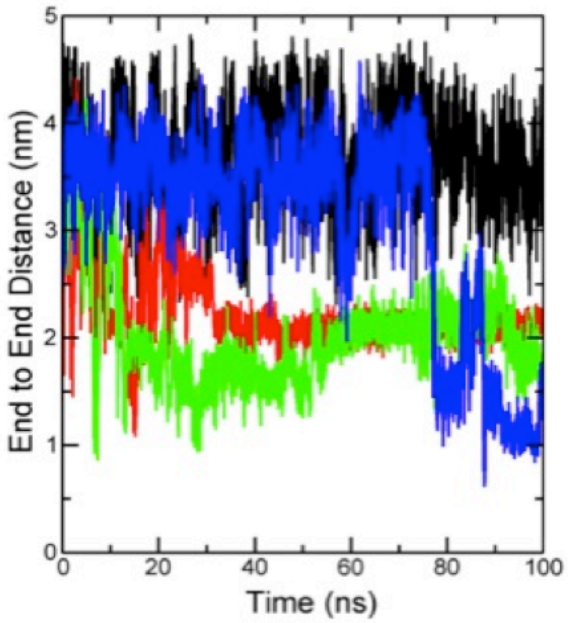

Figure 2. ssDNA end-to-end distance through representative simulations from each force field, at 0.2 M NaCl salt concentration. ParmBSC0 in black, AMBER99 in blue, CHARMM27 in red, GROMOS G53A6 in green. AMBER force fields generally show greater stability of initial structure.

4.3.1.1 CHARMM and GROMOS

In both CHARMM27 and 53A6, the ssDNA tends to show much greater flexibility than in the AMBER force fields, and quickly loses any of the original B-helical structure it begins with. Both

force fields show the formation of a compact, globular structure within the first 10-20 ns of simulation. This structure appears to be random, showing some level of base stacking between adjacent bases and non-Watson-Crick base-pairing. Hydrogen bonding patterns show interactions both between the base components themselves, and between the bases and the ssDNA backbone phosphates, with the latter being the more prevalent of the two. These interactions are transient, lasting less than 10 ns; as such, the complex does not appear to show any long-term structure beyond forming a globule. Radius of gyration calculations were performed across all force fields at all salt concentrations, and are given in appendix 2. The radius of gyration for the initial configuration in all force fields was ~1.3 nm, and for the compact structure formed after simulation using CHARMM27 and 53A6 was ~0.9 nm.

4.3.1.2 AMBER99 and ParmBSC0

In both AMBER99 and ParmBSC0, the initial conformation of the ssDNA can be retained for substantial lengths of time. In AMBER99, this can be for up to 80 ns, and this retention occurs for at least 40 ns in 33% of simulations performed. In the other 66% of simulations this configuration is lost within 30 ns. ParmBSC0 retains this initial helical structure in some cases for the whole 100 ns of simulation time, and shows retention for more than 40 ns in 80 % of simulations performed. Clearly the AMBER-type force fields show some degree of favouring a helical-type structure, further demonstrated by the formation in 40% of all simulations of a quasi-double helix, shown in figure 3 below.

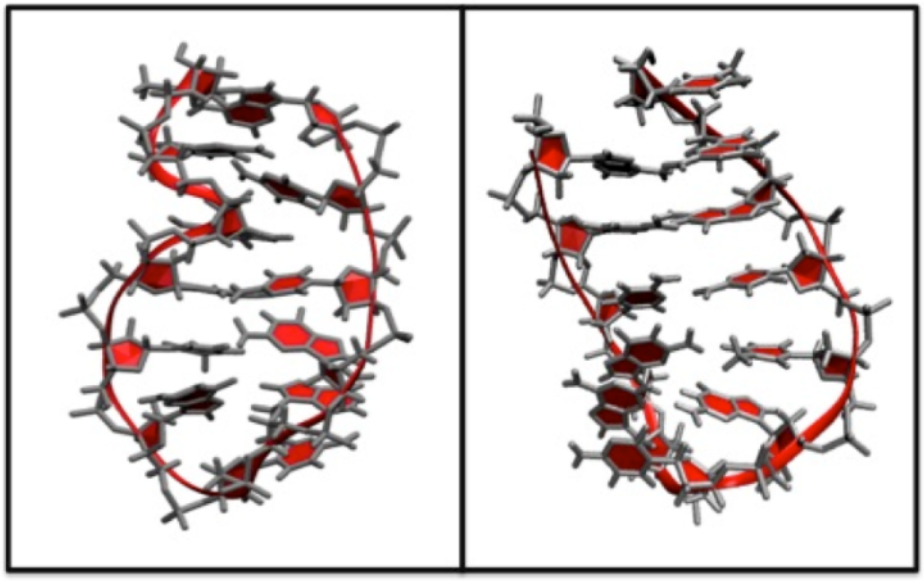

Figure 3. Quasi-helical structures formed in AMBER99 (left) and ParmBSC0 (right).

This structure appears to be stabilised by hydrogen bonding between the bases of the strand. This bonding is non-Watson-Crick, with any base being able to pair with any other. These hydrogen

bonds are relatively stable, and last between 30 and 40 ns. The structure, once formed, does not unravel within the simulation times used, and did not appear to show any dependence on salt concentrations. The observed radii of gyration for the AMBER-type force fields were on average 0.95 Å in the helical-type configuration. Representative ending configurations from all force fields are given in figure 4 below, illustrating the general differences between the two types of behaviour observed.

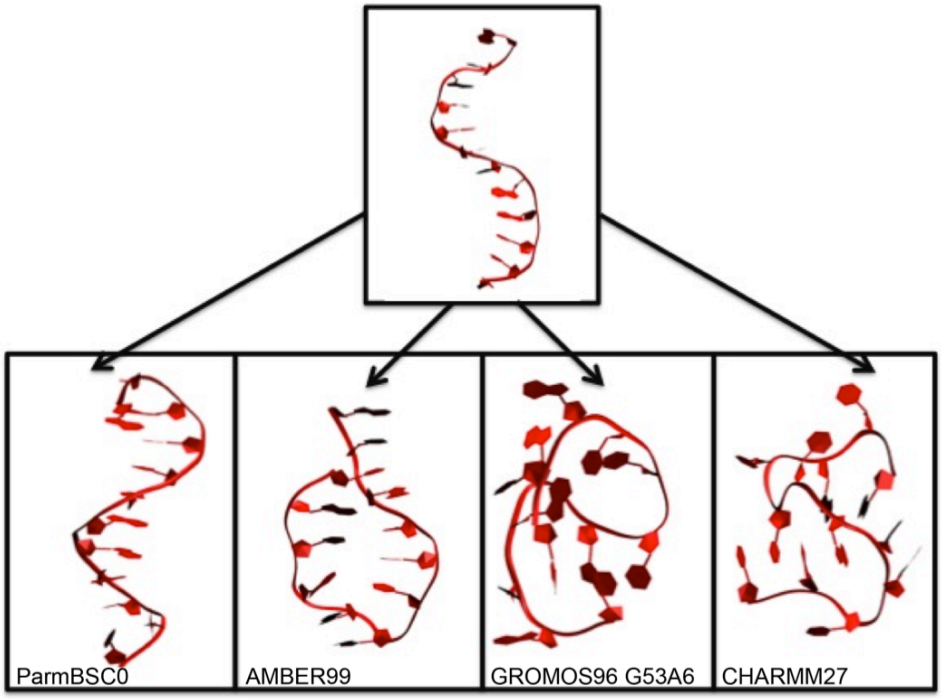

Figure 4. Representative structures of starting and ending configurations of ssDNA. Particularly notable is the AMBER force field family's ability to retain initial structure or form quasi-helical structures.

Also considered was whether these behaviours were simply a result of the starting ssDNA structure, and whether starting from a non-helical piece of ssDNA would show the same patterns of behaviour. Three simulations were performed for each force field using an ssDNA that had been extended using the GROMACS pull code, such that its structure was no longer B-helical. The CHARMM27 and 53A6 force fields showed much the same patterns of behaviour, forming globular structures relatively quickly. The AMBER-type force fields no longer show the same retention of initial conformation; however, they still show formation of helical-type structures by the end of simulation in 50% of cases.

4.3.2 Base-base parameter analysis

Comparison of backbone phosphate-phosphate distances with experimental data was used to determine whether the ssDNA configurations formed were within observed parameters.

Experimental measurements show that, while the phosphate-phosphate distance can fluctuate, this distance is usually considered to be between 5.9 and 7 Å, as calculated directly from multiple crystal structures of DNA (*173*). Measurements from simulation are given below in figure 5.

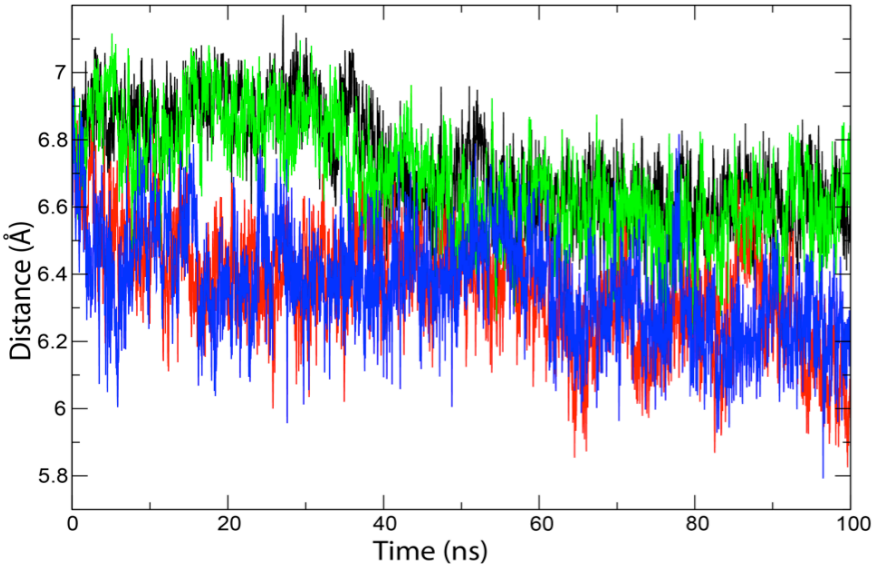

Figure 5. Average phosphate-phosphate distances across all force fields at 0.2 M salt concentration. ParmBSC0 in black, AMBER99 in green, CHARMM27 in red, 53A6 in blue.

Again, broadly speaking, the results cluster into two patterns of behaviour, with the AMBER-type force fields on average having a slightly higher phosphate-phosphate distance. Average distances for CHARMM27 were 6.41 Å with a standard deviation of 0.42 Å; this was similar to the values for 53A6 which were 6.40 Å with a standard deviation of 0.51 Å. Both AMBER99 and ParmBSC0 show an average phosphate-phosphate distance of 6.76 Å, with standard deviations of 0.4 Å and 0.36 Å respectively. The lower standard deviations perhaps reflect the lower flexibility observed in these force fields. All observed values across all force fields are within experimental parameters, but the AMBER-type force fields are closer to the 7 Å expected in the canonical double helix.

Further calculations were performed using the 3DNA web interface (*174, 175*) in order to determine base-step parameters, which provide another potential comparison with an experimental observable. The data from the calculations on the AMBER99 simulations are given in table 2 below, and are in good agreement with the observed values for a double helix, again highlighting the propensity for AMBER force fields to maintain this configuration. Such calculations were not possible for the GROMOS and CHARMM force fields due to the disordered nature of the globular structures formed.

Base step	Dickerson (0 ns)	20 ns	40 ns	60 ns	80 ns	100 ns
C-G	3.62					
G-C	3.46	2.93	3.1		3.16	
C-G	2.85	3.17	3.06	3.44	3.48	3.38
G-A	3.38	2.84	3.49	3.09	3.67	2.74
A-A	3.36	3	3.52	3.42	2.99	3.81
A-T	3.22	3.27	3.29	3.42	3.5	3.54
T-T	3.2	3.28	3.84	2.83	4.01	3.05
T-C	3.37	3.82	3.57	3.31	3.32	3
C-G	3.4	3.91	3.32	3.82	3.54	3.6
G-C	3.96	3.9	3.12	3.1	3.17	4.08
C-G	3.02					
Avg.	3.35	3.35	3.37	3.30	3.43	3.4

Table 2 Base-step parameters in AMBER99 compared to experimental values. In general, these remain close to the values from the experimental structure (column 2), which are calculated from the PDB structure 1BNA (164). Blank cells indicate either outlying values or values that could not be calculated due to unusual configurations.

Given this general propensity for AMBER-type force fields to retain their initial helical configurations, and also to often form a similar structure even after losing the initial configuration, it may be that such parameter sets are unsuitable for simulations of ssDNA in solution. While both CHARMM27 and 53A6 show similar behaviour, 53A6 was chosen for further simulation work. The GROMACS CHARMM27 implementation runs much slower than 53A6 due to the presence of Lennard-Jones parameters for water hydrogens (*176*), and as such 53A6 was used due to its relative speed.

4.3.3 ssDNA in contact with protein

The ssDNA-protein complex remains stable throughout the length of the simulation across all force fields, and does not separate over the course of 100 ns. However, some minor differences were observed across force fields. The initial system and final systems for each force field are shown in figure 6 below. The GROMOS force field shows the largest deviation from the initial structure, with both protein and nucleic acids having noticeable changes. The ends of the DNA across all force fields have some freedom of movement, but again GROMOS DNA seems to show greater movement compared to both AMBER-type force fields.

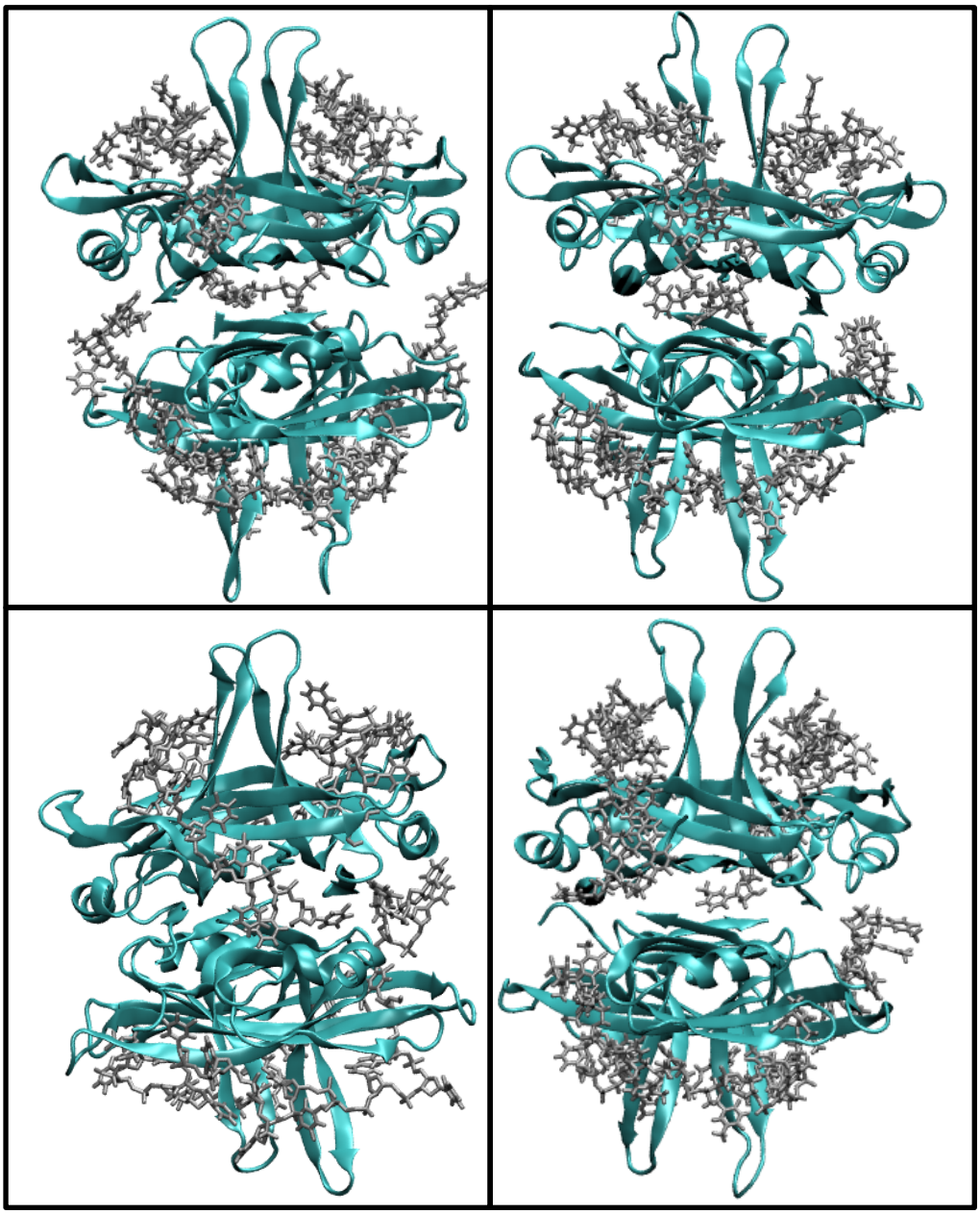

Figure 6. Initial (top left) and final conformations of *plasmodium falciparum* ssDNA binding protein across all force fields. Top right: AMBER99, bottom left: GROMOS96, bottom right: ParmBSC0. Complex remains stable in all force fields, but the minor helices tend to show small disruptions.

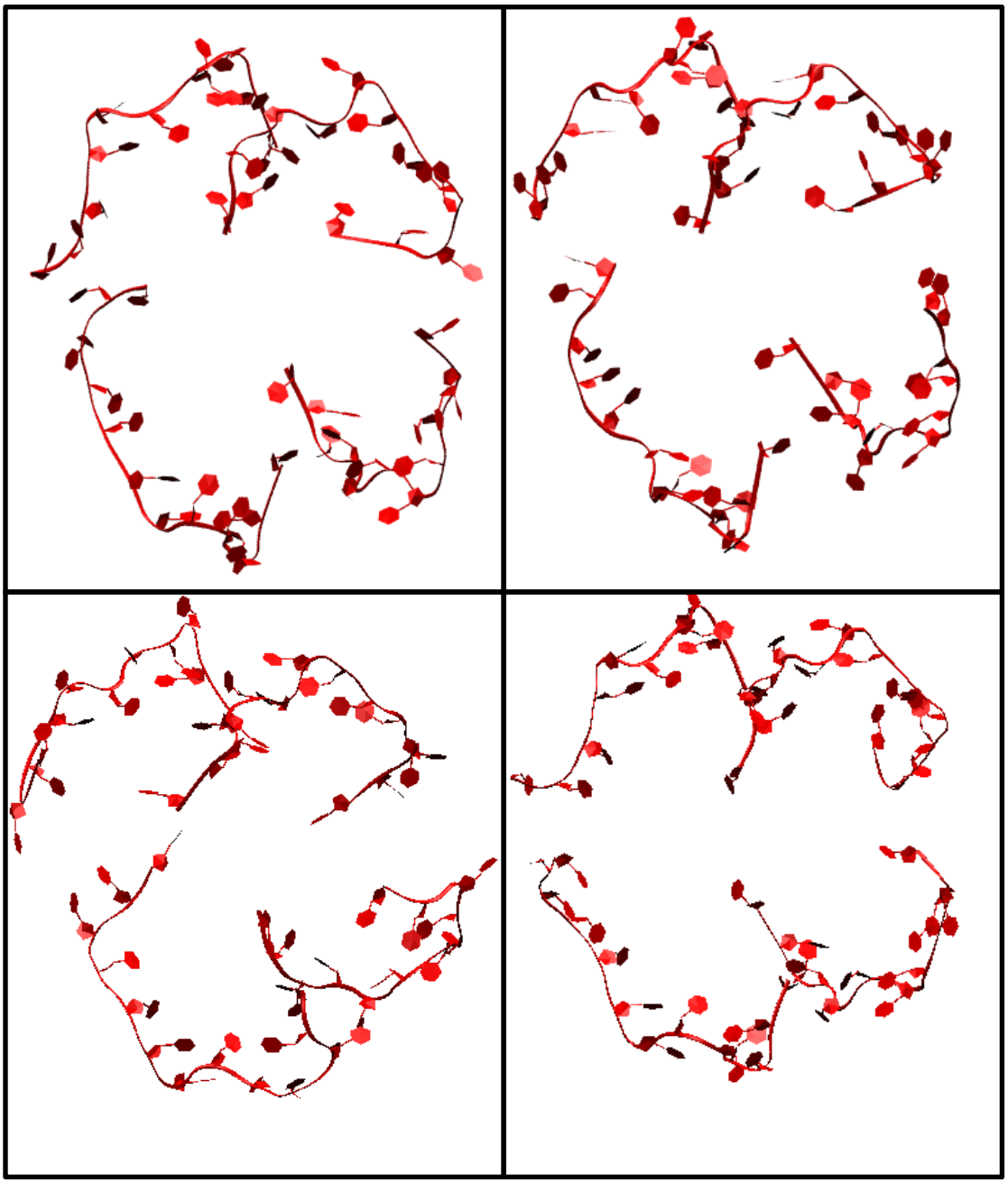

Figure 7. Initial (top left) and final configurations for nucleic acid simulated using AMBER99 (top right), GROMOS 53A6 (bottom left) and ParmBSC0 (bottom right). While the general configuration is conserved across all forcefields, some regions of the GROMOS DNA show a greater degree of movement.

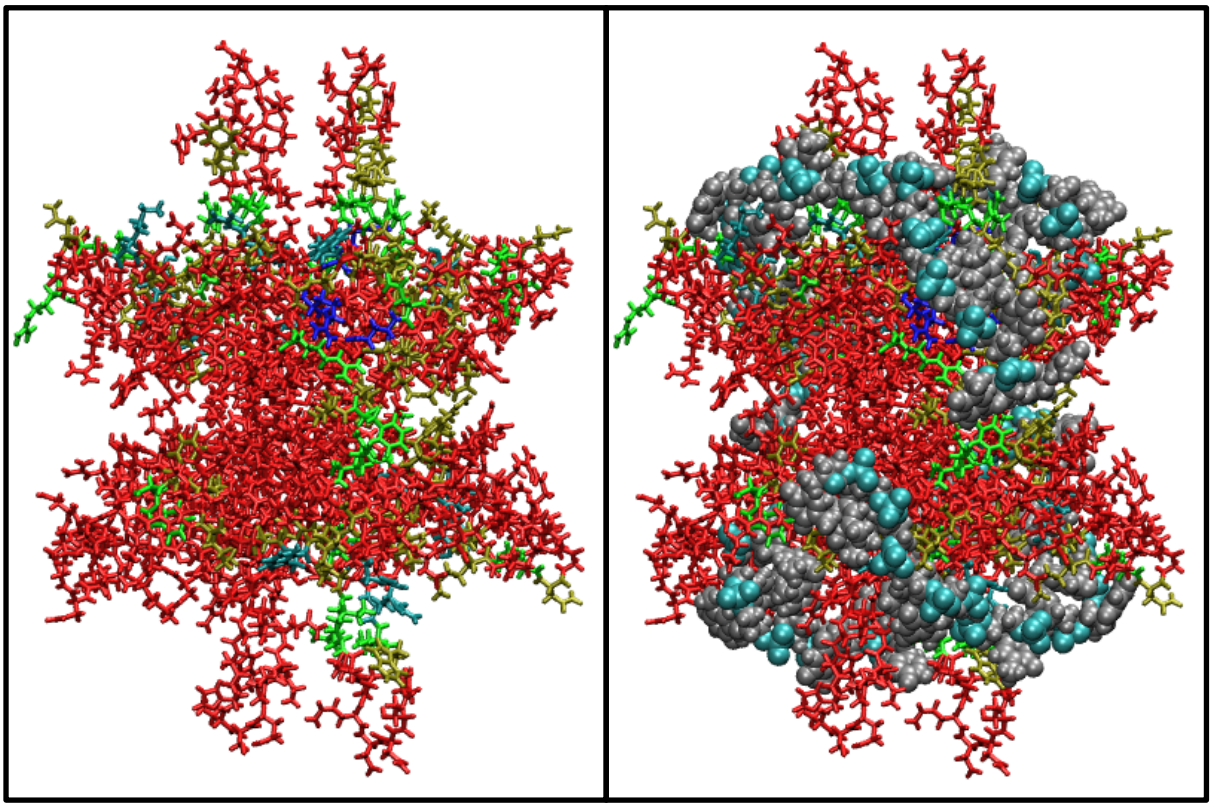

Figure 8. Initial frame contact map between protein and DNA (AMBER99 force field). Left: DNA not visible; right: DNA visible as silver van der Waals spheres, with backbone phosphates highlighted in cyan. Contact colourings for the protein are as an RGB scale. Red represents the least number of contacts, blue the most.

The initial contact map between protein and DNA is given in figure 8 both with and without the protein component displayed. As might be expected from a single-stranded binding protein, many of the contacts are with the positive amino acids arginine and lysine.

The stability of the individual components was examined across all force fields using root mean square deviation (RMSD) calculations, where the deviation of the structure of interest from the starting structure across a trajectory is calculated by least squares fitting to the initial structure. The results for the protein component are given in figure 9 below.

Proteins simulated in AMBER99 and ParmBSC0 show similar patterns of behavior, with an average RMSD of approximately 0.25 nm. This is expected, as the protein component of each force field is identical. The GROMOS protein has a greater average RMSD across all simulations, at approximately 0.36 nm, suggesting slightly greater flexibility of the protein component relative to the AMBER force fields. However, a value of 0.3 nm is considered to indicate some degree of similarity across structures and therefore an acceptable level of movement (*177*).

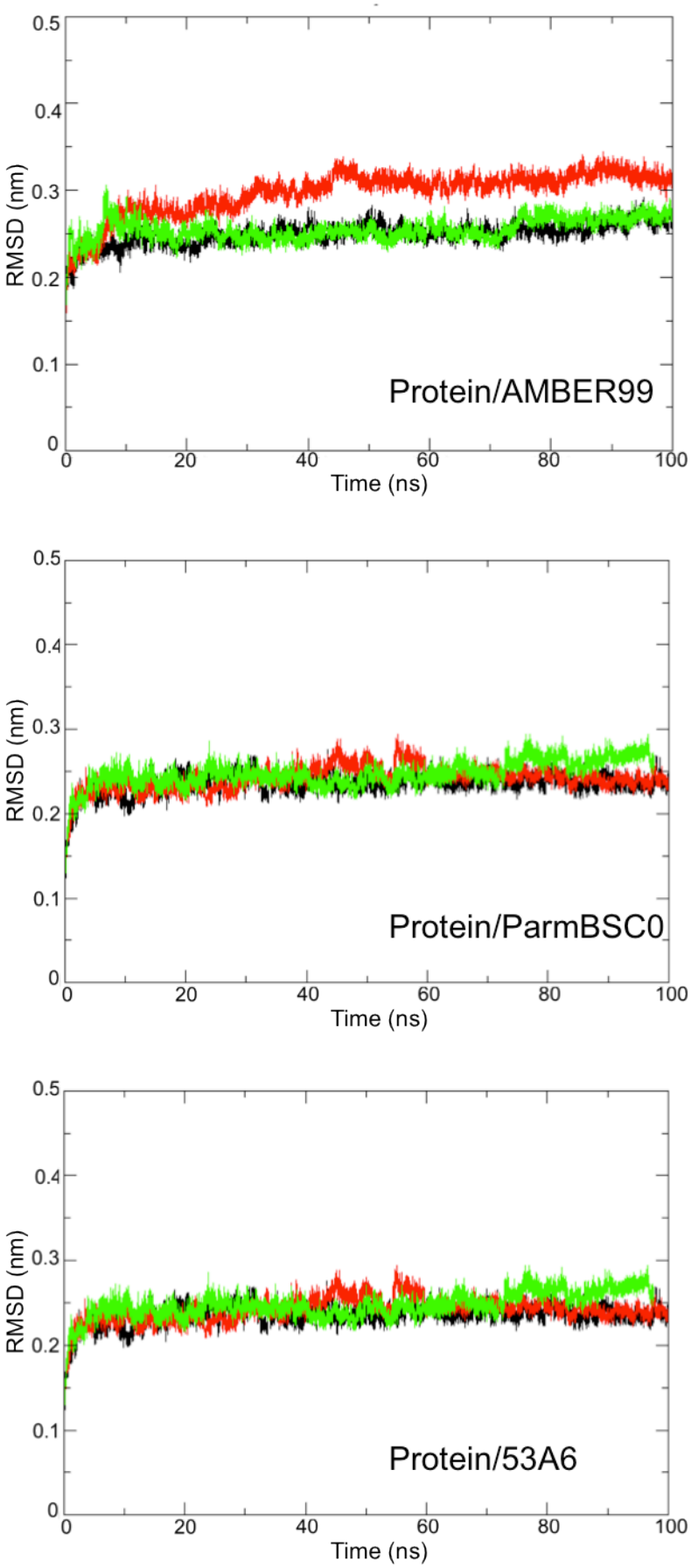

Figure 9, above: RMSDs in the protein component of each simulation. Top, AMBER99; middle; ParmBSC0; bottom, GROMOS96. Different colours represent different trajectories.

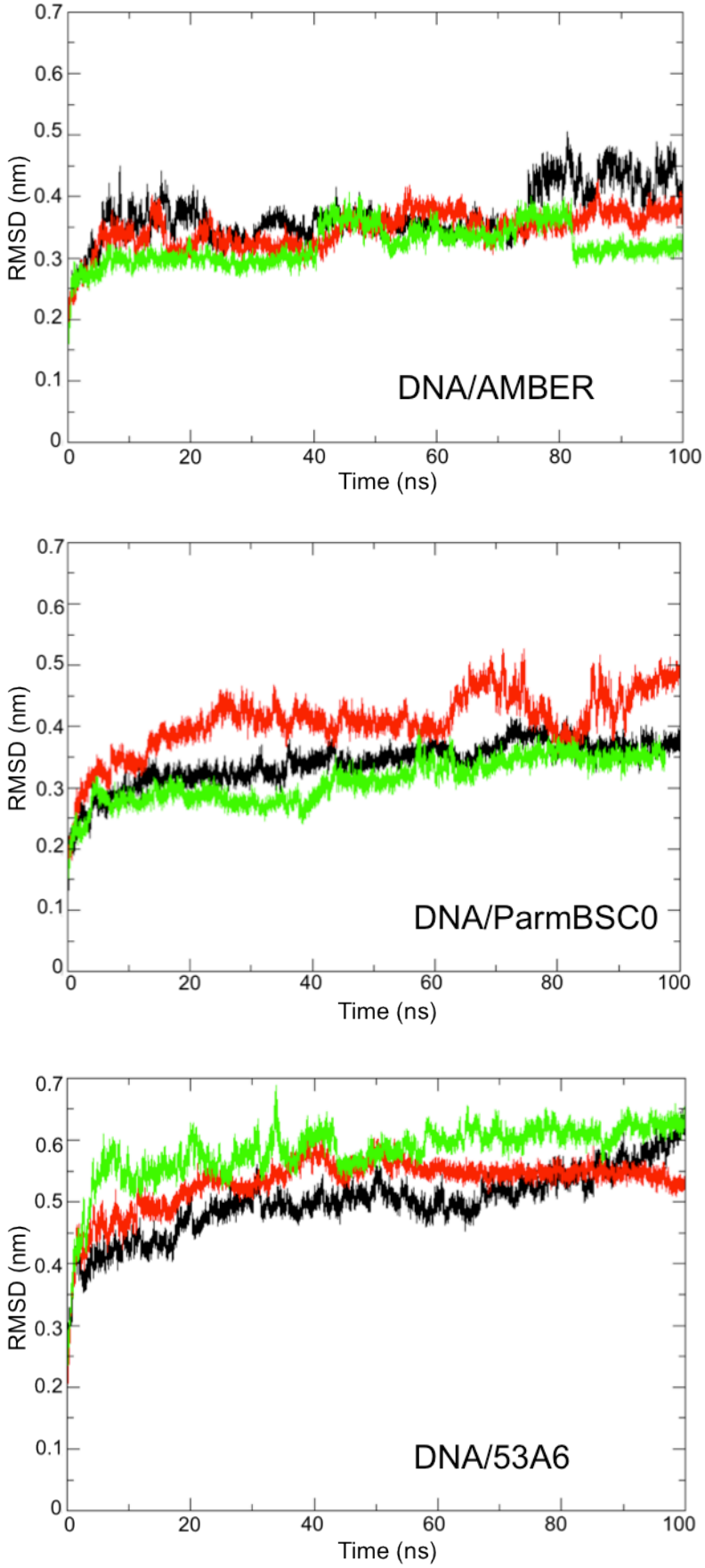

Figure 10, above: RMSDs in the DNA component of each simulation. Top, AMBER99; middle; ParmBSC0; bottom, GROMOS96. Different colours represent different trajectories

The flexibility of ssDNA in contact with protein was also examined through RMSD calculations, shown in figure 10, above. Again, the AMBER force field family show less deviation relative to starting structure compared to the GROMOS force field, with the AMBER force fields having an average RMSD of 0.35 nm by 100 ns of simulation time compared to an average of 0.55 nm for GROMOS DNA. As noted previously, the backbone torsions of the AMBER force field family can exaggerate the conformational rigidity of the DNA. Both force field families appear to suffer from the ends of DNA strands having more freedom to move and thus forming the basis for some of this fluctuation as they move away from the protein; however, this appears to be slightly more prevalent in GROMOS DNA, and the value of 0.55 nm suggests that the final structure will have in part deviated from the initial structure, suggesting some degree of instability.

Root mean square fluctuation calculations can give a residue-by-residue breakdown of the standard deviation of atomic positions relative to starting structure. Representative RMSFs for each force field are given below in figure 11. The GROMOS 53A6 force field shows the greatest fluctuation in a single amino acid, M229 – which is an n-terminal residue of the complex – as well as peaks that are on average higher than in AMBER99 or ParmBSC0. Calculation of the number of contacts between protein and DNA shows a large number of contacts between the protein and the DNA as a whole, with no specific region dominating this interaction. A contact is defined using a distance cutoff of 3.5 Å between the two groups. As such, defining whether a peak is the result of a DNA interaction is difficult: both peaks and troughs can show a large number of DNA-protein contacts; some residues at large peaks, such as at x=140 on all graphs, are not in contact with the DNA at all. Based on structural and contact data, it appears that peaks arise from terminal residues, some residues in contact with DNA, and in some cases from the more flexible loop regions that extend from the protein, for example around residue N253.

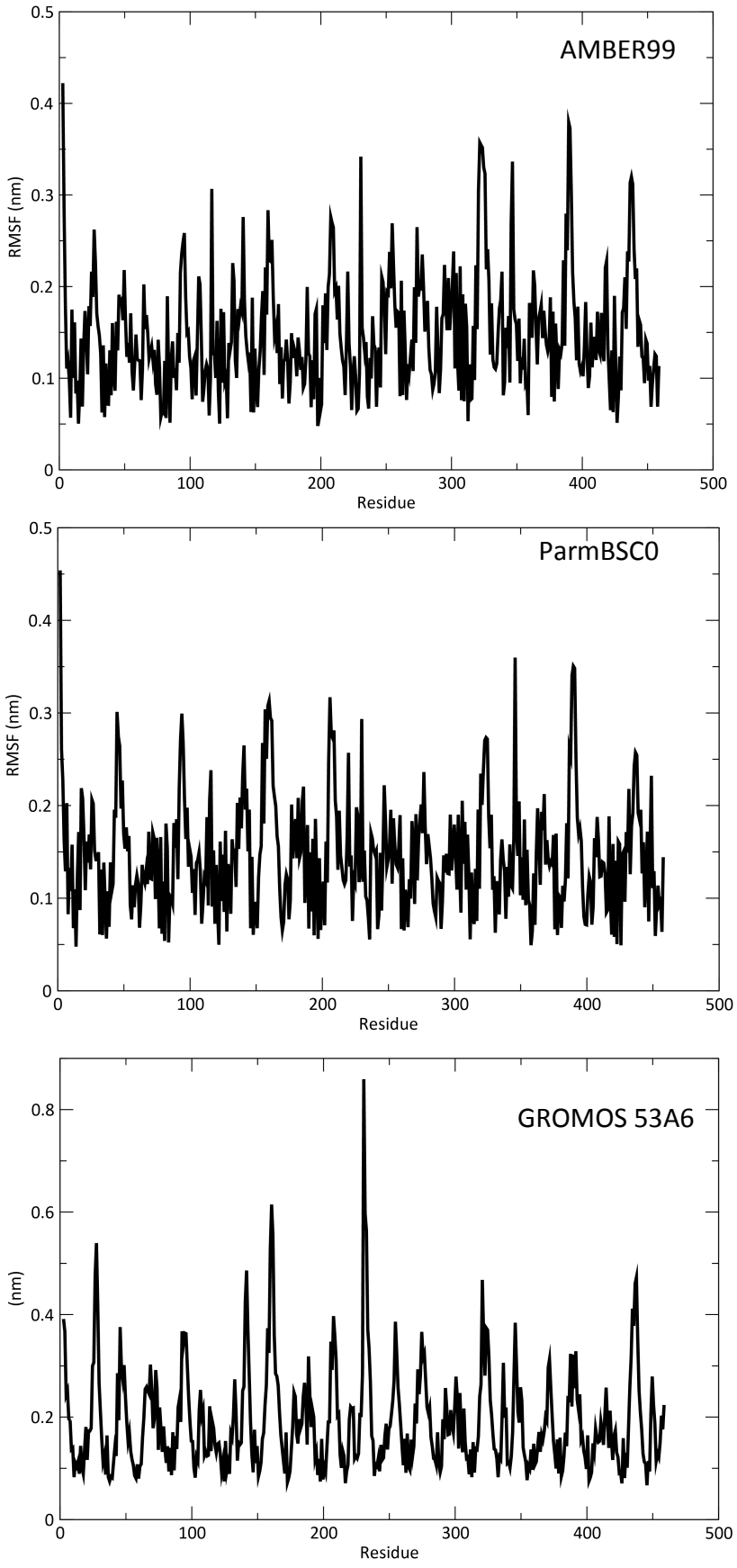

Figure 11. RMSF values for residues in the single-stranded binding protein, simulated in AMBER99 (top), ParmBSC0 (middle) and GROMOS 53A6 (bottom).

The evolution of the secondary structure of the protein was traced through time using the DSSP program (*178, 179*). The results across all trajectories are given in appendix 1. The AMBER99 and ParmBSC0 force fields show greater stability in the beta-sheet regions that form the core of the protein. While the GROMOS-simulated protein does remain generally stable, there appears to be more of a tendency particularly for the ends of the sheet regions to lose some definition as the simulation progresses. The GROMOS plots also are generally 'noisier', with minor drifts in structure that correct quickly. Across all force fields there can be some loss of secondary structure in the helical regions of the protein between residues 65-75, 178-188, 293-303 and 408-418, although the GROMOS force field appears again to suffer from this more. The GROMOS 53A6 force field, and the GROMOS96 force field family in general, are known to favour β-sheets, particularly with particle mesh Ewald electrostatics, or at least to disfavour α-helices (*180, 181*).

The number of protein-ssDNA hydrogen bonds was also calculated, using a distance criterion of 0.35 nm and an angle cut-off of 30 degrees. A representative result from each force field is given below in figure 12:

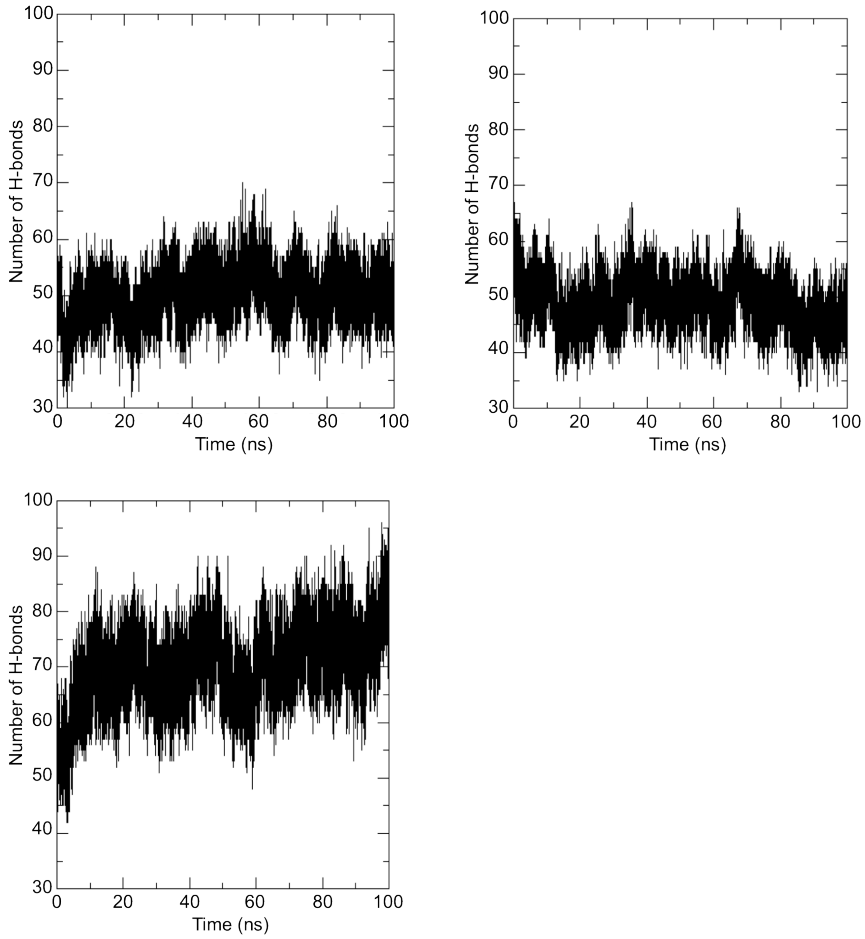

Figure 12. Number of hydrogen bonds observed using a distance cut-off of 0.35 nm and an angle cutoff of 30 degrees. Top left: AMBER99. Top right: ParmBSC0. Bottom left: GROMOS96.

Both the AMBER and ParmBSC0 force fields show a similar hydrogen bonding pattern, with approximately 50 hydrogen bonds being formed between ssDNA and protein at any given timestep. This is in contrast to the GROMOS force field, where approximately 70 hydrogen bonds are formed at any given timestep. This is likely due to the greater flexibility of GROMOS DNA allowing a greater number of bonds to be formed at any point; on average, there is also a greater standard deviation in the number of bonds in the GROMOS force field compared to the AMBER force fields (7.1 for G53A6 compared to 4.6 for AMBER99). In general, it appears that the whole complex has greater freedom of movement when using the GROMOS force field compared to the AMBER force fields.

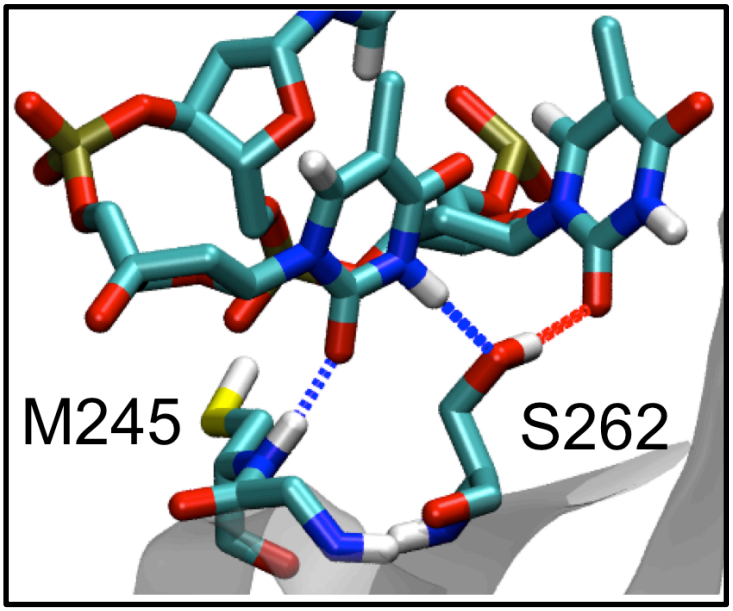

Figure 13. Typical base hydrogen bonding patterns in protein-DNA simulations, in the GROMOS force field. Hydrogen bonds can be formed between the bases and either the backbone (as in M245, left) or between sidechains and bases (S262, right).

Hydrogen bonding between the protein and DNA occurs mostly between the protein sidechains and the DNA, for a total of 75% of hydrogen bonding across all force fields; protein backbone-DNA interactions comprise about 25% of the total. Figure 13 above shows the typical types of hydrogen bond: backbone-DNA and sidechain-DNA, with both being donors and acceptors.

DNA phosphate-protein interactions also form a substantial part of the total hydrogen bonding contribution, approximately 35% of the total. Like most DNA binding proteins, the *plasmodium falciparum* SSBP has a high number of the positive amino acids arginine and lysine, with 12 lysine and 8 arginine residues per subunit. As the complex is a tetramer, this is a total of 80 positively charged residues in the whole complex, out of a total of 500 residues. Over two thirds of the DNA phosphate-protein hydrogen bonds are with the lysine and arginine residues; it can be assumed

that there is a substantial electrostatic contribution to this interaction as well as the hydrogen bonding itself.

4.4 Conclusions

While the AMBER force field family is generally regarded as superior in the context of nucleic acid simulations, the results as applied to single-stranded DNA tend to suggest that the limited conformational flexibility of the backbone compared to other force fields should be a point of concern.

The above results show that single-stranded DNA in solution simulated using the AMBER force fields has a preference for a helical or quasi-helical structure, which suggests that the parameterization of the AMBER favours this helical conformation, likely as a result of their being parameterized against double helical configurations. The GROMOS and CHARMM force field families avoid this, showing greater conformational flexibility. It should be noted that it may not be possible to state which is technically 'correct', but it appears that this tendency may be a force field artifact.

The pattern of AMBER force field rigidity versus GROMOS flexibility is continued in the protein-DNA studies. GROMOS DNA shows a greater conformational flexibility than AMBER; this is also reflected in the tendency to form a greater number of hydrogen bonds due to the ability of the DNA to shift and form new bonding configurations. The AMBER force fields show a lower number of hydrogen bonds and a lesser flexibility. However, once again, it is difficult to say which, exactly, is 'correct', but the fact that there are two patterns of behavior perhaps warrants further investigation.

For future work using ssDNA, it was decided to use the GROMOS force field due to the lack of quasi-helical structure formation, as well as the faster simulation time compared to the somewhat more impractical CHARMM force fields

Chapter 5: Conformational analysis of ssDNA within the α-hemolysin pore

Abstract: It has been demonstrated that small DNAs can adopt a limited, easily characterisable set of conformations within a nanopore. However, these small DNAs are not representative of the lengths of DNA that can be used in a typical sequencing experiment. Furthermore, a longer DNA will be able to adopt a much broader range of conformations within a nanopore, and as such the methods used previously to characterise these conformations may no longer be successful. The conformations of ssDNA within the wildtype and several mutant nanopores are characterised using cluster analysis, and the potential effects of the resulting conformations on translocation are considered. Of particular note was the fact that ssDNA may exit the nanopores in a slightly different order to which it entered, which may have an effect on strand readings and therefore may be of consequence to sequencing experiments.

5.1 Introduction

The ability of DNA to form multiple conformations within a pore, demonstrated in chapter 3, has several implications for the design of strand sequencing devices. In particular, the question remains that if a simple hexamer can adopt a variety of conformations within a simple nanopore, then what effect does increasing the length of the strand have on the conformational flexibility? Furthermore, does this have implications for work on sequencing an entire strand of DNA? And, given the ability of DNA to loop, does the possibility that it may translocate in an order not necessarily identical to the starting order have any effect on the reliability of a sequencing read?

Within a nanopore, it is generally established that constriction sites form the basis of read-head recognition of a strand for strand sequencing. These constriction sites already exist within the pore, in the form of the K147/E111 constriction site at the upper mouth of the transmembrane beta-barrel. A second exists at the base of this beta-barrel. However, further constriction sites can be introduced within the pore by mutagenesis. This is done for two reasons: firstly, in order to attempt to improve the resolution between bases by adding to the read capacity (*17*); and secondly, as a consequence of introducing molecular 'brakes' in order to slow down a translocating strand (*20*). The brakes in question often take the form of arginine residues, and thus are rather substantial in terms of steric bulk, occluding part of the pore.

However, as is known from our prior work with DNA hexamers, these additions to the pore, particularly in the form of positive residues, have a substantial effect on observed DNA conformations (*153*). DNA hexamers are far shorter than any strand used in sequencing, and have limited conformational flexibility. Because of this, we sought to examine and characterise the conformations adopted by a longer strand of ssDNA within a nanopore using a cluster analysis method. Cluster analysis seeks to identify similar structures based on their similarity; similarity is determined from a root mean square deviation analysis. This will in theory allow identification of similar conformations across multiple simulation trajectories.

5.2 Methods

The model pore generated in the previous chapters was again employed. Simulations used the 53A6 revision of the GROMOS96 force field (*155*). The simulation system consisted of the model pore embedded in a ~1000-atom methane slab, with one ssDNA dodecamer (*182*) of sequence 5'-ACCGACGTCGGT-3'. This strand of ssDNA was partly pre-threaded into the pore by the first two bases, as preliminary work on 12mer strands had shown that pore entry can take a long time due

to the energetic barrier created by the constriction site. Each base of the strand will be referred to for the rest of the chapter in the following manner:

5’ A1−C2−C3−G4−A5−C6−G7−T8−C9−G10−G11−T12 3’

The ssDNA was prethreaded in the 5’ orientation, with bases A1 and C2 beyond the constriction site. The system was solvated with ~10,000 SPC water molecules, with ~200 Na+ and Cl- ions added for a total concentration of 1 M. All simulations were run for 50,000,000 steps with a timestep of 2 fs, for a total of 100 ns. Each system was repeated 5 times, for a total of 6 simulations for each. Van der Waals interactions were truncated using a cutoff at 1 nm. PME was used for electrostatic interactions, with a short-range cutoff of 1 nm. An electric field of strength 0.1 V nm^{-1} was applied. Cluster analysis used the GROMOS clustering method (*183*) as implemented in the GROMACS g_cluster tool. All generated trajectories for a system were concatenated, and then clustered with an initial RMSD cutoff of 0.4 nm; the resulting structures were reclustered with a cutoff of 0.8 nm to find clusters of broad similarity. Mutant pores were generated using the MODELLER package (*151*) as previously. Three mutant pores were generated: G119R, G119K and G119W, representing the two positive residue types, and a large but non-polar residue. The latter was chosen as questions remained as to whether it was the charged nature of residues that induced folding, or simply a steric effect.

5.3 Results

5.3.1 ssDNA translocation behaviour

The translocation behaviour of the ssDNA as it moved through the nanopore was characterised by marking the time at which the centre of mass of each individual base passes the centre of mass of the N123 ring at the base of the pore, this being the terminal ring of residues of the pore. Any brief excursions, where a base would pass this point but then return to within the pore, were discounted.

A frequent observation within the wildtype nanopore was that base A1, which enters the pore first, is not necessarily the base to exit first. In four out of six simulations, base C2 or C3 were the first base to exit the pore, while only in two simulations did A1 exit first. A potential explanation for this phenomenon is that base A1 would frequently form hydrogen bonds with N123 and N121, resulting in a ‘tethering’ of the base to the pore similar to the phosphate-lysine tethering

observed in previous chapters. This interaction is short-lived, lasting between 2 to 5 ns, but is sufficient to allow the following bases time to pass.

Based on this observation, further analysis considered the translocation times of each base relative to the first base to exit the pore, allowing comparisons between individual bases. Across the wildtype mutants, there is a relatively constant rate of translocation, with the major exception of a single simulation with a time lag of around 50 ns between bases A5 and C6 exiting the pore. The relative translocation times across all simulations that completed translocation are given in figure 1 below.

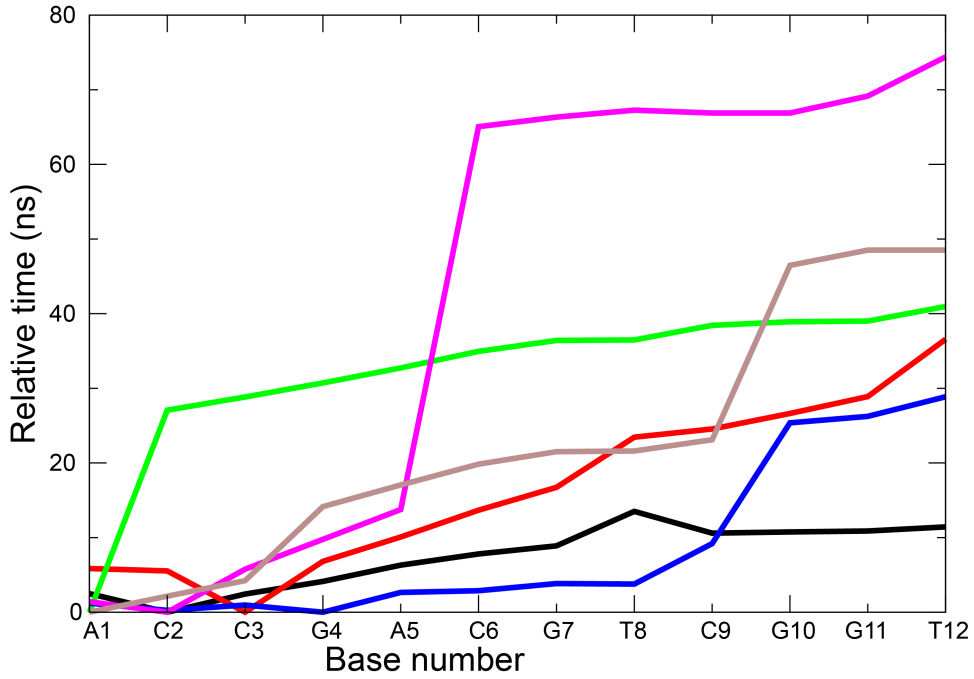

Figure 1. Time evolution data for the wildtype nanopore. Different colours represent different trajectories.

For the G119K mutant, two of the six total simulations show short tethering events where bases C2 and C3 exit the pore first, while another two show exit of the first three bases within 1 ns of each other, indicating that they are exiting in a more coiled conformation. Out of the six simulations performed, three completed translocation within 100 ns, and the time evolution data is given below in figure 2. Average base translocation times in these simulations were more varied than wildtype, with values of 1.3 ns, 2.8 ns and 6.5 ns per base. In the former, translocation occurred in a relatively constant rate. For the latter two simulations the translocation occurred in a staggered manner, with one simulation showing a time difference of 40 ns between bases G7 and T8. This staggered translocation mechanism was also observed in two simulations that showed entry of the ssDNA into the pore but did not complete translocation fully in 100 ns.

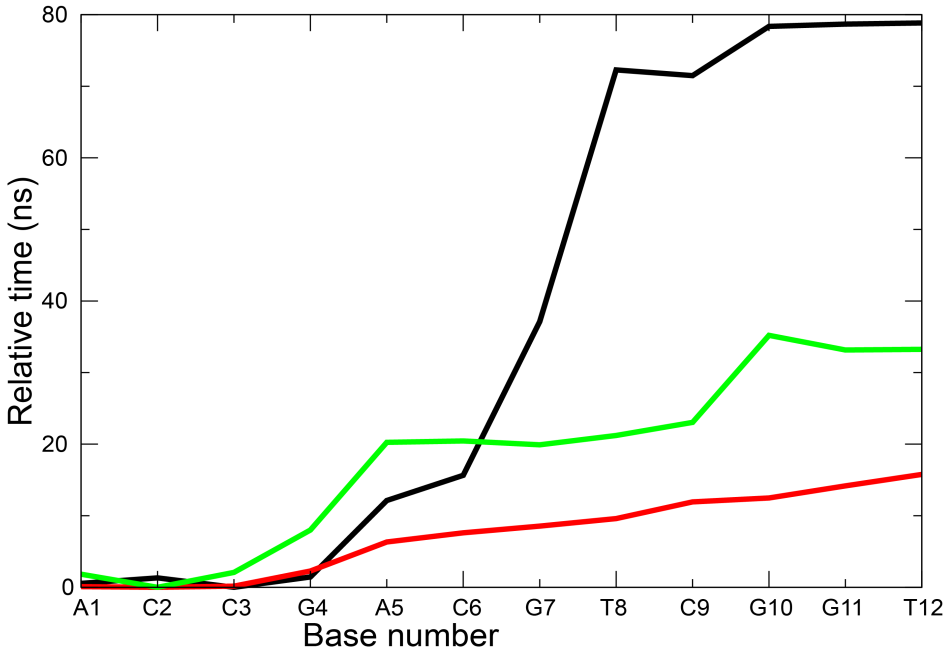

Figure 2. Time evolution data for the G119K mutant nanopore. Different colours represent different trajectories.

In the G119R mutant system, tethering of the ssDNA via base A1 to the N123 ring of residues occurs in four out of six total simulations, with bases C2 and C3 being the first to exit the pore. This tethering usually lasts between 3 and 10 ns, but in one simulation that did not complete translocation the interaction lasted for 63 ns after base C2 had exited the pore; in another simulation, the first and second bases exit the pore within 300 ps of each other. The time evolution data for the three systems that completed translocation is given in figure 3 below, with average translocation times per base of 1.3 ns, 2.7 ns and 5.5 ns. Two simulations show a relatively constant rate of translocation, while the third shows a more staggered rate. In this third simulation it is base A5 that is first to exit the pore.

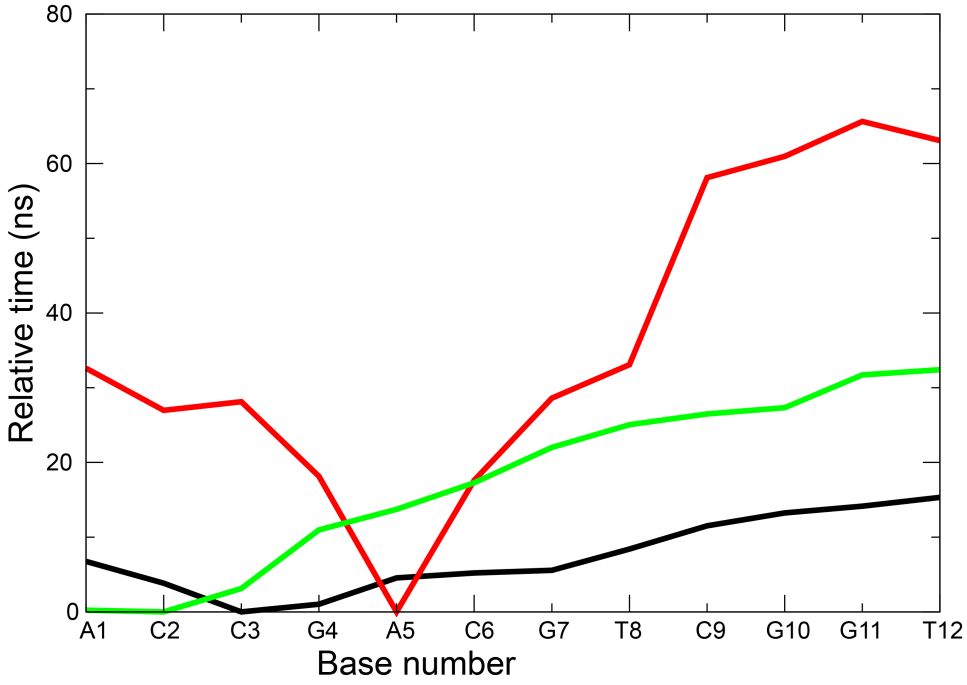

Figure 3. Time evolution data for the G119R mutant. Different colours represent different trajectories.

5.3.2 ssDNA conformations within the protein nanopore

Conformations adopted by the ssDNA within the mutant nanopores were characterised using cluster analysis. In the wildtype system, one major cluster was observed, accounting for 40% of all simulation time, or 240 ns. This is shown in figure 4, below.

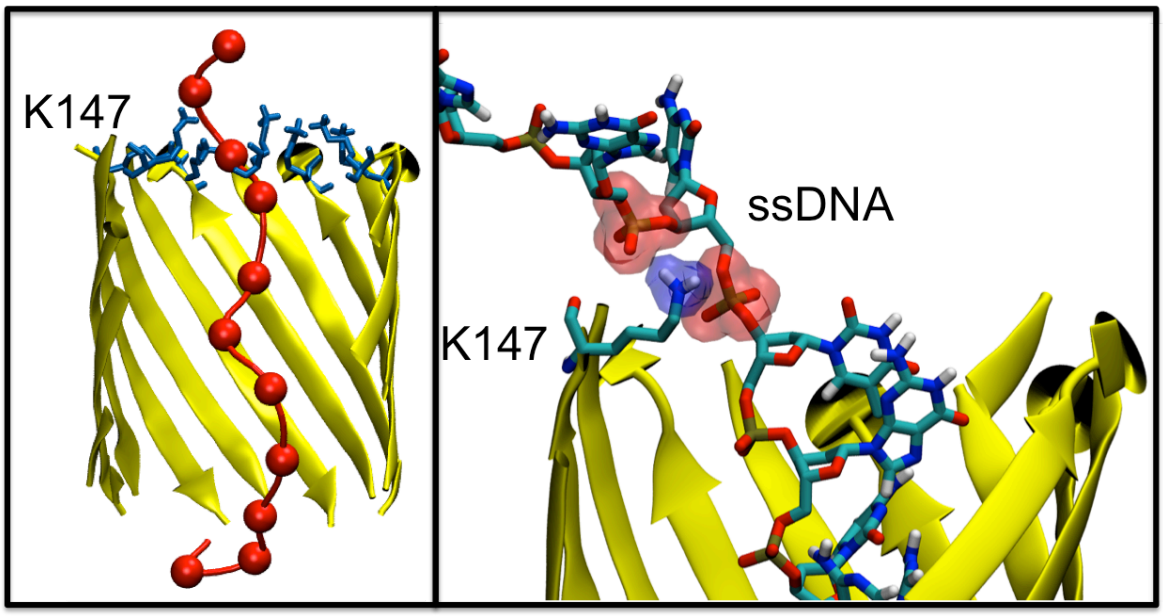

Figure 4. Main cluster in wildtype pore (left) and interaction of ssDNA backbone with lysine sidechain (right). Left panel: backbone phosphate as red sphere, backbone as red tube, protein in yellow cartoon representation, lysine 147 in blue liquorice representation. Right panel: backbone phosphate and lysine sidechain highlighted in red and blue surfaces respectively. ssDNA coloured by atom type.

This primary cluster consists of the ssDNA in an extended configuration within the nanopore (figure 4, left panel), with a DNA end-to-end distance of between 3.75 and 4.5 nm. This

conformation appears to be partly as the result of tethering of the ssDNA to the ring of residues formed by K147/E111, similar to the tethering observed previously in ssDNA hexamers. These interactions primarily consist of hydrogen bonding between the K147/E111 sidechains and the ssDNA bases, usually G10-T12; however, electrostatic interactions are also present, as shown in figure 4, right panel. In this instance the lysine sidechain intersperses between the backbone phosphates of G10 and G11, forming a relatively long-lasting interaction of some tens of nanoseconds.

In the G119K mutant, two major clusters are observed, comprising 42% and 22% of simulation time (252 ns and 132 ns respectively). The clusters observed are shown in figure 5. The first cluster consists of a coiled ssDNA structure within the pore, which forms multiple interactions with the sidechains of K119. These interactions are electrostatic in nature, with the amine group of the lysine sidechains again interspersing between backbone phosphates. Some transient hydrogen bonding was also observed between DNA bases and these amine groups, with lifetimes of a few nanoseconds. The second cluster appears to consist of the ssDNA tethered to the K119 ring, generally by electrostatic interactions, with the majority of the ssDNA having exited the pore.

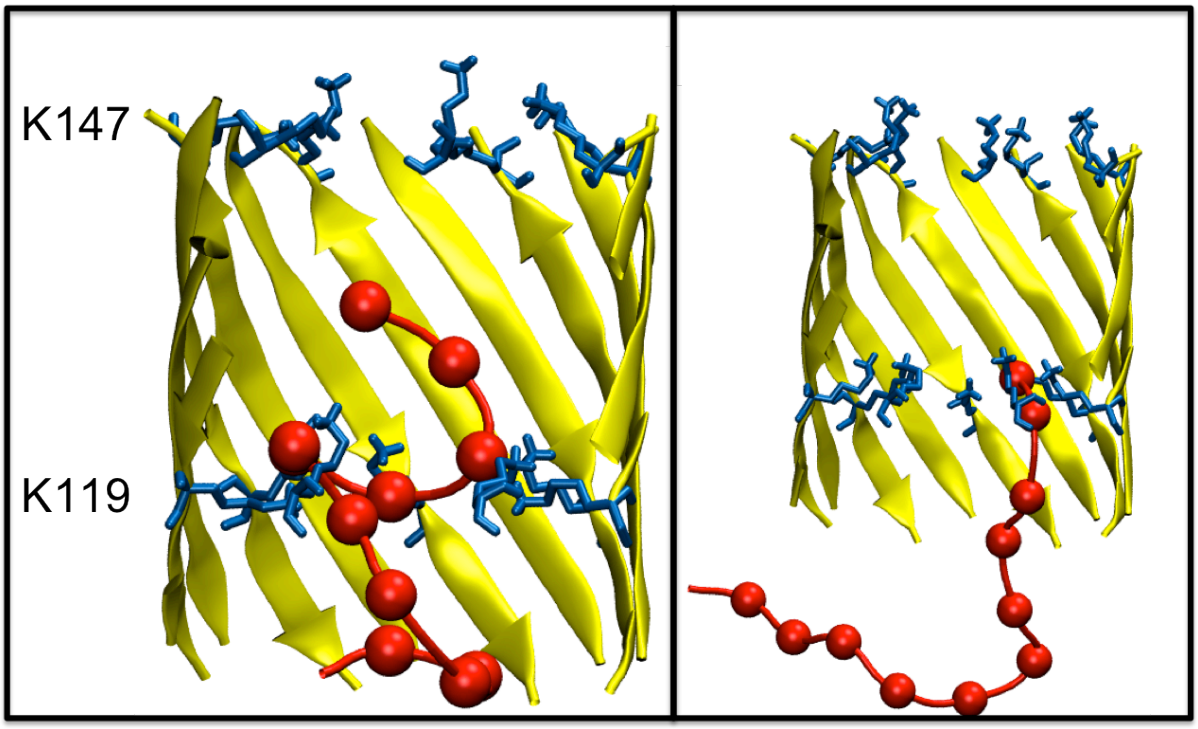

Figure 5. main clusters in G119K mutant. Left panel: cluster formed by wrapping of DNA around K119. Right panel: cluster formed by tethering of ssDNA strand to K119. DNA backbone phosphates in red sphere, DNA backbone in red ribbon trace. Protein in yellow cartoon representation, lysine residues in blue.

For the G119R mutant, three large clusters are observed, two of which are in close proximity to R119, and are shown below in figure 6. Similar to the G119K mutant, these two clusters in proximity to R119 are a coiled structure in which multiple interactions are formed between the ssDNA and the arginine sidechains (figure 6, left panel), and a tethered structure, where the terminal bases of the ssDNA strand remain in contact with the ring of charged residues formed by R119 (figure 6, centre panel). The former accounts for approximately 40% of simulation time, around 240 ns, while the latter accounts for 10% or so of simulation time, or around 60 ns. In both cases, both electrostatic and hydrogen bonding contributions between the ssDNA and the protein sidechains appear to be important, with an average of 3 hydrogen bonds being formed between protein and ssDNA at any one point, but this increases to 5 in the case of the first cluster, in which the ssDNA coils around the sidechains of R119. Interspersion of the arginine sidechains between the phosphate backbone occurs as in both the G119K example and as in the ssDNA hexamers, shown in figure 7. A third cluster was also observed in this mutant, accounting for 21% of simulation time, where the ssDNA forms a looped structure, but not necessarily in contact with the protein (figure 6, right panel). This structure often appears to form as a result of the second cluster, where the 3' end of the strand remains in contact with and tethered to the R119 ring of residues while the majority of the strand has free movement in solution. This freedom of movement can allow the 5' end of the strand to coil such that it comes into contact with bases G7 or T8, forming this looped structure. This structure is stable enough that it persists in solution, even after the ssDNA has exited the pore.

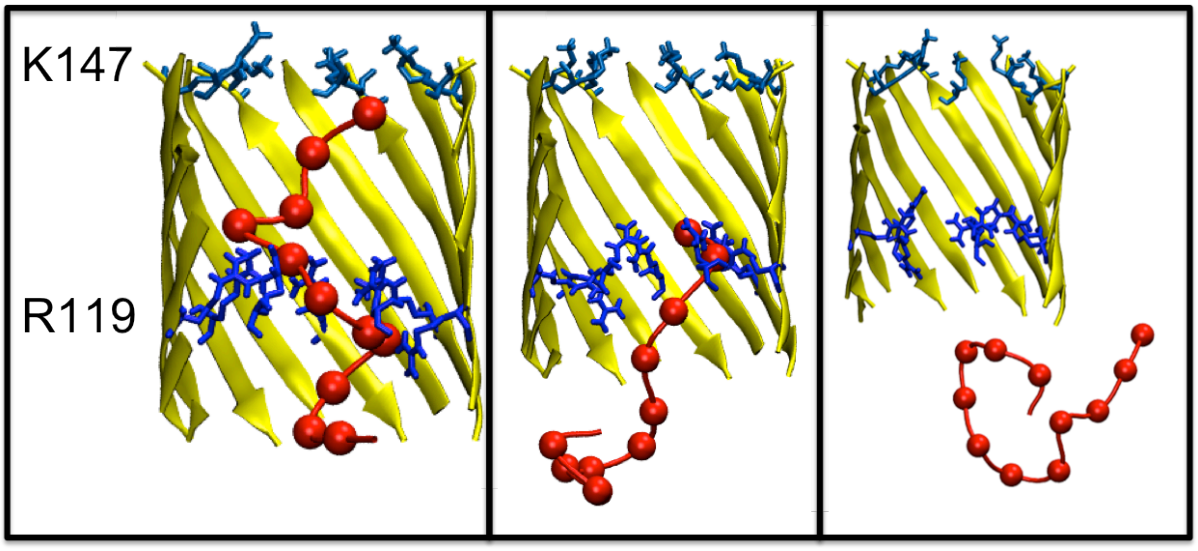

Figure 6. Clusters observed in G119R mutant. Left panel: wrapped cluster formed when ssDNA forms multiple interactions with R119. Middle panel: Tethered cluster formed by interaction of 3' bases with R119. Right panel: solution cluster formed when ssDNA loops and interacts with itself.

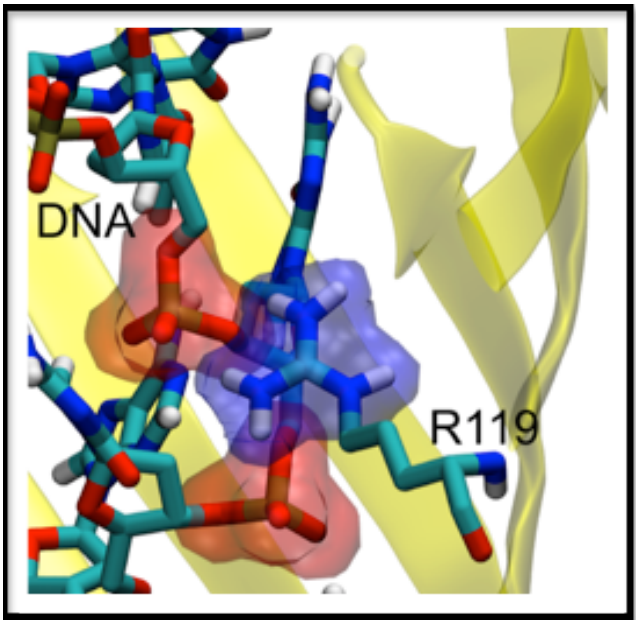

Figure 7. Interaction of DNA backbone with R119. Phosphate and guanidinum in red and blue surface respectively, pore in yellow cartoon representation. ssDNA coloured by atom type.

A non-charged mutant, G119W, was also studied. Thus far, all interactions and induced conformations appear to form as the result of interactions with the charged sidechains of either lysine or arginine, but whether there is a steric component to this induction was thus far unclear. Two broad clusters were observed in this mutant, shown in figure 8 below. The first cluster has the ssDNA in an extended configuration within the pore, but with at most only the first 4 bases translocating beyond the ring of residues formed by W119. This accounts for approximately 67%, or 400 ns of simulation time (figure 8, left panel). The second cluster shows the ssDNA strand interacting mostly with the upper surface of the pore, having not yet translocated beyond the barrier formed by W119 (figure 8, right panel). This second cluster accounts for 25% of simulation time, or 150 ns. This interaction with the upper surface of the pore does not appear to occur in any other simulations, and is likely a consequence of the ssDNA being unable to translocate further. This mutant appears to be non-translocating, and while partial threading of the ssDNA strand through this ring of residues occurs in several individual simulations, translocation does not complete in any. Extension of these simulations by 100 ns does not reveal any further translocation.

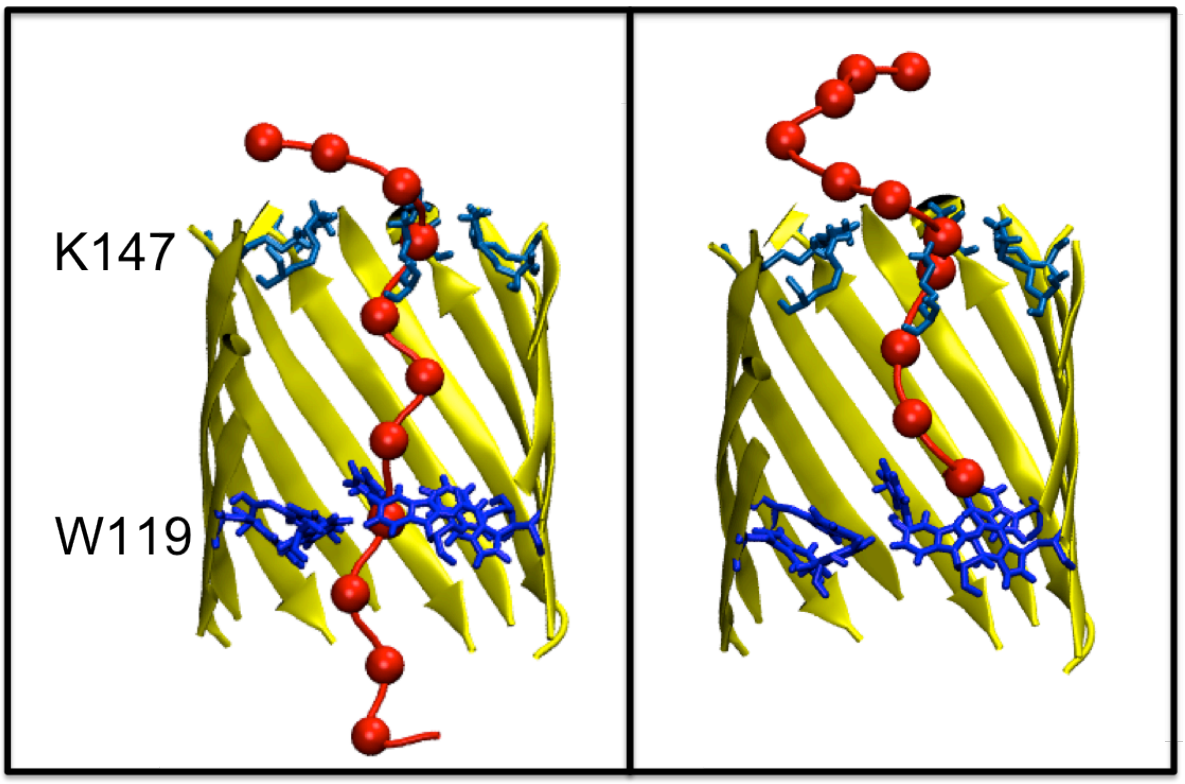

Figure 8. Main clusters observed in G119W mutant. Left: ssDNA straight within the pore, partially translocated beyond W119. Right: ssDNA in contact with W119 but not moving beyond the ring formed by W119; DNA backbone phosphates as red spheres, DNA backbone in red ribbon trace, protein in yellow cartoon representation, K147 in cyan, W119 in blue.

5.4 Discussion

The tendency for the ssDNA to not necessarily exit the pore in the order that it entered was observed across both charged mutants and wildtype, although the charged mutants exaggerate this effect further. This appears to be as a result of the charged sidechains forming interactions with the first bases of the strand while the later bases pass through, thus forming a looped structure, with this interaction lasting tens of nanoseconds in some cases. However, this disruption of order can still occur in the wildtype protein, likely as a result of base A1 forming interactions with N121 or N123 near the base of the pore. It is possible that this looping and disordered exit of the ssDNA from the nanopore could have some effect on the reliability of strand sequencing, since the ordering of the strand is no longer preserved. This also leads to more than one base occupying the mouth of the barrel at the same time, which has also been observed in previous studies (*150*). If this deformation occurs at the same point as the current reading, this again could have implications for read reliability. Whether these observations are only a result of strand entry into the pore is unclear, since these observations are only made at the 5’ end of the

strand, and whether they would be significant in the middle of a long strand warrants further investigation.

Conformational analysis of the ssDNA within the pore shows distinct patterns of behaviour between the charged and uncharged nanopores. The charged nanopores induce significant deformations in the ssDNA, shown by the large clusters in G119K and G119R where the ssDNA wraps around the mutated charged residues. The wildtype and G119W nanopores instead have a dominant cluster that consists of a strand in the extended configuration. It is, of course, possible for more folded configurations to form in the uncharged mutants and extended configurations to form in the charged mutants, but these are by no means the dominant configurations. The deformations in G119K/G119R appear to be the result of the ssDNA maximising the number of electrostatic interactions between the backbone phosphates and the positive sidechains.

The observed translocation mechanisms across all mutants involve some degree of the "binding and sliding" mechanism described in chapter 3. The translocations themselves are often staggered, due to interaction of the ssDNA bases with the pore itself. For example, in one system with an average translocation time of 4 ns per base, bases G11 and T12 remain tethered to K147 for ~45 ns, thus leading to a particularly non-uniform translocation rate. The residues most often interacted with across all simulations are the K147 ring at the mouth of the pore, and the R119/K119 residues in the charged mutants. This non-uniform DNA translocation rate has been noted in other studies (*150*). The binding and sliding mechanism does appear to be more common in the charged mutants due to the presence of the extra interaction sites, and this slows down translocation, shown in part by the fact that fewer simulations of the charged mutants complete translocation in 100 ns relative to the wildtype systems.

There is little difference between the G119K and G119R mutants in terms of ssDNA translocation properties. It does appear that the interaction between G119R and ssDNA is a slightly stronger interaction than G119K, with hydrogen bonds and charge-charge interactions forming between the R119 ring and the ssDNA in some 73% of the total simulation time, compared with 65% of the total translocation time between ssDNA and K119 in the G119K mutant.

The G119W mutant nanopore shows some further insight into translocation mechanisms. Both G119R and G119W have similar pore dimensions at the constriction site formed by the mutant residues, but W119 forms more of a physical barrier to translocation, with few bases translocating past this point, and most coiling of the ssDNA occurring above rather than around the W119 ring. The discrepancy between the two mutants appears to occur as a result of the electric field applied

to the system, which allows the arginine sidechains to align slightly with the field and thus marginally opening the pore. Tryptophan, being nonpolar, does not align with the field and thus the pore remains narrower. In addition, base A1 does not interact with W119 once it has passed the barrier formed by this residue, and thus the coiling observed in other mutants does not occur, with the ssDNA remaining in an extended configuration. The hydrophobic nature of these residues likely contributes to the lack of translocation due to no favourable interactions being formed between the sidechain and the ssDNA.

5.5 Conclusions

It has been shown that the sidechains of the nanopore have a direct effect on the conformations adopted by an ssDNA strand. Charged residues in particular induce deformations in the DNA into a non-linear conformation, frequently wrapped around the sidechains of the residues in question. The ability of the ssDNA to exit the nanopore in an order that it did not enter in is notable in that it may alter the sequence observed in experiment. The wildtype nanopore appears to allow ssDNA to translocate in an extended conformation rather than the induced coiling observed in the charged mutants, and translocation occurs more quickly. Introduction of large but hydrophobic residues does not induce the same deformations of the ssDNA as charged residues, and indeed appears to completely block translocation instead of merely slowing it.

Chapter 6: Free energies of translocation through the α-hemolysin transmembrane barrel

Abstract: Translocation of ssDNA through nanopores has been discussed in previous chapters. However, one aspect that has not been explored thus far is the free energies of translocation. Free energy is a property that drives much of the translocation dynamics discussed previously, with the free energy at a given point determining the likelihood and speed of translocation past that point. In this chapter umbrella sampling calculations are performed using an extended model of the α-HL transmembrane barrel. The components of a nucleotide are considered individually, i.e. phosphoric acid and the bases are separated. The results presented include phosphate, cytosine and a preliminary adenine potential of mean force profile. The results show similarities between the two bases studied, as well as a general contrast in the features between the charged phosphate and the nonpolar bases.

6.1 Introduction

In previous chapters, the effects of DNA length, pore mutation and DNA parameter sets on translocation behaviours have been discussed. Important information can be gleaned from such studies, but one aspect overlooked thus far has been of free energies of translocation. Free energies determine the 'favourability' of different states or pathways and the likelihood of that state or pathway being adopted. Formally, free energy is defined as the energy of a system available to do work.

The *internal* energy of a system is formally defined as the energy contained within a thermodynamic system; that is, the energy required to create a system. The internal energy of a system is the sum of two components, the potential energy and the kinetic energy, or mathematically:

$$U = U_{pot} + U_{kin} \tag{1}$$

Where U is the total internal energy, U_{pot} is the potential energy component and U_{kin} is the kinetic energy component. Kinetic energy is the energy of motion within the components of the system; potential energy is the energy contained within the static components, such as bonds or electrostatic potentials.

However, this does not define the total energy of a system: the total energy of a system also includes the energy required to displace the surrounding environment where the system exists. This total energy is referred to as *enthalpy*, H, such that:

$$H = U + PV \tag{2}$$

Where H is system enthalpy, U is internal energy, P is the absolute pressure and V is the system volume.

The first law of thermodynamics, which postulates the conservation of energies in thermodynamic processes, states that '*The change in internal energy of a system is equal to the heat added to the system minus the work done by the system on its surroundings*', or mathematically:

$$\Delta U = Q - W \tag{3}$$

Where ΔU is the change in internal energy, Q is heat absorbed by the system and W is the work done by the system. Energy is conserved by the fact that the energy change of the system is equivalent to the energy added minus the energy removed. Therefore, to change the internal energy, some form of external action must be applied. In an isolated system, therefore, the total ΔU is always zero as no energy transfer from the surrounding environment is possible.

The first law does not take into account directional preferences of processes. For example, heat transfer will spontaneously occur from an area of high temperature to an area of low temperature; however, the reverse is not true. Reversal of this process requires input of external energy. The first law has no provision for this, as the total internal energy is constant. This requires the introduction of the concept of *entropy*. The second law of thermodynamics formulates the concept of entropy, such that:

$$\Delta S = \frac{Q}{T} \tag{4}$$

Where Q is heat absorbed by a system to drive a process, T is temperature and ΔS is the change in entropy. The change in entropy in an equilibrated, isolated system is therefore zero, as no external heat can be added to further drive a process. Entropy is a measure of disorder in a system, and always remains the same (in a system in equilibrium) or increases. Another definition is the energy *unavailable* to do work, being related to the random, disordered, thermal energy of a system.

Free energy can therefore be defined from the above equations. If the internal energy of a system is the total energy, and the entropic component is the unavailable energy, then free energy is the difference. There are two basic types, Helmholtz free energy and Gibbs free energy. Helmholtz free energy is the energy required to create a system in the absence of pressure or volume changes; however, unlike the definition of internal energy given above, Helmholtz free energy includes a term, TS, which is the product of temperature and entropy, and is an expression of the energy that can be obtained from heat transfer from the system's environment. Helmholtz free energy is therefore defined as:

$$F = U - TS \tag{5}$$

where F is the Helmholtz free energy, U is the internal energy, T is temperature and S is the entropy of the system. Creating a higher entropy state means that less work is required to create the system as a whole.

Helmholtz free energies assume no change in pressure or volume. Gibbs free energy is the energy required to create the system and displace the volume required by the system, expressed as

$$G = U - TS + PV \tag{6}$$

However, as noted above in equation 2, this includes the definition of enthalpy. Therefore, Gibbs free energy is most commonly expressed as

$$G = H - TS \tag{7}$$

Where G is the Gibbs free energy, H is the enthalpy, T is temperature and S is entropy. This formal definition is the absolute free energy of a system from creation from a negligible volume.

Absolute Gibbs free energy values are less useful than a measured change in free energy. This change in energies is essentially the energy gained or lost from a change in system state. This is expressed as

$$\Delta G = \Delta H - T\Delta S \tag{8}$$

Where ΔG is the change in Gibbs free energy, ΔH is the change in enthalpy, T is the temperature, and ΔS is the change in entropy. This measure is particularly useful as it allows, for example, the quantification of energy changes between two states. A negative free energy indicates that this change is favourable, and will occur spontaneously, while a positive energy indicates an unfavourable, non-spontaneous change that requires energy to drive it.

A potential of mean force (PMF) is a type of free energy calculation that provides a free energy value as a function of a reaction coordinate, which is calculated from the average force of all configurations of a system (*184*). PMF calculations have been used in a variety of contexts, from simple ion-ion PMFs (*185*) to the calculation of free energies of drug molecules across a lipid bilayer (*146*). They have proven useful in the study of ion permeation through pores by defining the regions that are selective for a particular ion; the same principle can be applied to the study of nucleotides through α-HL.

Ion permeation through gramicidin A channels has been studied using potential of mean force calculations (*186*), where comparisons between potassium and chloride were drawn. Gramicidin A is part of the gramicidin group of bactericidal peptides. Gramicidins are peptides composed of 15 amino acids which form ion-permeable pores in the bacterial cell membrane by forming a dimeric structure. The results showed a marked free energy preference for potassium over anions or divalent cations within the channel, and the calculated data is in good agreement with experimental data. The selectivity of cations over anions is partly thought to be a result of the delta-negative backbone carbonyl groups of the protein, which form much of the channel, being unable to solvate anions as they pass through. This also appears to result in disruption of the pore. The monovalent cation selectivity is thought to be caused by the excess solvent required to solvate divalent cations – this excess water is found to disrupt the membrane and the channel in a bid to completely solvate the ion, and a large free energy barrier for divalent cations was observed.

Multi-dimensional PMF calculations were also used to compare free energies in the KcsA selectivity filter to measure the differences between the binding energies of sodium and potassium (*187*). A multi-dimension PMF uses two or more reaction coordinates when calculating the free energy.

Phosphate permeation through the OprP phosphate channel of *Pseudomonas aeruginosa* has been studied using PMF calculations using umbrella sampling and the GROMACS package (*188*). Like all known outer membrane proteins, OprP is a beta-barrel, and consists of 16 strands. It is specifically a porin, and therefore a trimer. OprP is expressed in conditions of phosphate starvation and is responsible for enhancing phosphate uptake into the bacterial cell. These calculations reveal two favourable energy wells within the centre of the channel where phosphate may bind during transport, between two regions of positive residues that serve as the phosphate carriers, described as a 'ladder'. A critical lysine residue between the two sites, K121, is thought to play a key role in mediating transport, being in contact with a transported phosphate in both positions, and is presumed to 'sweep' the phosphate through the pore. The pore also was shown to discriminate energetically between phosphate and chloride, unusually among anion transporters, and to transport in a unidirectional fashion.

Applied to the α-HL channel, potential of mean force calculations could be of use in determining the relative free energies of translocation for each nucleotide. PMFs for nucleotides in α-HL have been calculated previously, using steered molecular dynamics and the Jarzynski identity (*142*);

however, umbrella sampling along the reaction coordinate of interest – the transmembrane axis of the protein – has not been attempted. Umbrella sampling is a technique that performs multiple independent simulations of an analyte of interest at different points along a reaction coordinate within a system. These simulations allow free movement of the analyte of interest in two planes, but restrain it in the third. Any force acting against this restraint at each position is recorded. If the analyte is in a favourable position this force will be minimal, while if it is not a greater force will be recorded as the analyte tries to move from that position. This umbrella potential creates a bias - the final step in the calculation is the removal of the bias and the reconstruction of the free energy profile. A schematic of an umbrella sampling simulation setup is given below in figure 1.

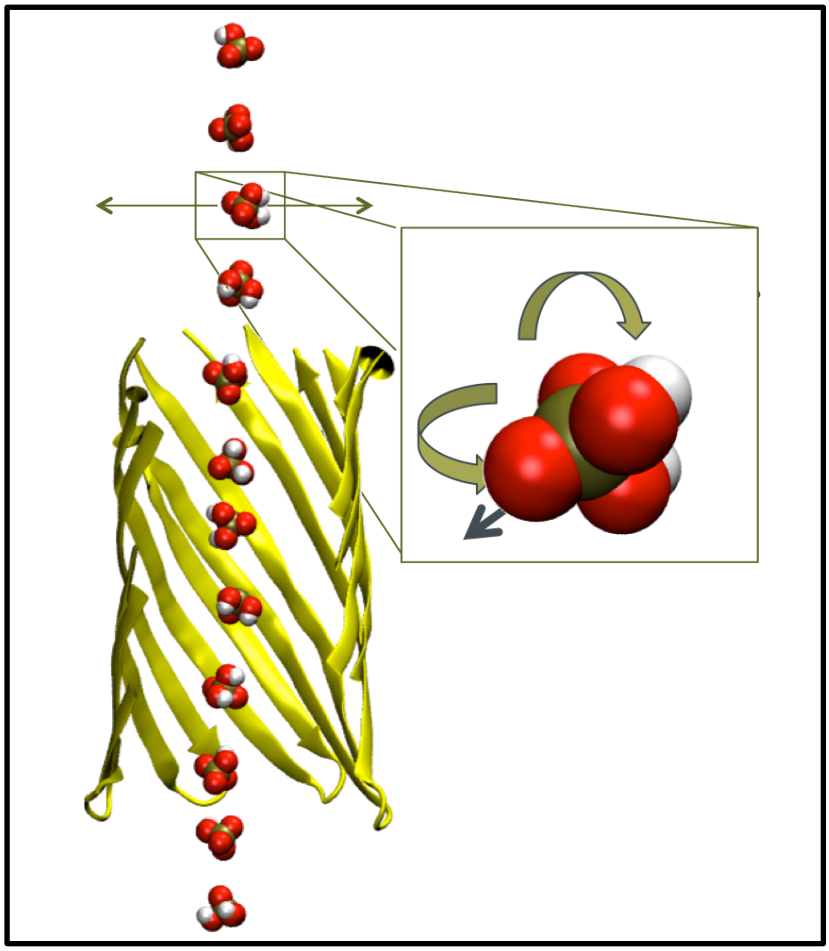

Figure 1. Schematic of an umbrella sampling simulation setup. Barrel shown as cut-away representation; water, ions and methane slab omitted for clarity. Individual phosphates are in separate simulations. Molecules of interest must sample all possible configurations. Total z axis length in the system of interest was approximately 8 nm.

Being able to identify the most favourable binding or interaction sites through umbrella sampling calculations could be useful in manipulating the channel to better serve as a DNA sequencing device. The difference between these favourable and unfavourable sites is the key to understanding which points serve as the detection sites for each base, and altering these could

allow better resolution between each base, a desirable goal in the development of nanopore sequencing.

6.2 Methods

Theoretically, a potential of mean force can be calculated for any molecule as long as all possible configurational states are sampled. However, this configurational sampling requires proportionally longer times for bigger molecules. While a calculation can be performed relatively simply and cheaply on a symmetrical atom or molecule such as methane, the required times become much longer once the complexity of the molecule of interest grows. Beyond a certain size, which can vary depending on the local environment, simulation becomes impractical.

This is also compounded by the type of calculation being performed. Umbrella sampling (*144*) is a rigorous method for potential of mean force calculations; however, while widely used, the drawback is in the time required. Both the system size and molecule complexity increase the simulation time required. A long simulation time per window thus reduced the usefulness of the technique if applied inefficiently.

Most of the efforts in the field of nanopore sequencing have concentrated on sequencing either through strands or through release of individual bases. However, even a full nucleotide is too large to consider because of the time required for it to sample all possible configurations. In a confined geometry such as a nanopore, this becomes more of a problem as the limited space may prevent full rotation in the same time period as might be expected in aqueous medium.

To overcome this, a simplified approach was again used. As a full nucleotide – base, phosphate and sugar – is too large for practical simulation due to size and therefore required simulation time, the decision was taken to split the nucleotide into its basic components and simulate for them individually, specifically the phosphate component (the charged component of the molecule) and the base component (the base being what determines the difference between each nucleotide). The structures of the simplified base components are given in figure 2 below.

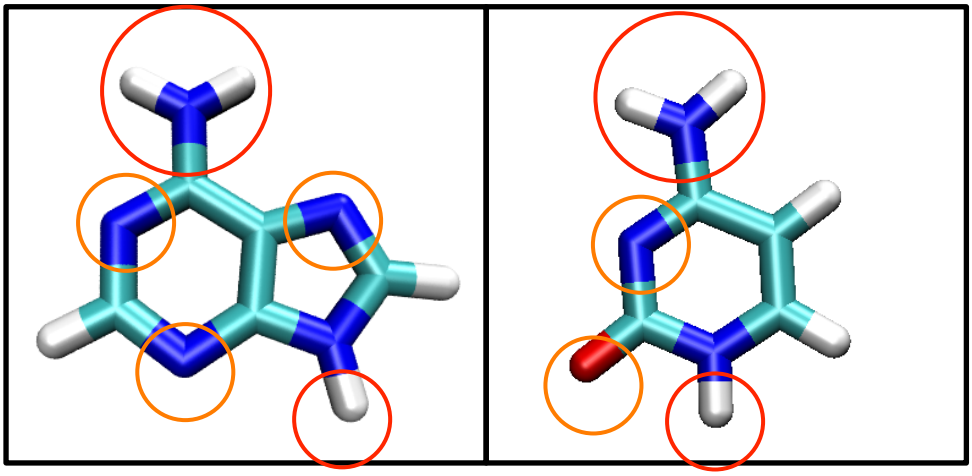

Figure 2. Structures of the simplified adenine (left) and cytosine (right) bases used in this study. Hydrogen bond donor sites are marked with red circles. Hydrogen bond acceptor sites marked with orange circles.

Use of these bases gives data for both a purine and a pyrimidine base; both are also the smallest bases of their type, hopefully reducing the sampling time required. Furthermore, both are bases from different base pairs, with adenine forming two and cytosine forming three hydrogen bonds with its associated pair – however, adenine normally has four potential donor and acceptor sites, and cytosine three. Both models introduce an extra hydrogen bonding site in the form of the hydrogen that replaces the connection to the sugar group, but these are equivalent in both models. As the relative free energies observed in a PMF calculation are in part dependent on interactions formed, this difference in the number of potential hydrogen bonding sites could be part of what distinguishes between these two bases. The relative sizes and hydrophobicities will likely form another cause of energy differences, with a more hydrophobic base being less likely to find an aqueous environment favourable, and seeking to minimise interactions with solvent wherever possible; in this instance, adenine is the more hydrophobic base.

Instead of calculating a potential of mean force for the whole protein, the barrel region was again chosen. Instead of using the shortened barrel of the model pore described previously, a barrel encompassing the whole length of the transmembrane region was created using the same procedures as in (*153*). The GROMOS 53A6 forcefield (*155*) was used.

The phosphate analogue chosen was dihydrogen phosphate, $H_2PO_4^-$. At neutral pH, approximately 70% of phosphoric acid is in this configuration, with the second hydrogen having a pKa value of 7.2. Like each individual phosphate in the backbone of a DNA strand, the dihydrogen form of phosphoric acid has an overall charge of -1. As such, $H_2PO_4^-$ is the closest phosphoric acid analogue to a DNA backbone, although a phosphate with a charge of -1 but no hydrogen atoms could also have been created as a model system. Each individual base was created using the existing full nucleotide parameters as a reference, and correcting for a total charge of zero.

Each PMF calculation consisted of ~140 individual umbrella sampling windows, spaced at 0.5 Å intervals. The required time for each window varied by molecule, from 100 ns in the case of the phosphate simulations to ~300 ns for the cytosine base simulations. A harmonic potential with a force constant of 1000 kJ mol^{-1} nm^{-2} was used to restrain each molecule in the Z axis of the simulation while allowing free movement in the X and Y planes; this force constant was used in (*146*). The weighted histogram analysis method (WHAM, (*145*)) was used to remove the bias and construct the PMF profile.

6.3 Results

6.3.1 Phosphate profile

The calculated profile for the dihydrogen phosphate ion is given below in figure 3. For reference, a schematic of major residue positions within the α-HL transmembrane β-barrel is given below that in figure 4.

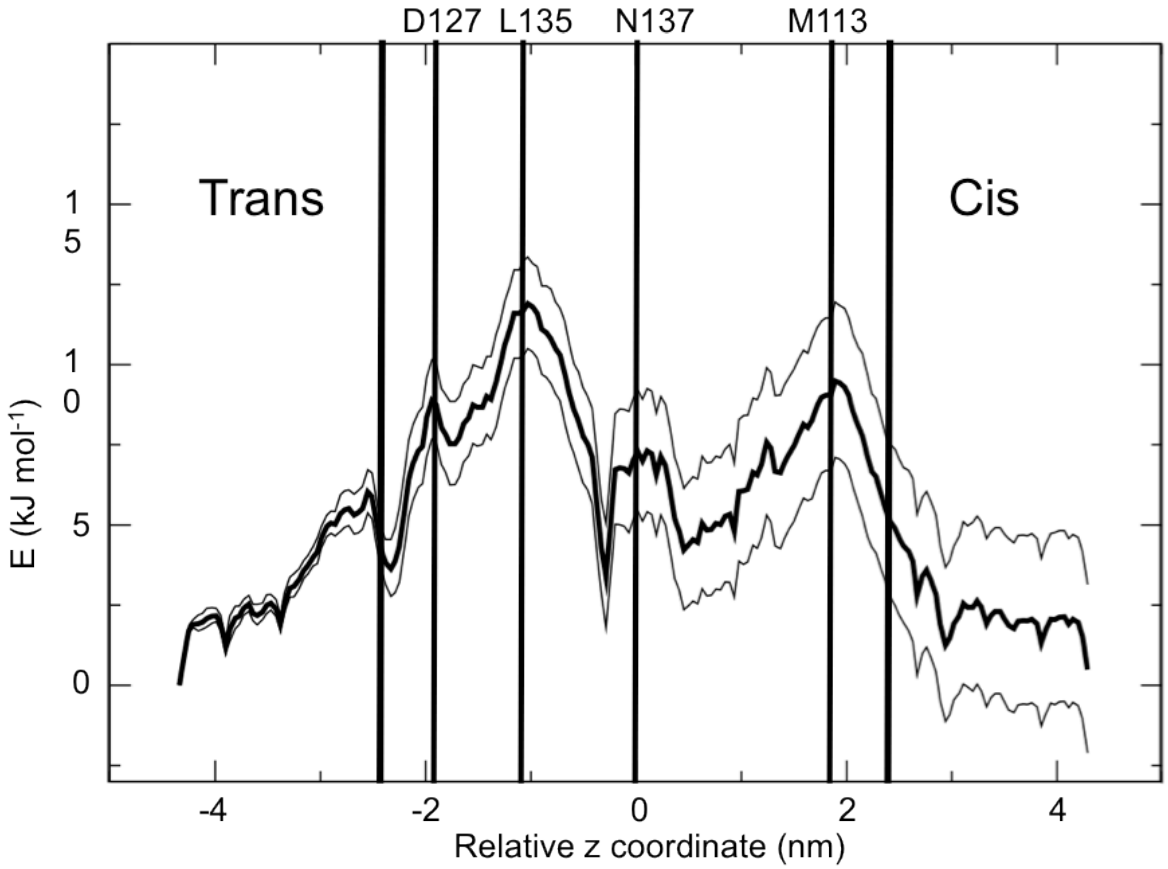

Figure 3. PMF profile for phosphate ion. Pore entrance and exit are marked; specific features are labelled. Upper and lower lines represent one standard deviation as calculated by bootstrapping error estimation.

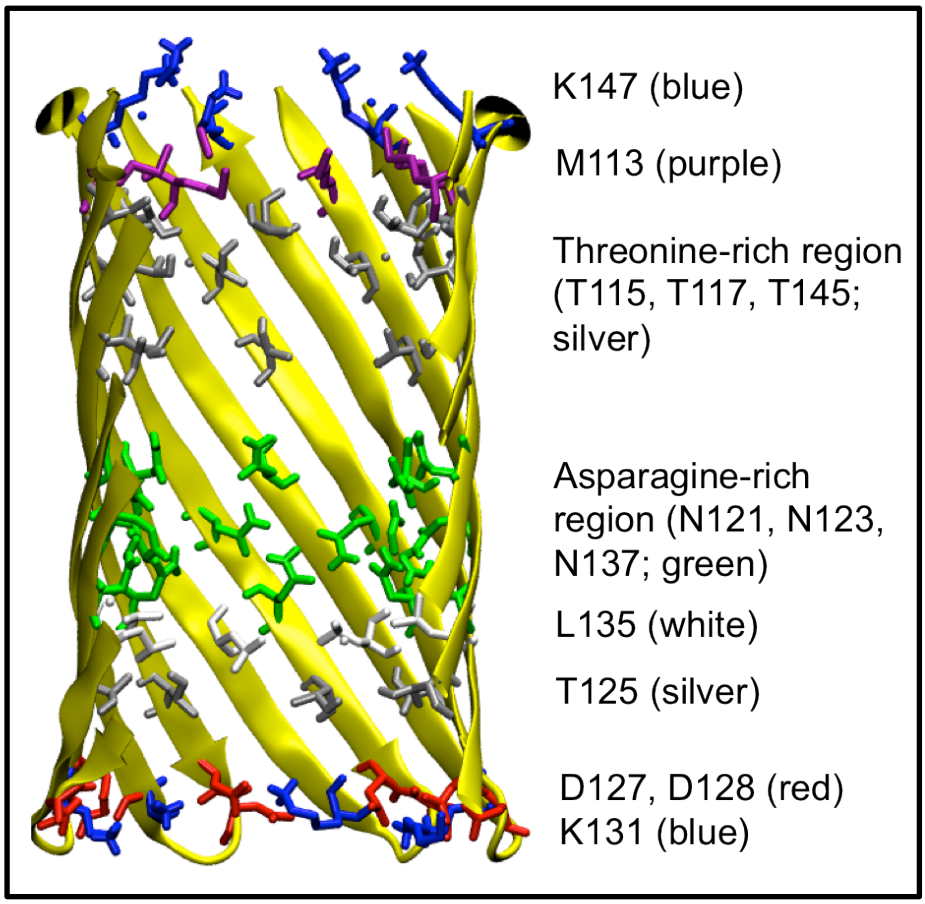

Figure 4. Important sidechains within the transmembrane β-barrel. Colours are given in the figure.

The phosphate profile reveals several features. The deepest trough, with a relative free energy value almost equivalent to bulk solvent, at z = -0.5 corresponds to a region rich in asparagine residues, specifically N137, N121 and N123. These asparagine residues create multiple favourable interaction sites for phosphate, with the windows at z = -0.55, z = -0.5 and z = -0.45 allowing the phosphate to sit within the pocket formed by the asparagine sidechains. This is illustrated in figure 5 below, with hydrogen bonds between the sidechains and the phosphate marked.

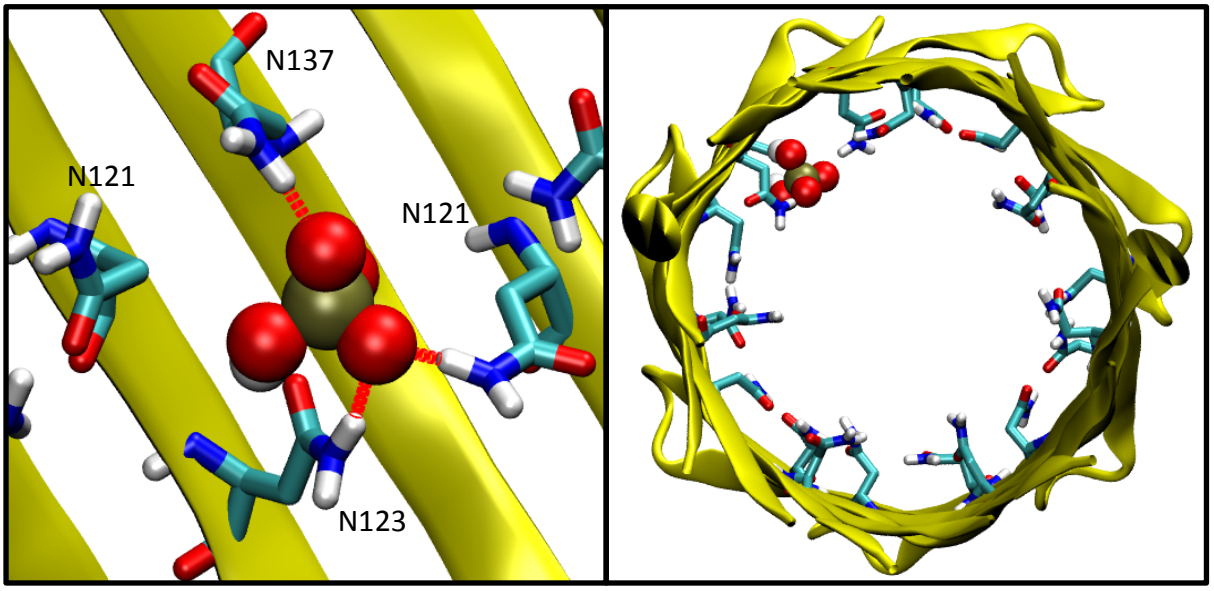

Figure 5. Left: Sidechain interactions between phosphate and asparagine within the pocket formed by N121, N123 and N137. Hydrogen bonds marked in red. Right: position of phosphate as seen from above the pore.

There are peaks at leucine 135 and aspartate 127, which are hydrophobic and negatively charged regions respectively, and therefore unfavourable to a negatively charged phosphate. Within the leucine 135 region the phosphate tends to remain closer to the centre of the pore, avoiding interaction with the leucine sidechains and not sitting close to the wall of the pore as in the asparagine region. The free energy at this position is 12 kJ mol^{-1} higher than in bulk solvent, the largest observed peak in the phosphate profile. This interaction is illustrated in figure 6.

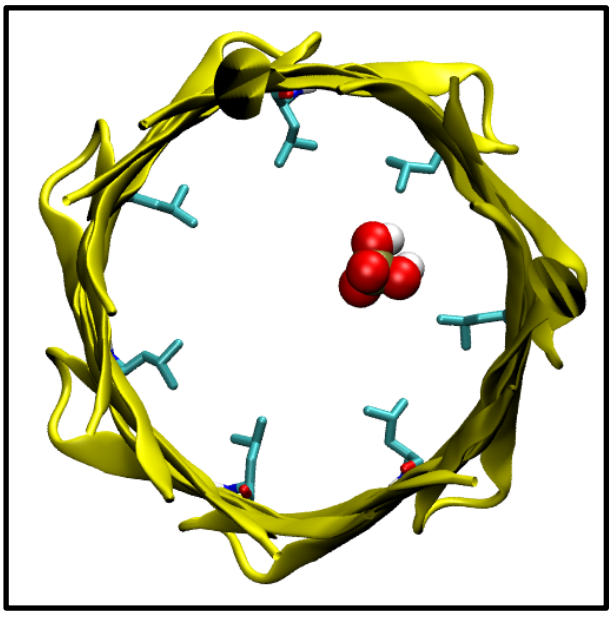

Figure 6. Positioning of the phosphate within the leucine 135 region of the pore. The phosphate samples within the centre of the pore but tends to avoid close interaction with the sidechains.

Another minor peak exists near methionine 113, another hydrophobic residue. The pattern of positioning of the phosphate within the barrel lumen is similar to the position near leucine 135. The peak itself, however, is smaller than the leucine peak by around 4 kJ mol^{-1}, likely due to the positioning of lysine 147, a positively charged residue directly above methionine 113. The interaction with this positive residue likely mitigates some of the unfavourability of the positioning of a negatively charged molecule in a hydrophobic environment. This positioning is shown in figure 7 below.

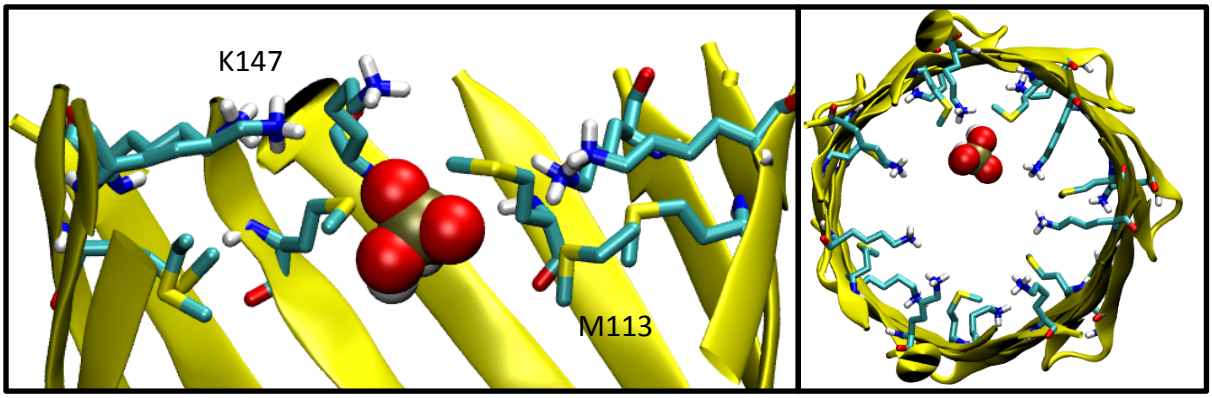

Figure 7. Positioning of phosphate in M113/K147 region. Left: interaction from side. Phosphate tends to avoid contact with methionine but interacts with lysine. Right: interaction from above.

6.3.2 Cytosine profile

The cytosine profile is given below in figure 8. The most favourable regions are at z = -1.8; curiously, this corresponds to a region near the charged residue D127 and the polar residue T125. The least favourable regions within the pore are around z = 0.5, with an energy of -4 kJ mol^{-1}. This corresponds to a pore region with multiple polar residues, specifically arginine and threonine, but is still a more favourable location than in bulk solvent. The least favourable regions overall are in bulk solvent, which is to be expected from a hydrophobic base. The solvent energies on both sides of the pore do not match, which is likely a consequence of the simulations not converging and thus requiring more simulation time. Block analysis of the umbrella sampling trajectories confirms that the simulations remain unconverged.

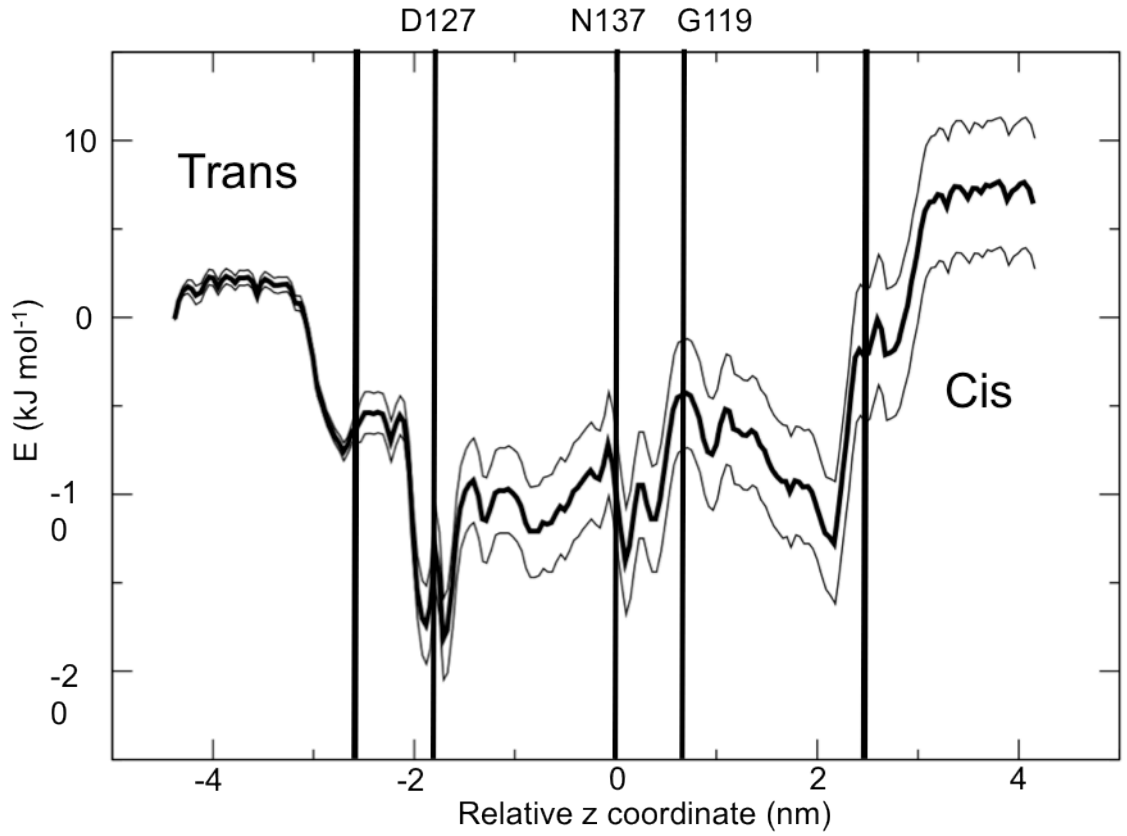

Figure 8. Cytosine PMF profile. Pore entrances and exits are marked, as are features of interest. Upper and lower lines represent one standard deviation as calculated by bootstrapping error estimation.

The interactions that result in the favourable region at z = -1.8 are displayed below in figure 6. The energies at this region are approximately -17.5 kJ mol^{-1} compared to bulk solvent, the most favourable energy observed for the cytosine base. This favourable interaction does not appear to arise as a result of favourable electrostatic interactions or hydrogen bonding, as at any one time there is only a single hydrogen bond formed between cytosine and usually threonine 125, which would account for, at most, 10 kJ mol^{-1} of this interaction – it would appear that much of the favourability arises either from the entropic effect of reducing water contacts with the base, or from van der Waals interactions between the base and the protein. Both of these factors may also be responsible as a whole for the greater favourability of the pore relative to bulk solvent. The location of the cytosine at this position is illustrated in figure 9, showing in particular the positioning of the base as close to the wall of the pore as possible. The presence of nearby aspartate residues does not appear to produce unfavourable energetics at this site, which might otherwise be expected of the interaction between a charged and a non-polar group. It may be that, while the presence of this aspartate is energetically unfavourable, the favourable contributions outweigh this.

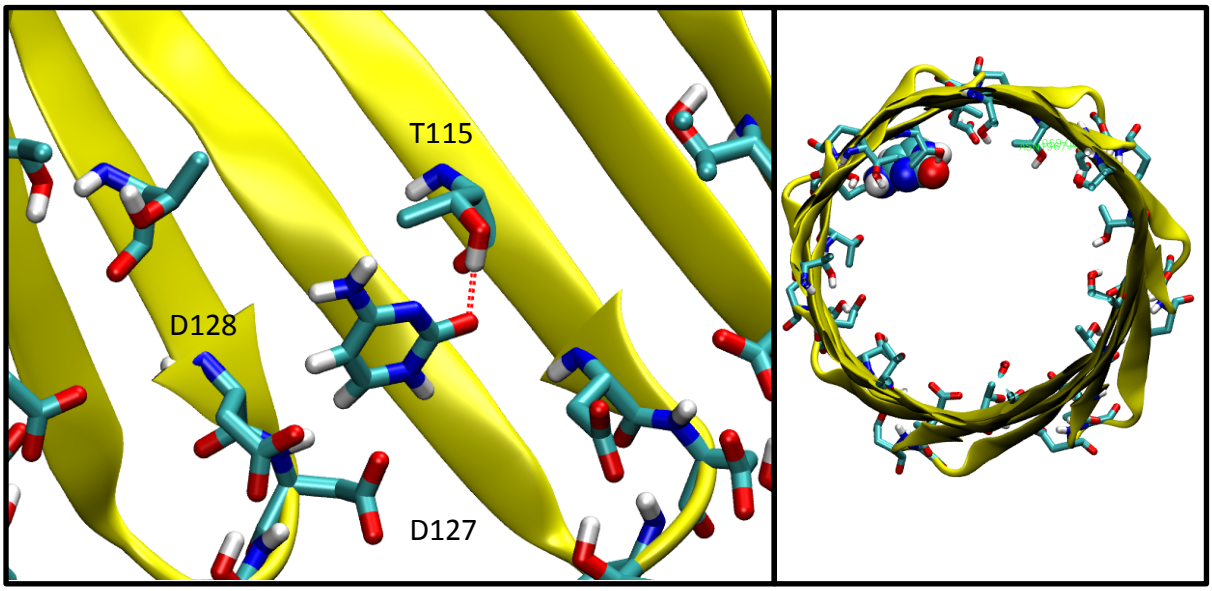

Figure 9. Interactions of cytosine at the most favourable region of the pore (left). Usually only a single hydrogen bond is formed with the protein at any given time. The frequent positioning of the cytosine at the side of the barrel (right) maybe suggests that the favourability is the result of lack of interaction with water.

The peak at z = +0.5 arises in a region of more polar residues. Nearby residues include N123, T117 and S129. It appears as though the cytosine base is too large to fit into the pocket formed by these residues, and as such is slightly more exposed to water than in the case of the D127 region. This may account for the large energy difference between both regions, although this could also be a function of the profile remaining unconverged. This interaction is shown in figure 10 below. Calculation of the average solvent accessible area of cytosine at this position gives a value of 0.97 nm^2, compared with a value of 0.79 nm^2 in the D127 region, which again could suggest the favourability of this area is the result of fewer water-base interactions, although whether this completely accounts for the 12 kJ mol^{-1} difference in energies is unclear.

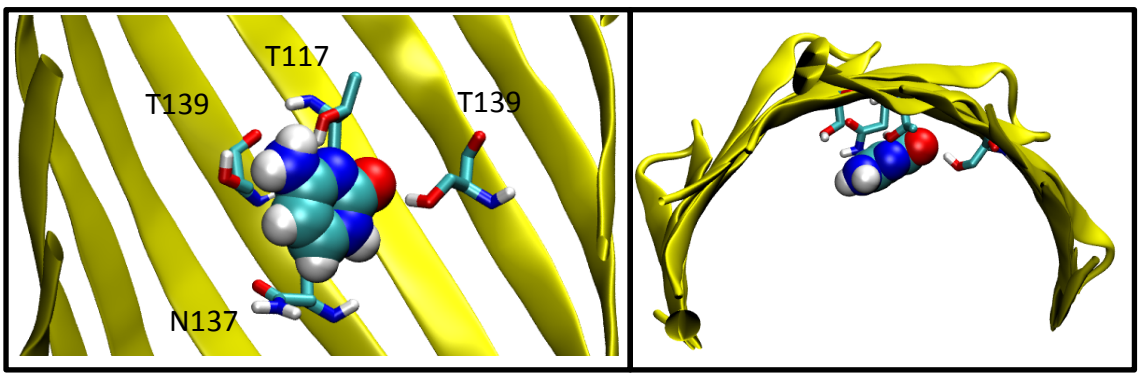

Figure 10: positioning of cytosine within the G119 area of the pore, in contact with T117, T139 and N137.

The features of the cytosine profile can be compared to those of the phosphate profile, and for the most part are opposites, with the troughs of one generally corresponding to the peaks of the other. This is shown in figure 11 below. In particular, the regions at z = -2 and around z = 0, as well as at both the cis and trans exits, are mostly opposites.

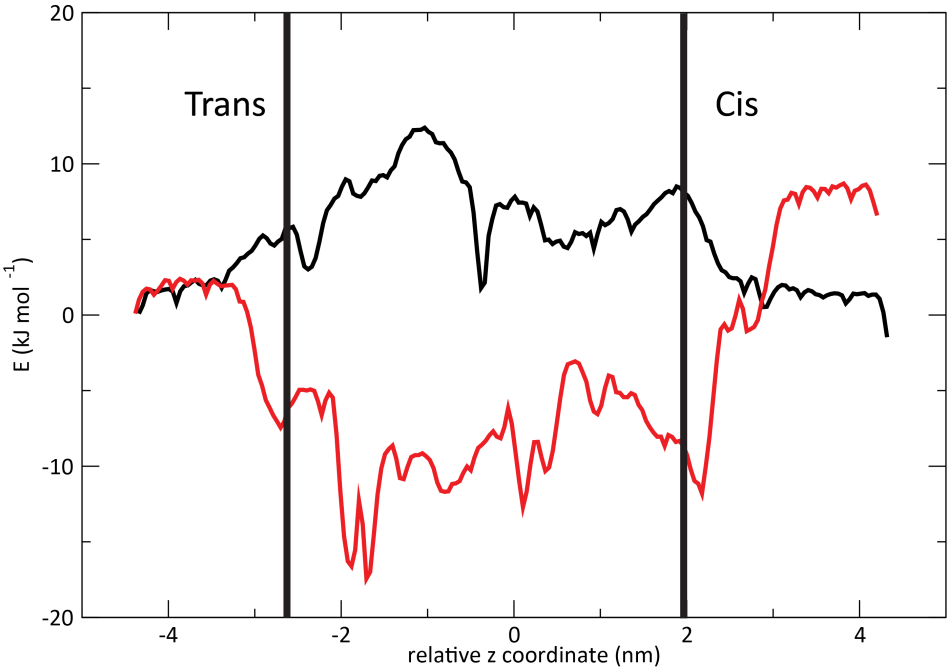

Figure 11: Superimposed phosphate (black) and cytosine (red) profiles.

6.3.3 Adenine profile

As of the time of writing, the adenine profile was only partially complete, with some windows unfinished – most windows had simulated to 100 ns, while several remained unsimulated due to energy minimisation issues. Lack of convergence was illustrated by block analysis. The profile is given in figure 12 below.

The incompleteness of the adenine profile makes drawing any conclusions difficult. However, superimposing the profile with the cytosine profile reveals that, even though the specific energies themselves differ, several of the features appear to be preserved across profiles, shown in figure 13. As a general rule, what was observed across all profiles was that an incomplete profile develops specific features very early in simulation time, and after this point it is the absolute energy values that change rather than the feature positions. While this does not allow conclusions to be drawn, it does allow very rough comparisons between molecules.

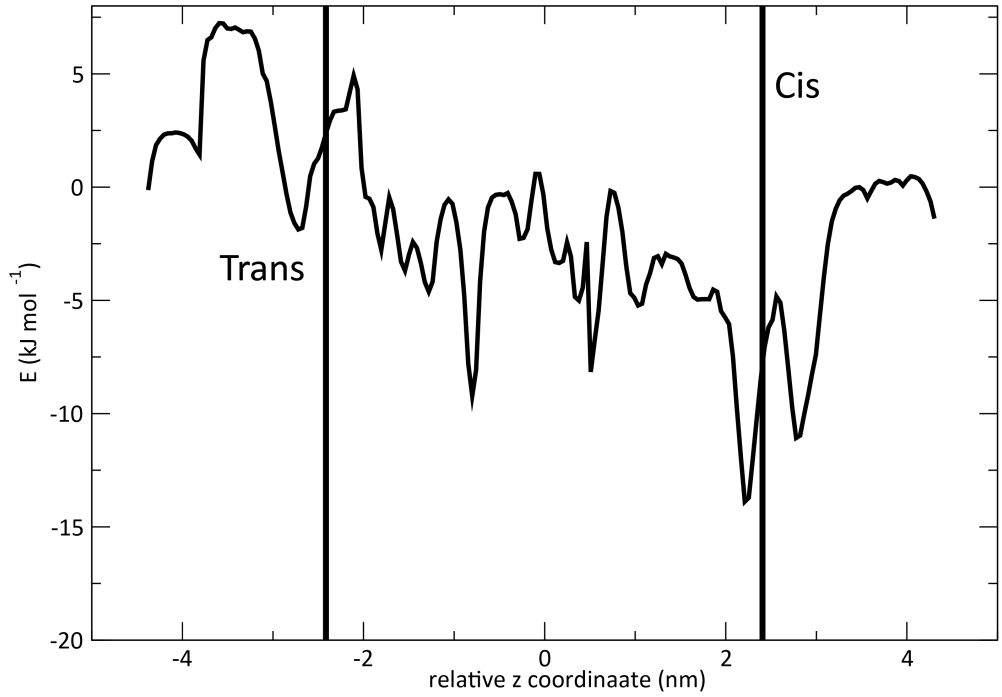

Figure 12. Incomplete adenine profile.

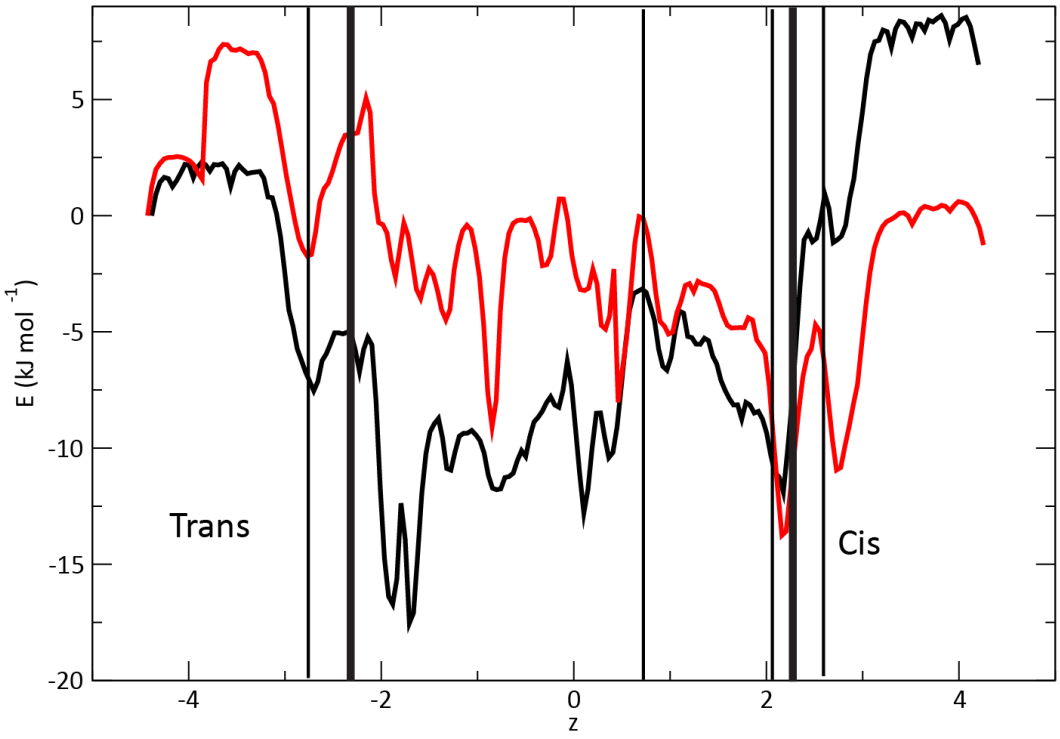

Figure 13. Superimposed cytosine (black) and adenine (red) profiles with particularly conserved features marked as thin lines.

Among others, features at z = -2.75 nm, z = 0.8 nm, z = 2 nm and z = 2.5 nm appear to be conserved, as well as other general superficial appearances of similarity at other points. These features correspond to the base of the pore at D127, the lower end of the threonine-rich region, and the constriction site formed by K147/E111 respectively. Both bases show energies within 10

kJ mol^{-1} of each other, with some minor exceptions of differences between large peaks and/or troughs, with the differences arising likely as a lack of sampling on the part of the adenine profile. The region between z = +0.8 and z = 2.5 shows good agreement between the profiles, with energies generally within 3 kJ mol^{-1} between the two.

6.4 Discussion and conclusions

The observed features within the nanopore can be broadly divided into two groups: the features observed with phosphate, and those observed in the hydrophobic bases. The phosphate profile reveals barriers particularly at hydrophobic regions such as that near L135, and a smaller barrier at M113. There are also minor barriers at sites with negatively charged residues, as is to be expected. Particularly favourable interactions appear to occur particularly in regions with many polar groups, such as near N123, where several sidechains can interact with the phosphate group at any one time. Care should be taken, however, in interpreting this data as it could be a consequence of the phosphate group interacting within a 'pocket' formed by the sidechains (see figure 2, above), and is therefore only valid for free but not backbone phosphates due to size constraints. However, due to the nature of the interactions observed, it is likely that this interaction will remain favourable in all cases.

The bases as a whole find the pore environment more favourable than bulk solvent with an energy difference of around 10 kJ mol^{-1}, which is a consequence of them being more hydrophobic in character and therefore less likely to find the bulk solvent environment suitable. The most favourable sites appear to be close to M113, a hydrophobic residue, and, interestingly, near D127/K131, both charged residues at a constriction site. As a similar site appears near the constriction formed by K117/E111, it actually may be a consequence of pore narrowing and therefore exclusion of water that causes these sites to be favourable. As there are known to be three regions with recognition sites within the barrel (*189*), two of which are in the regions including these particularly favourable sites, it could be that these specifically favourable residues play a role in the base recognition process. This may also be due to their nature as constriction sites, and thus already have a substantial impact on observed currents. The third recognition site is known to be in the central region of the barrel, which lacks a constriction site, but includes the asparagine-rich region and possibly L135, which has been shown to have a large effect on phosphate energies in the profile calculated above. The base profiles calculated thus far do not appear to show any distinguishing features within this region, with the possible exception in the cytosine profile of a change in energies of 7.5 kJ mol^{-1} within a short distance around z = 0. This is close to the favourable binding pocket observed for phosphate, which exists around z = -0.2;

while this is far from conclusive, it may suggest some basis for this recognition site, being linked to the presence of the large number of asparagine residues around this point. The ability of all bases to form a substantial number of hydrogen bonds, and this area containing many hydrogen bond donors/acceptors, could be part of this recognition process. As this site is known to be slightly less effective at recognising differences between bases (*17*), it may be that its effects are more subtle and less easy to distinguish in a profile such as those above.

Compared to the phosphate PMF profile in OprP reported in (*188*), the calculated profile for phosphate displayed above shows opposite behaviour in that the pore environment is generally less favourable than bulk solvent, particularly around hydrophobic residues. However, the favourability is increased in areas with large quantities of polar residues, but not to any level that would be considered largely favourable. The regions within OprP that show the most favourability show a free energy of approximately -17 kJ mol^{-1}, compared to the most favourable energy observed here of around +1.5 kJ mol^{-1}. Even the somewhat less favourable regions within OprP still show free energies of anywhere between approximately 0 and -12 kJ mol^{-1}, while the most unfavourable site within α-HL for phosphate has an energy of +12 kJ mol^{-1}. While α-HL is known to be slightly anion selective (*16*), this appears to only have been measured for the monovalent chloride anion. The large differences in energies between the proteins is most likely accounted for by the differences in positive residue distribution between the two, with OprP including specific 'ladders' of residues that carry the phosphate, while the α-HL barrel only includes polar residues within the lumen and one ring of positive residues at each exit. The phosphate used above is also the -1-charge variant, instead of the -3 charge variant used in the OprP simulations. A change in the phosphate charge may have an effect on the profile due to altering of the favourability of the electrostatic interactions, which may be worth further investigation.

Eventually, two or more converged base simulations would reveal the different sites within the channel that are responsible for discrimination between bases – these differences would allow identification of the key areas for manipulation that would lead to better resolution when detecting DNA. However, as the simulations of the bases remain unconverged this level of detail is not yet observable.

6.5 Future work

While the phosphate profile is essentially complete, all base profiles remain incomplete and require further simulation. The cytosine profile has been simulated for 300 ns per window but block analysis shows it to be unconverged. The adenine profile currently has 100 ns per window

and therefore is far from convergence, considering adenine is slightly larger than cytosine and in theory would require a proportionally longer simulation time due to the requirement of sampling all rotational configurations available to the base within the system. Future work would consist of extending these simulations until all profiles are converged. Further work would also consider the profiles for thymine and guanine for completeness.

Further analysis work could also be performed, particularly with regard to specific hydrogen bonding patterns between the bases and the pore. It is maybe expected that these interactions play a major role in DNA transport, being the primary source of interaction between the bases themselves and the protein nanopore.

References

1. Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W. (1972) Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein, *Nature 237*, 82-88.
2. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., and Ysebaert, M. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene, *Nature 260*, 500-507.
3. Maxam, A., and Gilbert, W. (1977) A new method for sequencing DNA, *Proc Natl Acad Sci USA 74*, 560-564.
4. Sanger, F., Nicklen, S., and Coulson, A. (1977) DNA sequencing with chain-terminating inhibitors, *Proc Natl Acad Sci USA 74*, 5463-5467.
5. Metzker, M. (2009) Sequencing technologies - the next generation, *Nat Rev Genet 11*, 31-46.
6. Branton, D., Deamer, D., Marziali, A., Bayley, H., Benner, S., Butler, T., di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S., Krstic, P., Lindsay, S., Ling, Z., Mastrangelo, C., Meller, A., Oliver, J., Pershin, Y., Ramsey, J., Riehn, R., Soni, G., Tabard-Cossa, V., Wanunu, M., Wiggin, M., and Schloss, J. (2008) The potential and challenges of nanopore sequencing, *Nat Biotechnol 26*, 1146-1153.
7. Kang, X.-F., Cheley, S., Guan, X., and Bayley, H. (2006) Stochastic detection of enantiomers, *J Am Chem Soc 128*, 10684-10685.
8. Astier, Y., Braha, O., and Bayley, H. (2006) Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter, *J Am Chem Soc 128*, 1705-1710.
9. Mardis, E. (2006) Anticipating the $1,000 genome, *Genome Biol 7*, 112.
10. Schneider, G., Kowalczyk, S., Calado, V., Pandraud, G., Zandbergen, H., Vandersypen, M., and Dekker, C. (2010) DNA translocation through graphene nanopores, *Nano Lett 10*, 3163-3167.
11. Li, J., Stein, D., McMullan, C., Branton, D., Aziz, M., and Golovchenko, J. (2001) Ion-beam sculpting at nanometre length scales, *Nature 412*, 166-169.
12. Merchant, C., Healy, K., Wanunu, M., Ray, V., Peterman, N., Bartel, J., Fischbein, M., Venta, K., Luo, Z., Johnson, A., and Drndic, M. (2010) DNA translocation through graphene nanopores, *Nano Lett 10*, 2915-2921.
13. Dekker, C. (2007) Solid-state nanopores, *Nature Nanotechnology 2,* 209-215

14. Klingelhoefer, J., Carpenter, T., and Sansom, M. (2009) Peptide nanopores and lipid bilayers: interactions by coarse-grained molecular-dynamics simulations, *Biophys J 96*, 3519-3528.

15. Kasianowicz, J. J., Brandin, E., Branton, D., and Deamer, D. W. (1996) Characterization of individual polynucleotide molecules using a membrane channel, *Proc Natl Acad Sci USA 93*, 13770-13773.

16. Song, L., Hobaugh, M., Shustak, C., Cheley, S., Bayley, H., and Gouaux, J. (1996) Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore, *Science 274*, 1859-1866.

17. Stoddart, D., Maglia, G., Mikhailova, E., Heron, A. J., and Bayley, H. (2010) Multiple base-recognition sites in a biological nanopore: two heads are better than one, *Angew Chem Int Ed Engl 49*, 556-559.

18. Cheley, S., Malghani, M., Song, L., Hobaugh, M., Gouaux, J., Yang, J., and Bayley, H. (1997) Spontaneous oligomerization of a staphylococcal α-hemolysin conformationally constrained by removal of residues that form the transmembrane β-barrel, *Protein Eng 10*, 1433-1443.

19. Maglia, G., Rincon-Restrepo, M., Mikhailova, E., and Bayley, H. (2008) Enhanced translocation of single DNA molecules through α-hemolysin nanopores by manipulation of internal charge, *Proc Natl Acad Sci USA 105*, 19720-19725.

20. Rincon-Restrepo, M., Mikhailova, E., Bayley, H., and Maglia, G. (2011) Controlled Translocation of Individual DNA Molecules through Protein Nanopores with Engineered Molecular Brakes, *Nano Lett 11*, 746-750.

21. Jett, J., Keller, R., Martin, J., Marrone, B., Moyzis, R., Ratliff, R., Seitzinger, N., Shera, E., and Stewart, C. (1989) High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules, *J Biomol Struct Dyn 7*, 301-309.

22. Subbarao, G., and van den Berg, B. (2006) Crystal structure of the monomeric porin OmpG, *J Mol Biol 360*, 750-759.

23. Chen, M., Khalid, S., Sansom, M. S. P., and Bayley, H. (2008) Outer membrane protein G: Engineering a quiet pore for biosensing, *Proc Natl Acad Sci USA 105*, 6272-6277.

24. Simpson, A., Tao, Y., Leiman, P., Badasso, M., He, Y., Jardine, P., Olson, N., Morais, M., Grimes, S., Anderson, D., Baker, T., and Rossman, M. (2000) Structure of the bacteriophage ϕ29 DNA packaging motor, *Nature 408*, 745-750.

25. Wendell, D., Jing, P., Geng, J., and Subramaniam, V. (2009) Translocation of double-stranded DNA through membrane-adapted phi29 motor protein nanopores, *Nature Nanotechnology 4,* 765-772

26. Faller, M., Niederweis, M., and Schulz, G. (2004) The structure of a mycobacterial outer-membrane channel, *Science 303*, 1189-1192.
27. Derrington, I., Butler, T., Collins, M. D., Manrao, E., Pavlenok, M., Niederweis, M., and Gundlach, J. (2010) Nanopore DNA sequencing with MspA, *Proc Natl Acad Sci USA 107*, 16060-16065.
28. Maglia, G., Restrepo, M. R., Mikhailova, E., and Bayley, H. (2008) Enhanced translocation of single DNA molecules through {alpha}-hemolysin nanopores by manipulation of internal charge, *Proc Natl Acad Sci USA 105*, 19720-19725.
29. Pawelek, P., Croteau, N., Ng-Thow-Hing, C., Khursigara, C., Moiseeva, N., Allaire, M., and Coulton, J. (2006) Structure of TonB in complex with FhuA, *E. coli* outer membrane receptor, *Science 312*, 1399-1402.
30. Mohammad, M., Howard, K., and Movileanu, L. (2011) Redesign of a plugged β-barrel membrane protein, *J Biol Chem 286*, 8000-8013.
31. Cheatham, T., and Kollman, P. (2000) Molecular dynamics simulations of nucleic acids, *Annu Rev Phys Chem 51*, 435-471.
32. Cheatham, T., Miller, J., Fox, T., Darden, T., and Kollman, P. (1995) Molecular dynamics simulations on solvated biomolecular systems: the particle mesh Ewald method leads to stable trajectories of DNA, RNA and proteins, *J Am Chem Soc 117*, 4193-4194.
33. Tidor, B., Irikura, K., Brooks, B., and Karplus, M. (1983) Dynamics of DNA oligomers, *J Biomol Struct Dyn 1*, 231-252.
34. Levitt, M. (1983) Computer simulation of DNA double-helix dynamics, *Cold Spring Harb Symp Quant Biol 47*, 251-262.
35. Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems, *J Chem Phys 98*, 10089-10092.
36. Ricci, C., de Andrade, A., Mottin, M., and Netz, P. (2010) Molecular dynamics of DNA: comparison of force fields and terminal nucleotide definitions, *J Phys Chem B 114*, 9882-9893.
37. Beveridge, D., Barreiro, G., Byun, K., Case, D., Cheatham, T., Dixit, S., Giudice, E., Lankas, F., Lavery, R., Maddocks, J., Osman, R., Seibert, E., Sklenar, H., Stoll, G., Thayer, K., Varnai, P., and Young, M. (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps, *Biophys J 87*, 3799-3813.
38. Dixit, S., Beveridge, D., Case, D., Cheatham, T., Giudice, E., Lankas, F., Lavery, R., Maddocks, J., Osman, R., Sklenar, H., Thayer, K., and Varnai, P. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II:

Sequences context effects on the dynamical structures of the 10 unique dinucleotide steps, *Biophys J 89*, 3721-3740.

39. Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T., Case, D., Cheatham, T., Dixit, S., Jayaram, B., Lankas, F., Laughton, C., Maddocks, J., Michon, A., Osman, R., Orozco, M., Perez, A., Singh, T., Spackova, N., and Sponer, J. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA, *Nucl Acids Res 38*, 299-313.
40. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T., Laughton, C., and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers, *Biophys J 92*, 3817-3829.
41. Pérez, A., Lankas, F., Luque, F., and Orozco, M. (2008) Towards a molecular dynamics consensus view of B-DNA flexibility, *Nucl Acids Res 36*, 2379-2394.
42. Foloppe, N., and MacKerell, A. (2000) All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data, *J Comp Chem 21*, 86-104.
43. Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J Am Chem Soc 117*, 5179-5197.
44. Wang, J., Cieplak, P., and Kollman, P. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?, *J Comp Chem 21*, 1049-1074.
45. Seibel, G., Singh, U., and Kollman, P. (1985) A molecular dynamics simulation of double-helical B-DNA including counterions and water, *Proc Natl Acad Sci USA 82*, 6537-6540.
46. Weiner, S., Kollman, P., Case, D., Singh, U., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins, *J Am Chem Soc 106*, 765-784.
47. Cheatham, T., Cieplak, P., and Kollman, P. (1999) A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat, *J Biomol Struct Dyn 16*, 845-862.
48. Soares, T., Hunenberger, P., Kastenholz, M., Krautler, V., Lenz, T., Lins, R., Oostenbrink, C., and van Gunsteren, W. (2004) An improved nucleic acid parameter set for the GROMOS force field, *J Comp Chem 26*, 725-737.
49. Harris, S., Laughton, C., and Liverpool, T. (2008) Mapping the phase diagram of the writhe of DNA nanocircles using atomistic molecular dynamics simulations, *Nucl Acids Res 36*, 21-29.

50. Mitchell, J., Laughton, C., and Harris, S. (2011) Atomistic simulations reveal bubbles, kinks and wrinkles in supercoiled DNA, *Nucl Acids Res 39*, 3928-3938.
51. Simonsson, T. (2001) G-quadruplex DNA structures - variations on a theme, *Biol Chem 382*, 621-628.
52. Haider, S., Parkinson, G., and Neidle, S. (2003) Structure of a G-quadruplex-ligand complex, *J Mol Biol 326*, 117-125.
53. Haider, S., Parkinson, G., and Neidle, S. (2008) Molecular dynamics and principal components analysis of human telomeric quadruplex multimers, *Biophys J 95*, 296-311.
54. Haider, S., and Neidle, S. (2009) A molecular model for drug binding to tandem repeats of telomeric G-quadruplexes, *Biochem Soc Trans 37*, 583-588.
55. Haider, S., and Neidle, S. (2010) Molecular modeling and simulation of g-quadruplexes and quadruplex-ligand complexes, *Methods Mol Biol 608*, 17-37.
56. Akhshi, P., Mosey, N., and Wu, G. (2012) Free-energy landscapes of ion movement through a g-quadruplex DNA channel, *Angew Chem 124*.
57. Wheatley, E., Pieniazek, S., Mukerji, I., and Beveridge, D. (2012) Molecular dynamics of a DNA Holliday junction: the inverted repeat sequence d(CCGGTACCG)4, *Biophys J 102*, 552-560.
58. Luger, K., Mader, A., Richmond, R., Sargent, D., and Richmond, T. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution, *Nature 389*, 251-260.
59. de Vlieg, J., Berendsen, H., and van Gunsteren, W. (1989) An NMR-based molecular dynamics simulation of the interaction of the *lac* repressor headpiece and its operator in aqueous solution, *Proteins 6*, 104-127.
60. MacKerell, A., and Nilsson, L. (2008) Molecular dynamics simulations of nucleic acid-protein complexes, *Curr Opin Struct Biol 18*, 194-199.
61. Zakrzewska, K., Bouvier, B., Michon, A., Blanchet, C., and Lavery, R. (2009) Protein-DNA binding specificity: a grid-enabled compuatational approach applied to single and multiple protein assemblies, *Phys Chem Chem Phys 11*, 10712-10721.
62. Schwabe, J. (1997) The role of water in protein-DNA interactions, *Curr Opin Struct Biol 7*, 126-134.
63. Reddy, C., Das, A., and Jayaram, B. (2001) Do water molecules mediate protein-DNA recognition?, *J Mol Biol 314*, 619-632.
64. Fuxreiter, M., Mezei, M., Simon, I., and Osman, R. (2005) Interfacial water as a "hydration fingerprint" in the noncognate complex of *Bam*HI, *Biophys J 89*, 903-911.
65. Duan, J., and Nilsson, L. (2006) Effect of Zn2+ on recognition and stability of the p53 DNA-binding domain, *Biochemistry 45*, 7483-7492.

66. Zou, X., Ma, W., Solov'yov, I., Chipot, C., and Schulten, K. (2011) Recognition of methylated DNA through methyl-CpG binding domain proteins, *Nucl Acids Res 40,* 2747-2758

67. Kollman, P. (1993) Free energy calculations: applications to chemical and biochemical phenomena, *Chem Rev 93*, 2395-2417.

68. Ajay, and Murcko, M. (1995) Computational methods to predict binding free energy in ligand-receptor complexes, *J Med Chem 38*, 4953-4967.

69. Ettig, R., Kepper, N., Stehr, R., Wedemann, G., and Rippe, K. (2011) Dissecting DNA-histone interactions in the nucleosome by molecular dynamics simulations of DNA unwrapping, *Biophys J 101*, 1999-2008.

70. Harris, S., Gavathiotis, E., Searle, M., Orozco, M., and Laughton, C. (2001) Cooperativity in drug-DNA recognition: a molecular dynamics study, *J Am Chem Soc 123*, 12658-12663.

71. Spiegel, K., Rothlisberger, U., and Carloni, P. (2004) Cisplatin binding to DNA oligomers from hybrid Car-Parrinello/molecular dynamics simulations, *J Phys Chem B 108*, 2699-2707.

72. Hannon, M. (2007) Supramolecular DNA recognition, *Chem Soc Rev 36*, 280-295.

73. Bostock-Smith, C., Laughton, C., and Searle, M. (1998) DNA minor groove recognition by a tetrahydropyrimidinium analogue of hoechst 33258: NMR and molecular dynamics studies of the complex with d(GGTAATTACC)2, *Nucl Acids Res 26*, 1660-1667.

74. Bostock-Smith, C., Harris, S., Laughton, C., and Searle, M. (2001) Induced fit DNA recognition by a minor groove binding analogue of Hoechst 33258: fluctuations in DNA A tract structure investigated by NMR and molecular dynamics simulations, *Nucl Acids Res 29*, 693-702.

75. Spitzer, G., Wellenzohn, B., Laggner, C., Langer, T., and Liedl, K. (2007) DNA minor groove pharmacophores describing sequence specific properties, *J Chem Inf Model 47*, 1580-1589.

76. Wang, H., and Laughton, C. (2009) Evaluation of molecular modelling methods to predict the sequence-selectivity of DNA minor groove binding ligands, *Phys Chem Chem Phys 11*, 10722-10728.

77. Nguyen, B., Neidle, S., and Wilson, W. (2008) A role for water molecules in DNA-ligand minor groove recognition, *Acc Chem Res 42*, 11-21.

78. Meistermann, I., Moreno, V., Prieto, M., Moldrheim, E., Sletten, E., Khalid, S., Rodger, P., Peberdy, J., Isaac, C., Rodger, A., and Hannon, M. (2002) Intramolecular DNA coiling mediated by metallo-supramolecular cylinders: differential binding of P and M helical enantiomers, *Proc Natl Acad Sci USA 99*, 5069-5074.

79. Khalid, S., Hannon, M., Rodger, A., and Rodger, P. (2006) Simulations of DNA coiling around a synthetic supramolecular cylinder that binds the DNA major groove, *Chem-Eur J 12*, 3493-3506.

80. Khalid, S., Hannon, M., Rodger, A., and Rodger, P. (2006) Shape effects on the activity of synthetic major-groove binding ligands, *J Molec Graphic Model 25*, 794-800.

81. Dolenc, J., Oostenbrink, C., Koller, J., and van Gunsteren, W. (2005) Molecular dynamics simulations and free energy calculations of netropsin and distamycin binding to an AAAAA DNA binding site, *Nucl Acids Res 33*, 725-733.

82. Dai, L., Mu, Y., Nordensklöld, L., and van der Maarel, J. (2008) Molecular dynamics simulation of multivalent-ion mediated attraction between DNA molecules, *Phys Rev Lett 100*, 118301.

83. Ricci, C., and Netz, P. (2009) Docking studies on DNA-ligand interactions: building and application of a protocol to identify the binding mode, *J Chem Inf Model 49*, 1925-1935.

84. Read, M., Harrison, R., Romagnoli, B., Tanious, F., Gowan, S., Reszka, A., Wilson, W., Kelland, L., and Neidle, S. (2001) Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors, *Proc Natl Acad Sci USA 98*, 4844-4849.

85. Murphy, M., Rasnik, I., Cheng, W., Lohman, T., and Ha, T. (2004) Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy, *Biophys J 86*, 2530-2537.

86. Oostenbrink, C., Villa, A., Mark, A., and van Gunsteren, W. (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6, *J Comp Chem 25*, 1656-1676.

87. Martínez, J., Elmroth, S., and Kloo, L. (2001) Influence of sodium ions on the dynamics and structure of single-stranded DNA oligomers: a molecular dynamics study, *J Am Chem Soc 123*, 12279-12289.

88. Sen, S., and Nilsson, L. (2001) MD simulations of homomorphous PNA, DNA, and RNA single strands: characterization and comparison of conformations and dynamics, *J Am Chem Soc 123*, 7414-7422.

89. Laio, A., and Parrinello, M. (2002) Escaping free-energy minima, *Proc Natl Acad Sci USA 99*, 12562-12566.

90. Piana, S. (2007) Atomistic simulation of the DNA helix-coil transition, *J Phys Chem A 111*, 12349-12354.

91. Yu, J., Ha, T., and Schulten, K. (2006) Structure-based model of the stepping motor of PcrA helicase, *Biophys J 91*, 2097-2114.

92. Broderick, S., Rehmet, K., Concannon, C., and Nasheuer, H. (2010) Eukaryotic single-stranded DNA binding proteins: central factors in genome stability, *Subcell Biochem 50*, 143-163.

93. Folmer, R., Nilges, M., Folkers, P., Konings, R., and Hilbers, C. (1994) A model of the complex between single-stranded DNA and the single-stranded DNA binding protein encoded by gene V of filamentous bacteriophage M13, *J Mol Biol 240*, 341-357.

94. Folmer, R., Nilges, M., Papavoine, C., Harmsen, B., Konings, R., and Hilbers, C. (1997) Refined structure, DNA binding studies, and dynamics of the bacteriophage Pf3 encoded single-stranded DNA binding protein, *Biochemistry 36*, 9120-9135.

95. Wallace, E., and Sansom, M. (2009) Carbon nanotube self-assembly with lipids and detergent: a molecular dynamics study, *Nanotechnology 20*, 045101.

96. Zheng, M., Jagota, A., Semke, E., Diner, B., McLean, R., Lustig, S., Richardson, R., and Tassi, N. (2003) DNA-assisted dispersion and separation of carbon nanotubes, *Nat Mater 2*, 338-342.

97. Zheng, M., Jagota, A., Strano, M., Santos, A., Barone, P., Chou, S., Diner, B., Dresselhaus, M., Mclean, R., Onoa, G., Samsonidze, G., Semke, E., Usrey, M., and Walls, D. (2003) Structure-based carbon nanotube sorting by sequence-dependent DNA assembly, *Science 302*, 1545-1548.

98. Johnson, R., Johnson, A., and Klein, M. (2008) Probing the structure of DNA-carbon nanotube hybrids with molecular dynamics, *Nano Lett 8*, 69-75.

99. Frischknecht, A., and Martin, M. (2008) Simulation of the adsorption of nucleotide monophosphates on carbon nanotubes in aqueous solution, *J Phys Chem C 112*, 6271-6278.

100. Sugita, Y., and Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding, *Chem Phys Lett 314*, 141-151.

101. Johnson, R., Kohlmeyer, A., Johnson, A., and Klein, M. (2009) Free energy landscape of a DNA-carbon nanotube hybrid using replica exchange molecular dynamics, *Nano Lett 9*, 537-541.

102. Gao, H., Kong, Y., Cui, D., and Ozkan, C. (2003) Spontaneous insertion of DNA oligonucleotides into carbon nanotubes, *Nano Lett 3*, 471-473.

103. Gao, H., and Kong, Y. (2004) Simulation of DNA-nanotube interactions, *Annu Rev Mater Res 2004*, 123-150.

104. Wadkins, R., Vladu, B., and Tung, C.-S. (1998) Actinomycin D binds to metastable hairpins in single-stranded DNA, *Biochemistry 37*, 11915-11923.

105. Rahman, K., James, C., Bui, T., Drake, A., and Thurston, D. (2011) Observation of a single-stranded DNA/pyrrolobenzodiazepine adduct, *J Am Chem Soc 133*, 19376-19385.

106. Zhao, X., Payne, C., Cummings, P., and Lee, J. (2007) Single-strand DNA molecule translocation through nanoelectrode gaps, *Nanotechnology 18*, 424018.

107. Zhao, X., Payne, C. M., and Cummings, P. T. (2008) Controlled translocation of DNA segments through nanoelectrode gaps from molecular dynamics, *J Phys Chem C 112*, 8-12.

108. Payne, C. M., Zhao, X., Vlcek, L., and Cummings, P. T. (2008) Molecular dynamics simulation of ss-DNA translocation between copper nanoelectrodes incorporating electrode charge dynamics, *J Phys Chem B 112*, 1712-1717.

109. Payne, C. M., Zhao, X., and Cummings, P. T. (2008) Electrophoresis of ssDNA through nanoelectrode gaps from molecular dynamics: impact of gap width and chain length, *J Phys Chem B 112*, 12851-12858.

110. Rappe, A., Casewit, C., Colwell, K., and Goddard III, W. (1992) UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulation, *J Am Chem Soc*

111. Heng, J., Aksimentiev, A., Ho, C., Marks, P., Grinkova, Y., Sligar, S., Schulten, K., and Timp, G. (2005) Stretching DNA using the electric field in a synthetic nanopore, *Nano Lett 5*, 1883-1888.

112. Heng, J., Aksimentiev, A., Ho, C., Marks, P., Grinkova, Y., Sligar, S., Schulten, K., and Timp, G. (2006) The electromechanics of DNA in a synthetic nanopore, *Biophys J 90*, 1098-1106.

113. Gracheva, M., Xiong, A., Aksimentiev, A., Schulten, K., Timp, G., and Leburton, J.-P. (2006) Simulation of the electric response of DNA translocation through a semiconductor nanopore-capacitor, *Nanotechnology 17*, 622-633.

114. Shi, X., Kong, Y., and Zhao, Y. (2005) Molecular dynamics simulation of peeling a DNA molecule on substrate, *Acta Mech Sinica 21*, 249-259.

115. Purnell, R., Mehta, K., and Schmidt, J. (2008) Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore, *Nano Lett 8*, 3029-3034.

116. Mathé, J., Aksimentiev, A., Nelson, D., Schulten, K., and Meller, A. (2005) Orientation discrimination of single-stranded DNA inside the α-hemolysin membrane channel, *Proc Natl Acad Sci USA 102*, 12377-12382.

117. Wells, D., Abramkina, V., and Aksimentiev, A. (2007) Exploring transmembrane transport through α-hemolysin with grid-steered molecular dynamics, *J Chem Phys 127*, 125101-125111.

118. Schrodinger, E. (1926) An undulatory theory of the mechanics of atoms and molecules, *Phys Rev 28*, 1049-1070.

119. Hine, N., Haynes, P., Mostofi, A., Skylaris, C.-K., and Payne, M. (2009) Linear-scaling density-functional theory with tens of thousands of atoms: expanding the scope and scale of calculations with ONETEP, *Comp Phys Comm 180*, 1041-1053.

120. Roszak, S., Keegstra, P., Hariharan, P., and Kaufman, J. (1988) Ab-initio MRD-CI calculations for breaking a chemical bond in a molecule in a crystal or other solid environment I. H3C-NO2 decomposition in nitromethane, *Int J Quant Chem 34*, 619-653.

121. Hess, B., Bekker, H., Berendsen, H., and Fraaije, J. (1997) LINCS: A linear constraint solver for molecular simulations, *Journal of Computational Chemistry*.

122. Swope, W., Andersen, H., Berens, P., and Wilson, K. (1982) A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters, *J Chem Phys 76*, 637.

123. Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation, *J Chem Theory Comput 4*, 435-447.

124. Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A., and Berendsen, H. (2005) GROMACS: Fast, flexible, and free, *J Comp Chem 26*, 1701-1718.

125. Roache, P. (1982) *Computational fluid dynamics*, Hermosa: Albuquerque, NM.

126. Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995) A 2nd generation force-field for the simulation of proteins, nucleic acids, and organic molecules, *J Am Chem Soc 117*, 5179-5197.

127. MacKerell, A., Bashford, D., Bellott, M., Dunbrack, R., Evanseck, J., Field, M., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., K, K., Lau, F., Mattos, C., Michnick, S., Ngo, T., Nguyen, D., Prodhom, B., Reiher III, W., Roux, B., Schlenkrich, M., Smith, J., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J Phys Chem B 102*, 3586-3616.

128. Scott, W., Huenenberger, P., Tironi, I., Mark, A., Billeter, S., Fennen, J., Torda, A., Huber, T., Krueger, P., and WF, v. G. (1999) The GROMOS biomolecular simulation program package, *J Phys Chem A 103*, 3596-3607.

129. Soares, T. A., Hünenberger, P. H., Kastenholz, M. A., Kräutler, V., Lenz, T., Lins, R. D., Oostenbrink, C., and van Gunsteren, W. F. (2005) An improved nucleic acid parameter set for the GROMOS force field, *J Comp Chem 26*, 725-737.

130. Lennard-Jones, J. (1924) On the determination of molecular fields - II. From the equation of state of a gas, *Proc R Soc Lond A 106*, 463-477.

131. Onsager, L. (1936) Electric moments of molecules in liquids, *J Am Chem Soc 58*, 1486-1493.

132. Ewald, P. (1921) Die Berechnung optischer und elekrostatischer Gitterpotentiale, *Annalen der Physik 369*, 253-287.

133. Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald - an n.log(n) method for Ewald sums in large systems, *J Chem Phys 98*, 10089-10092.

134. Perera, L., Berkowitz, M., Darden, T., and Lee, H. (1995) A smooth particle mesh Ewald method, *J Chem Phys 103,* 8577-8593

135. Berendsen, H., Postma, J., van Gunsteren, W., Dinola, A., and Haak, J. (1984) Molecular dynamics with coupling to an external bath, *J Chem Phys 81*, 3684-3690.

136. Hoover, W. (1985) Canonical dynamics: equilibrium phase-space distributions, *Phys Rev A 31*, 1695-1697.

137. Nose, S. (1984) A unified formulation of the constant temperature molecular-dynamics methods, *J Chem Phys 81*, 511-519.

138. Parrinello, M., and Rahman, A. (1981) Polymorphic transitions in single crystals: a new molecular dynamics method, *J App Phys 52*, 7182-7191.

139. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs, *J Comp Chem 26*, 1668-1688.

140. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and de Vries, A. H. (2007) The MARTINI force field: Coarse grained model for biomolecular simulations, *J Phys Chem B 111*, 7812-7824.

141. Brünger, A., Clore, G., Gronenborn, A., and Karplus, M. (1986) Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: Application to crambin, *Proc Natl Acad Sci USA 83*, 3801-3805.

142. Martin, H., Jha, S., Howorka, S., and Coveney, P. (2009) Determination of Free Energy Profiles for the Translocation of Polynucleotides through α …, *J Chem Theory Comput 5*, 2135-2148.

143. Denning, E., and Woolf, T. (2008) Double bilayers and transmembrane gradients: a molecular dynamics study of a highly charged peptide, *Biophys J 95*, 3161-3173.

144. Torrie, G., and Valleau, J. (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling, *J Comp Phys 23*, 187-199.

145. Kumar, S., Rosenberg, J., Bouzida, D., Swendsen, R., and Kollman, P. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method, *J Comp Chem 13*, 1011-1021.

146. Chew, C. F., Guy, A., and Biggin, P. C. (2008) Distribution and Dynamics of Adamantanes in a Lipid Bilayer, *Biophys J 95*, 5627-5636.

147. Giudice, E., Varnai, P., and Lavery, R. (2003) Base pair opening within B-DNA: free energy pathways for GC and AT pairs from umbrella sampling simulations, *Nucl Acids Res 31*, 1434-1443.

148. Karplus, M., and McCammon, J. (2002) Molecular dynamics simulations of biomolecules, *Nat Struct Biol 9*, 646-652.

149. Aksimentiev, A., and Schulten, K. (2005) Imaging α-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the electrostatic potential map, *Biophys J 88*, 3745-3761.

150. Wells, D. B., Abramkina, V., and Aksimentiev, A. (2007) Exploring transmembrane transport through α-hemolysin with grid-steered molecular dynamics, *J Chem Phys 127*, 125101.

151. Sali, A., and Blundell, T. L. (1993) Comparative Protein Modeling by Satisfaction of Spatial Restraints, *J Mol Biol 234*, 779-815.

152. Jayawardhana, D., Crank, J., Zhao, Q., Armstrong, D., and Guan, X. (2009) Nanopore Stochastic Detection of a Liquid Explosive Component and Sensitizers Using Boromycin and an Ionic Liquid Supporting Electrolyte, *Anal Chem 81*, 460-464.

153. Bond, P. J., Guy, A. T., Heron, A. J., Bayley, H., and Khalid, S. (2011) Molecular Dynamics Simulations of DNA within a Nanopore: Arginine-Phosphate Tethering and a Binding/Sliding Mechanism for Translocation, *Biochemistry 50*, 3777-3783.

154. van Gunsteren, W. F. (1996) *The GROMOS96 Manual and User Guide*.

155. Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6, *J Comp Chem 25*, 1656-1676.

156. Dutzler, R., Wang, Y., Rizkallah, P., Rosenbusch, J., and Schirmer, T. (1996) Crystal structures of various maltooligosaccharides bound to maltoporin reveal a specific sugar translocation pathway, *Structure 4*, 127-134.

157. Ye, J., and Van Den Berg, B. (2004) Crystal structure of the bacterial nucleoside transporter Tsx, *The EMBO Journal 23*, 3187-3195.

158. Zimmerli, U., and Koumoutsakos, P. (2008) Simulations of Electrophoretic RNA Transport Through Transmembrane Carbon Nanotubes, *Biophys J 94*, 2546-2557.

159. Jeon, I., Lee, J., Andrade, J., de Gennes, P., Hermens, H., and Biochem, A. (1991) Arginine-Mediated RNA Recognition: The Arginine Fork, *Science 252*, 1167-1171.

160. Cygan, R., Liang, J.-J., and Kalinichev, A. (2004) Molecular models of hydroxide, oxyhydroxide, and clay phases and the development of a general force field, *J Phys Chem B 108*, 1255-1266.

161. Pèrez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers, *Biophys J 92*, 3817-3829.

162. Sen, S., and Nilsson, L. (2001) MD simulations of homomorphous PNA, DNA and RNA single strands: characterization and comparison of conformations and dynamics, *J Am Chem Soc 123*, 7414-7422.

163. Wang, J., CIEPLAK, P., and KOLLMAN, P. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?, *J Comp Chem 21*, 1049-1074.

164. Feller, S. E., and MacKerell, A. D. (2000) An Improved Empirical Potential Energy Function for Molecular Simulations of Phospholipids, *J Phys Chem B 104*, 7510-7515.

165. Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. (1981) Structure of a B-DNA dodecamer: conformation and dynamics, *Proc Natl Acad Sci USA 78*, 2179-2183.

166. Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossvary, I., Wong, K. F., Paesani, F., Vanicek, J., Liu, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.-J., Hornak, V., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., Kollman, P. A., and Roberts, B. P. (2010) *Amber 11*, University of California.

167. Ricci, C. G., de Andrade, A. S., Mottin, M., and Netz, P. A. (2010) Molecular dynamics of DNA: comparison of force fields and terminal nucleotide definitions, *J Phys Chem B 114*, 9882-9893.

168. Lange, O. F., van der Spoel, D., and de Groot, B. L. (2010) Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data, *Biophys J 99*, 647-655.

169. Antony, E., Weiland, E., Korolev, S., and Lohman, T. (2012) Plasmodium falciparum SSB tetramer wraps single stranded DNA with similar topology but opposite polarity to e. coli SSB, *J Mol Biol 420*, 269-283.

170. Shamoo, Y. (2001) Single-stranded DNA-binding proteins, *Encyclopedia of Life Sciences*.

171. Prusty, D., Dar, A., Priya, R., Sharma, A., Dana, A., Choudhury, N., Rao, N., and Dhar, S. (2010) Single-stranded DNA binding protein from human malarial parasite *Plasmodium falciparum* is encoded in the nucleus and targeted to the apicoplast, *Nucl Acids Res 38*.

172. Auffinger, P., Cheatham III, T., and Vaiana, A. (2007) Spontaneous formation of KCl aggregates in biomolecular simulations: a force field issue?, *J Chem Theory Comput 3*, 1851-1859.

173. Murphy, M. C., Rasnik, I., Cheng, W., Lohman, T. M., and Ha, T. (2004) Probing Single-Stranded DNA Conformational Flexibility Using Fluorescence Spectroscopy, *Biophys J 86*, 2530-2537.

174. Lu, X.-J., and Olson, W. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures, *Nucl Acids Res 31*, 5108-5121.

175. Zheng, G., Lu, X.-J., and Olson, W. (2009) Web 3DNA - a web server for the analysis, reconstruction and visualization of three-dimensional nucleic-acid structures, *Nucl Acids Res 37*, W240-W246.

176. Bjelkmar, P., Larsson, P., Cuendet, M., Hess, B., and Lindahl, E. (2010) Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models, *J Chem Theory Comput 6*, 459-466.

177. Malorov, V., and Crippen, G. (1994) Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins, *J Mol Biol 235*, 625-634.

178. Joosten, R., Te Beek, T., Krieger, E., Hekkelman, M., Hooft, R., Scheider, R., Sander, C., and Vriend, G. (2010) A series of PDB related databases for everyday needs, *Nucleic Acids Res 39*, D411-D419.

179. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers 22*, 2577-2637.

180. Matthes, D., and de Groot, B. (2009) Secondary structure propensities in peptide folding simulations: a systematic comparison of molecular mechanics interaction schemes, *Biophys J 97*, 599-608.

181. Cino, E., Choy, W.-Y., and Karttunen, M. (2012) Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations, *J Chem Theor Comput 8*, 2725-2740.

182. Rozenberg, H., Rabinovich, D., Frolow, F., Hegde, R. S., and Shakked, Z. (1998) Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets, *Proc Natl Acad Sci USA 95*, 15194-15199.

183. Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1999) Peptide Folding: When Simulation Meets Experiment, *Angew Chem Int Ed 38*, 236-240.

184. Roux, B. (1995) The calculation of the potential of mean force using computer simulations, *Comp Phys Comm 91*, 275-282.

185. Patey, G., and Valleau, J. (1975) A Monte Carlo method for obtaining the interionic potential of mean force in ionic solution, *J Chem Phys 63*, 2334-2340.

186. Allen, T., Andersen, O., and Roux, B. (2006) Molecular dynamics - potential of mean force calculations as a tool for understanding ion permeation and selectivity in narrow channels, *Biophys Chem 124*, 251-267.

187. Egwolf, B., and Roux, B. (2010) Ion selectivity of the KcsA channel: a perspective from multi-ion free energy landscapes, *J Mol Biol 401*, 831-842.

188. Pongprayoon, P., Beckstein, O., Wee, C., and Sansom, M. (2009) Simulations of anion transport through OprP reveal the molecular basis for high affinity and selectivity for phosphate, *Proc Natl Acad Sci USA 106*, 21614-21618.

189. Stoddart, D., Heron, A., Mikhailova, E., Maglia, G., and Bayley, H. (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore, *Proc Natl Acad Sci USA*.

Appendix 1. Secondary structure plots

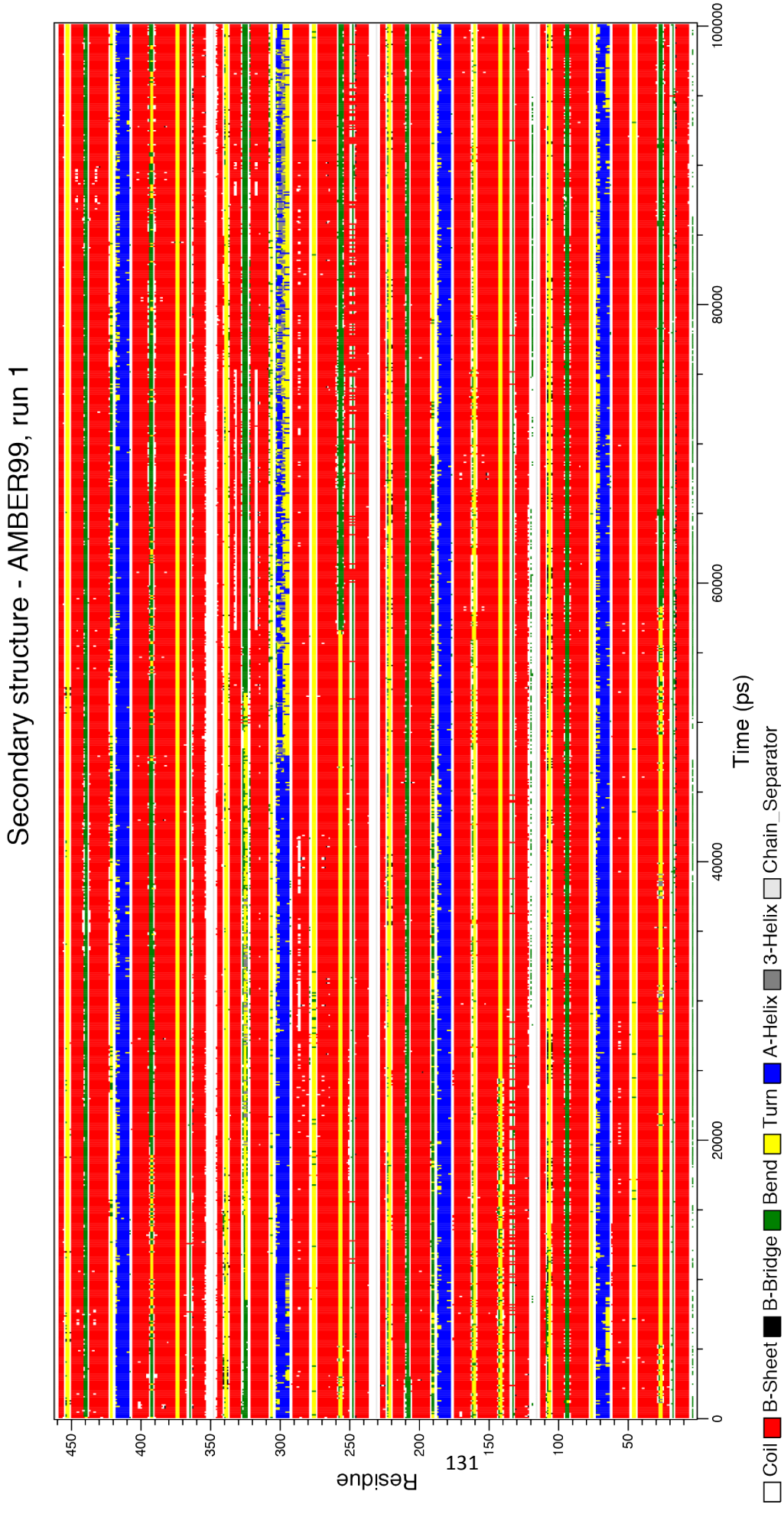

Figure 1. DSSP plot for AMBER99 force field, first simulation

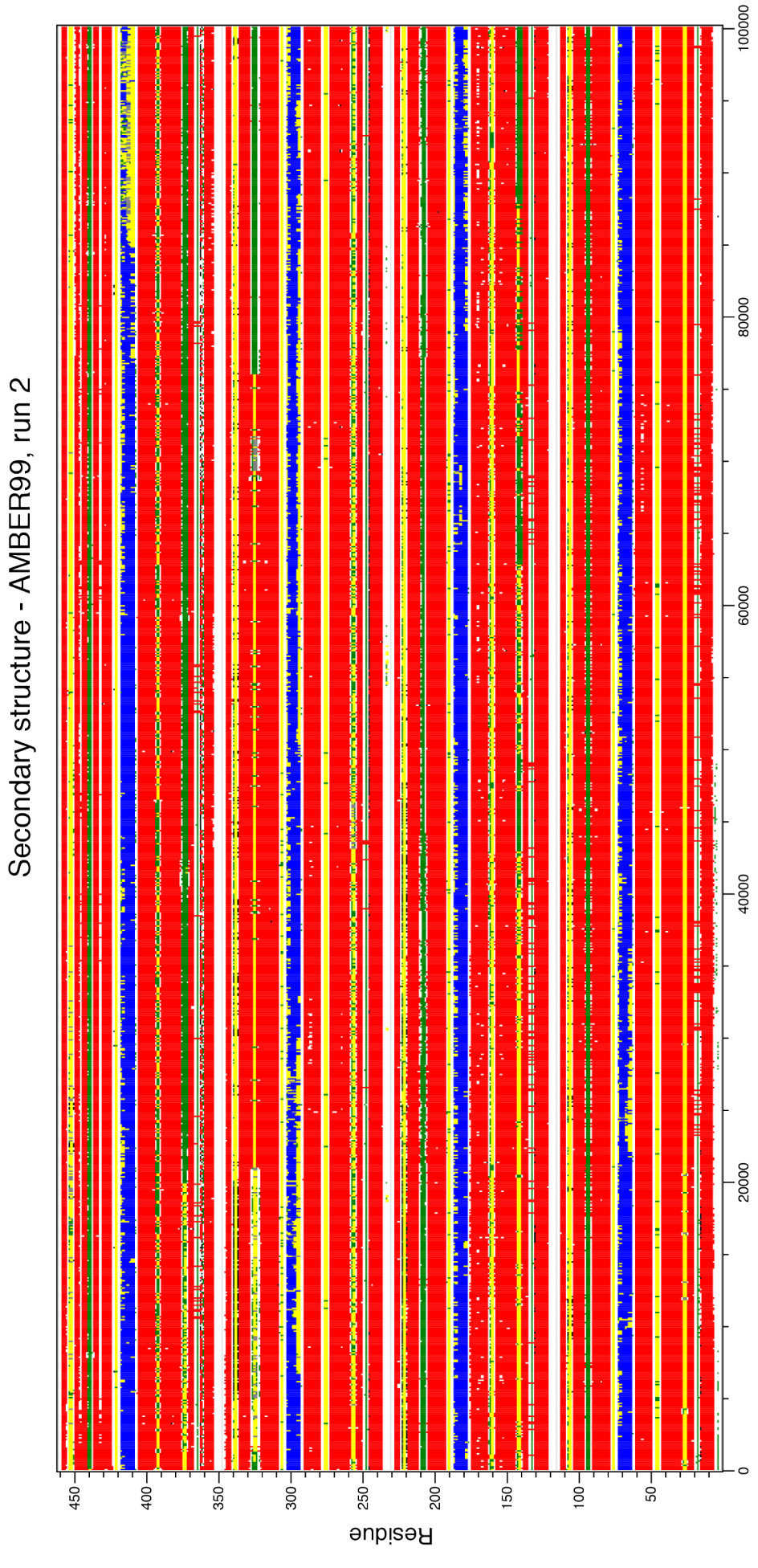

Figure 2. DSSP plot for AMBER99 force field, second simulation

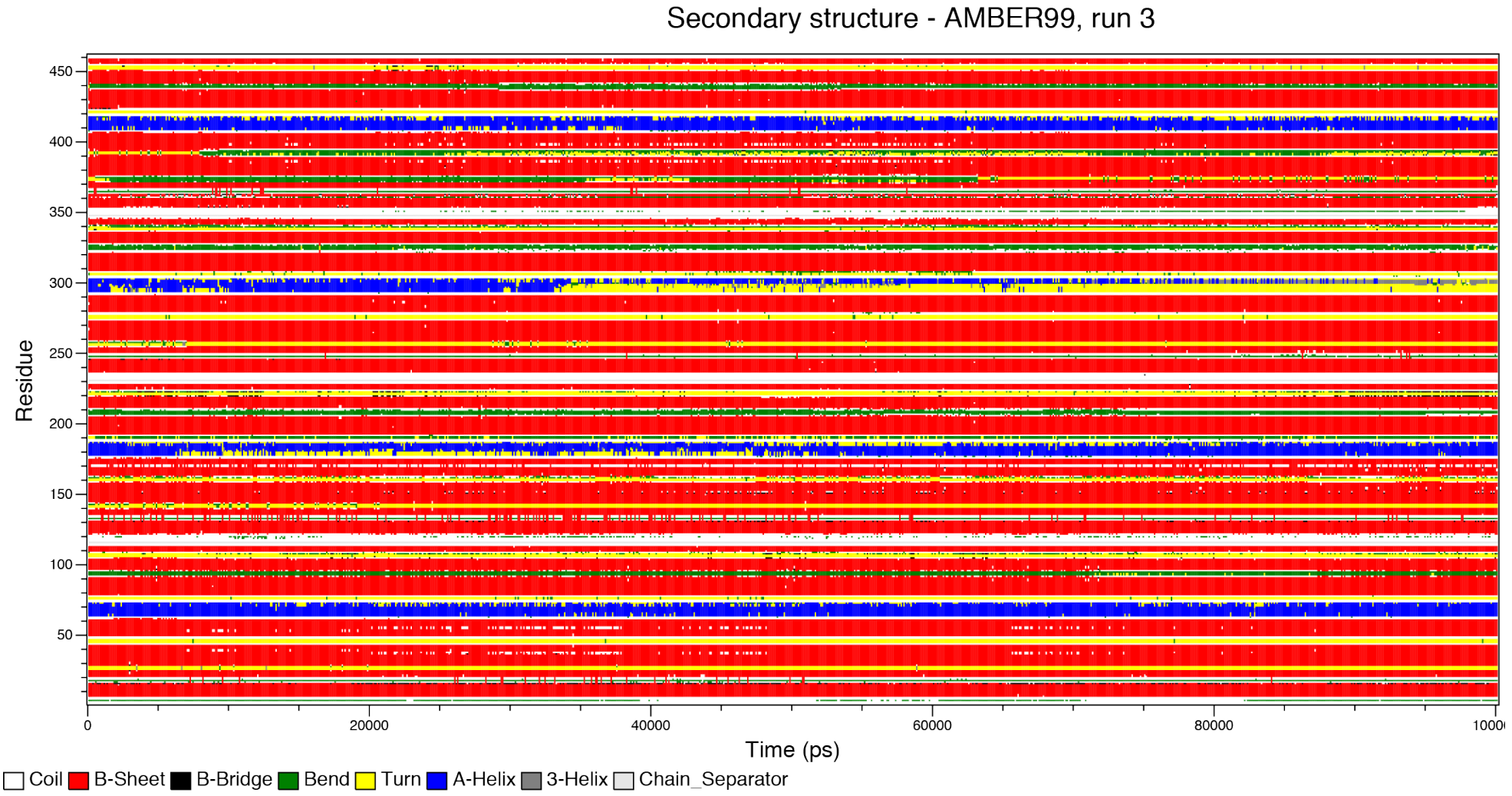

Figure 3. DSSP plot for AMBER99 force field, third simulation

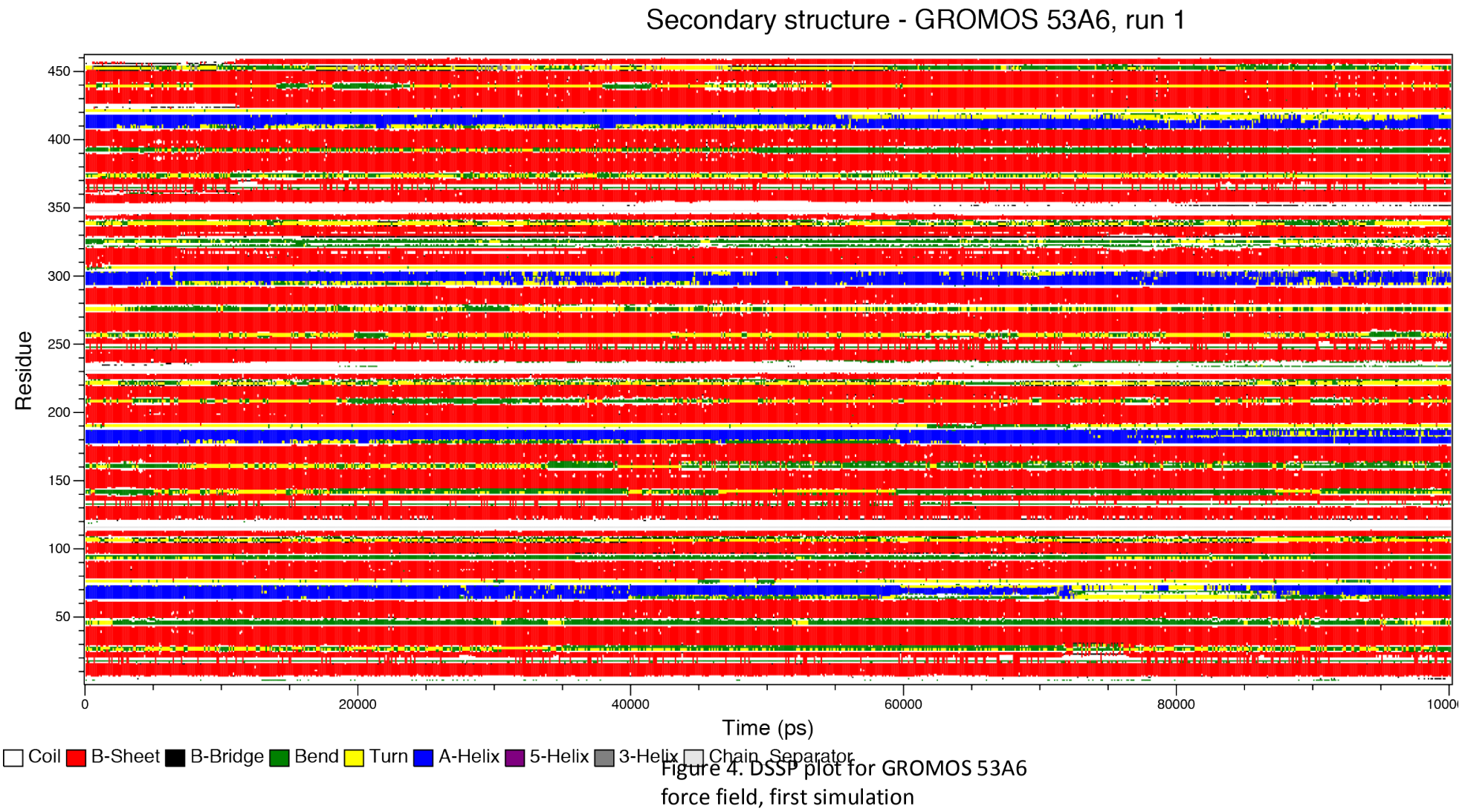

Figure 4. DSSP plot for GROMOS 53A6 force field, first simulation

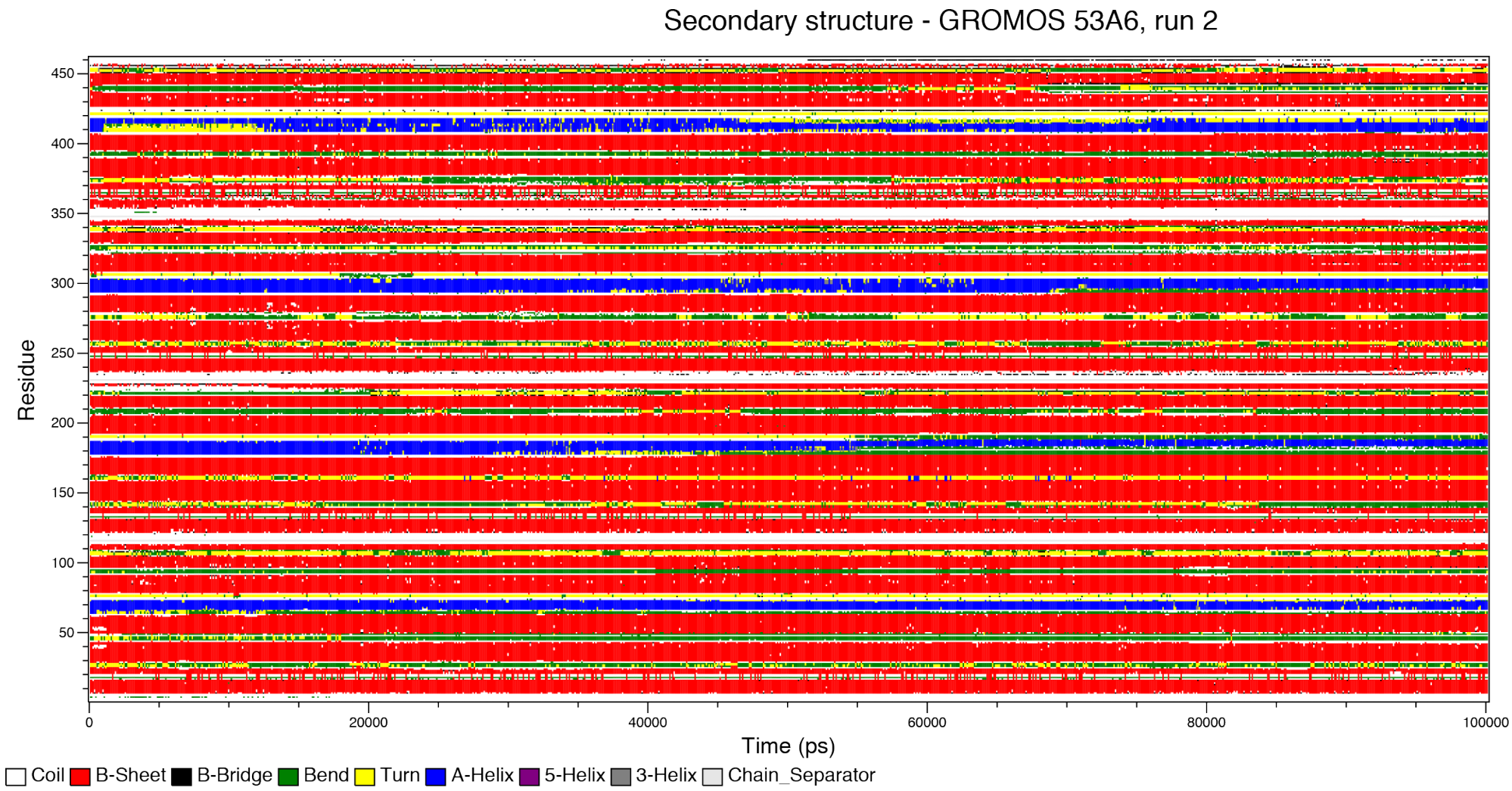

Figure 5. DSSP plot for GROMOS 53A6 force field, second simulation

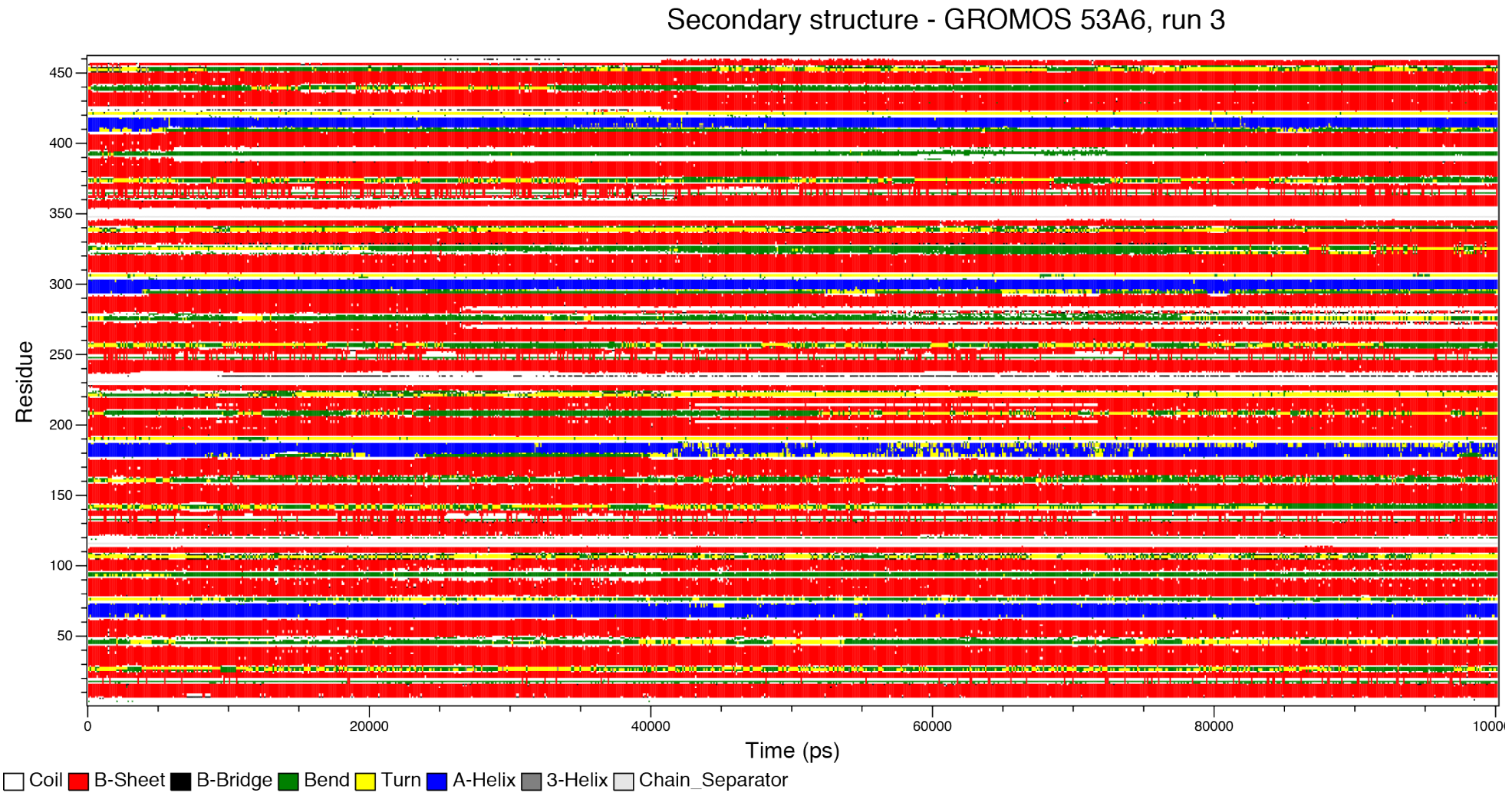

Figure 6. DSSP plot for GROMOS 53A6 force field, third simulation

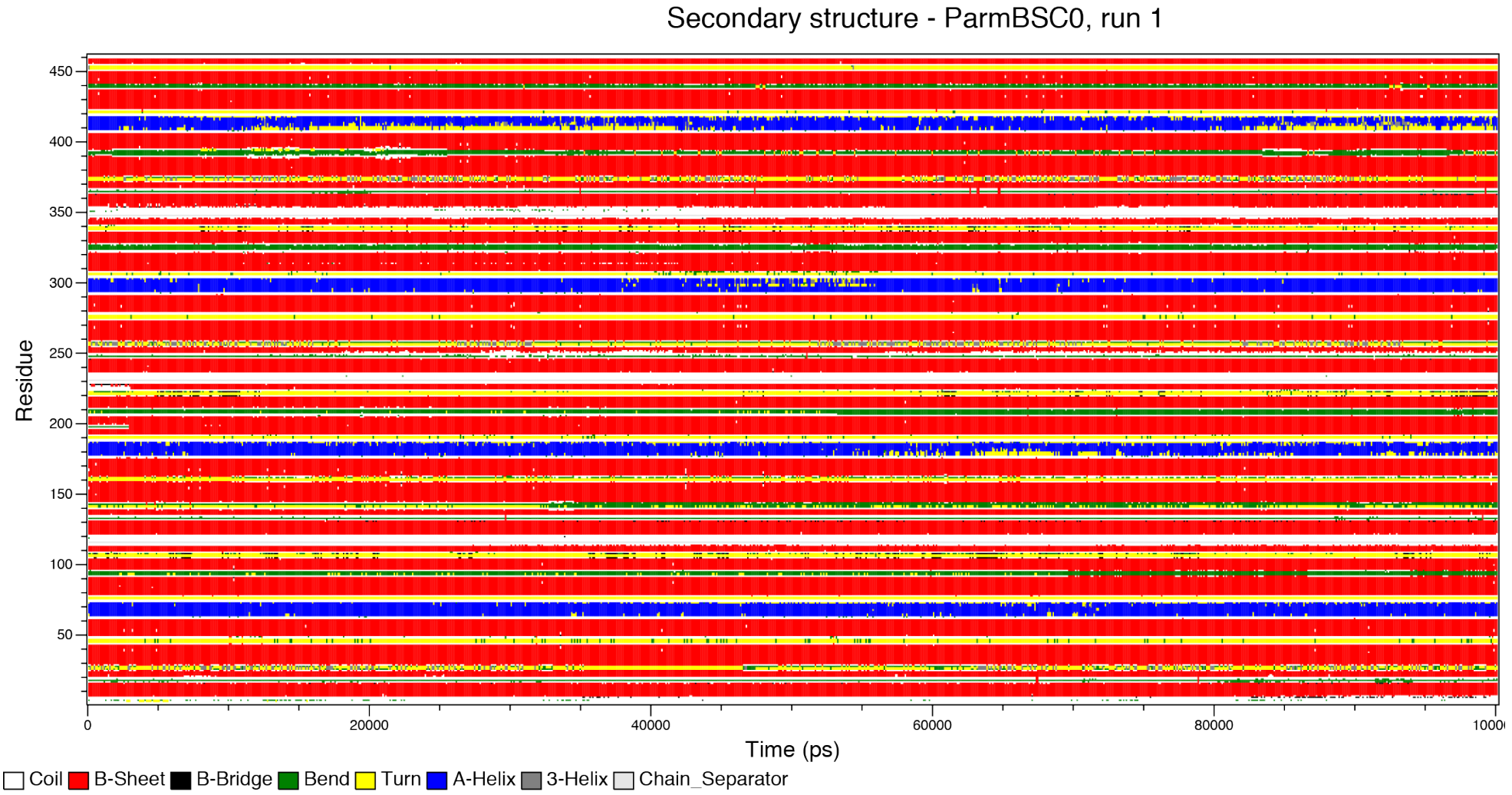

Figure 7. DSSP plot for ParmBSC0 force field, first simulation

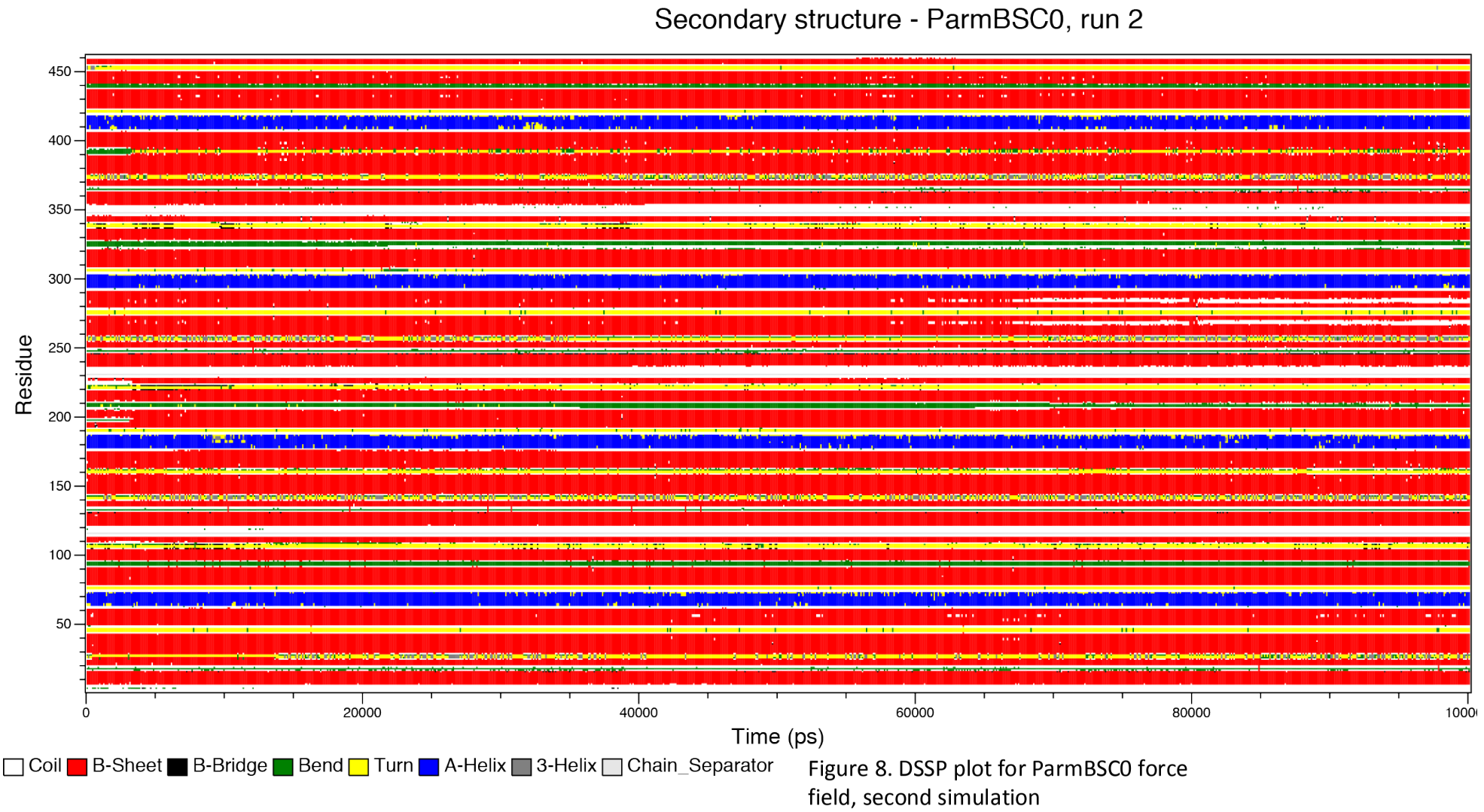

Figure 8. DSSP plot for ParmBSC0 force field, second simulation

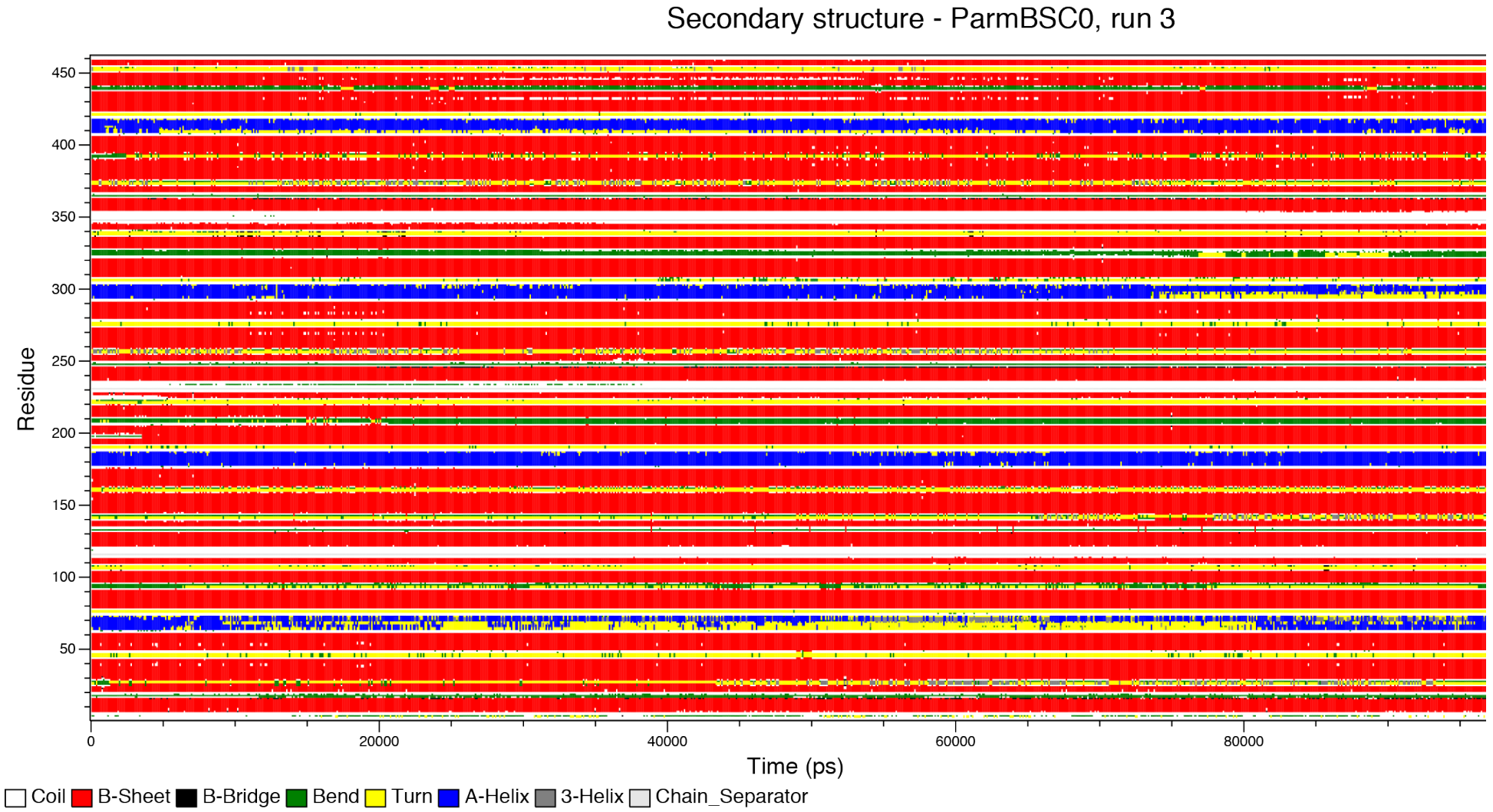

Figure 9. DSSP plot for ParmBSC0 force field, third simulation

Appendix 2: Radius of gyration calculations of ssDNA in different force fields and salt conentrations

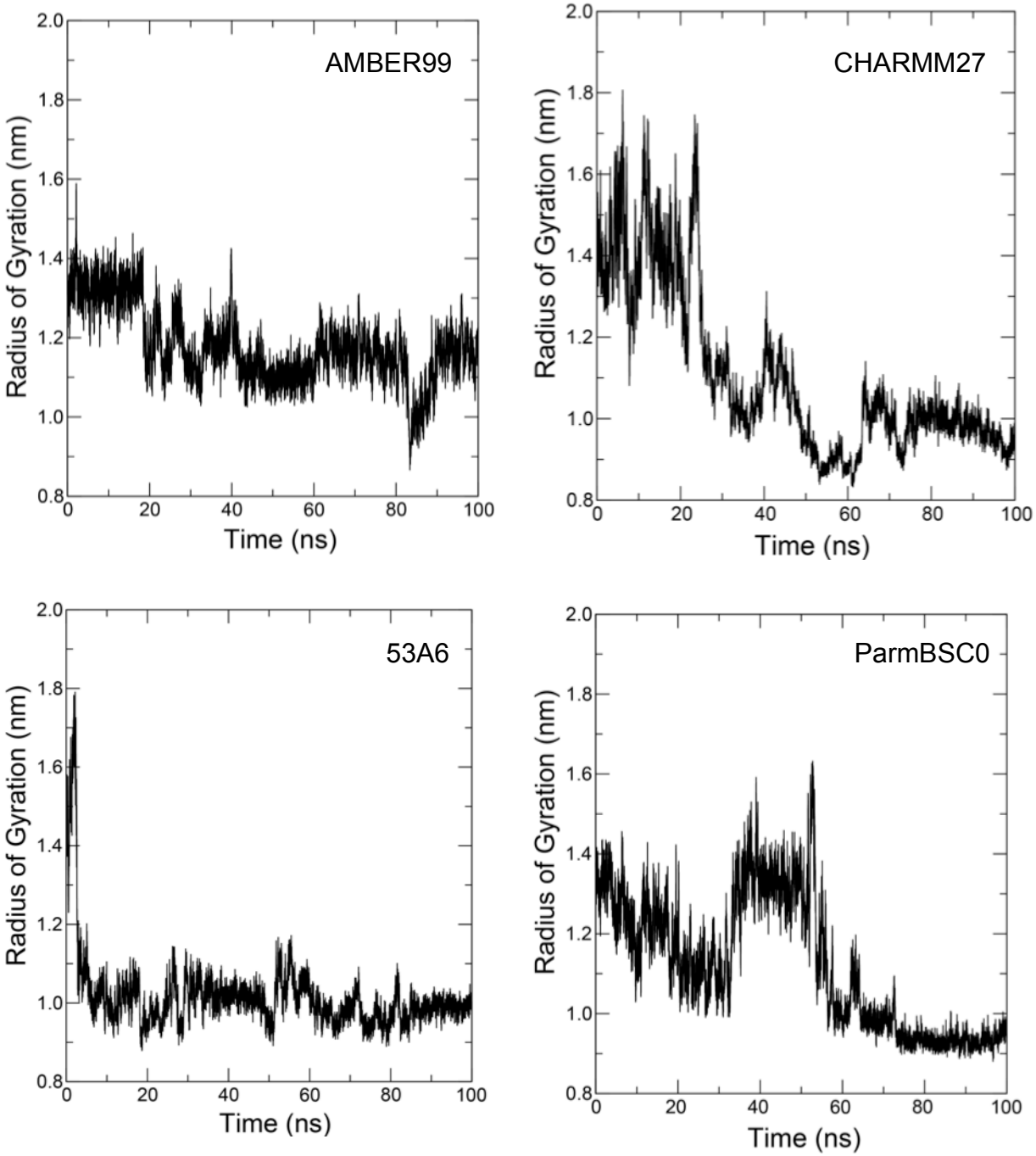

Figure 1. Radius of gyration of ssDNA across all force fields, at 0 M salt concentration. Top left: AMBER99, top right: CHARMM27, bottom left: GROMOS96 53A6, bottom right: ParmBSC0

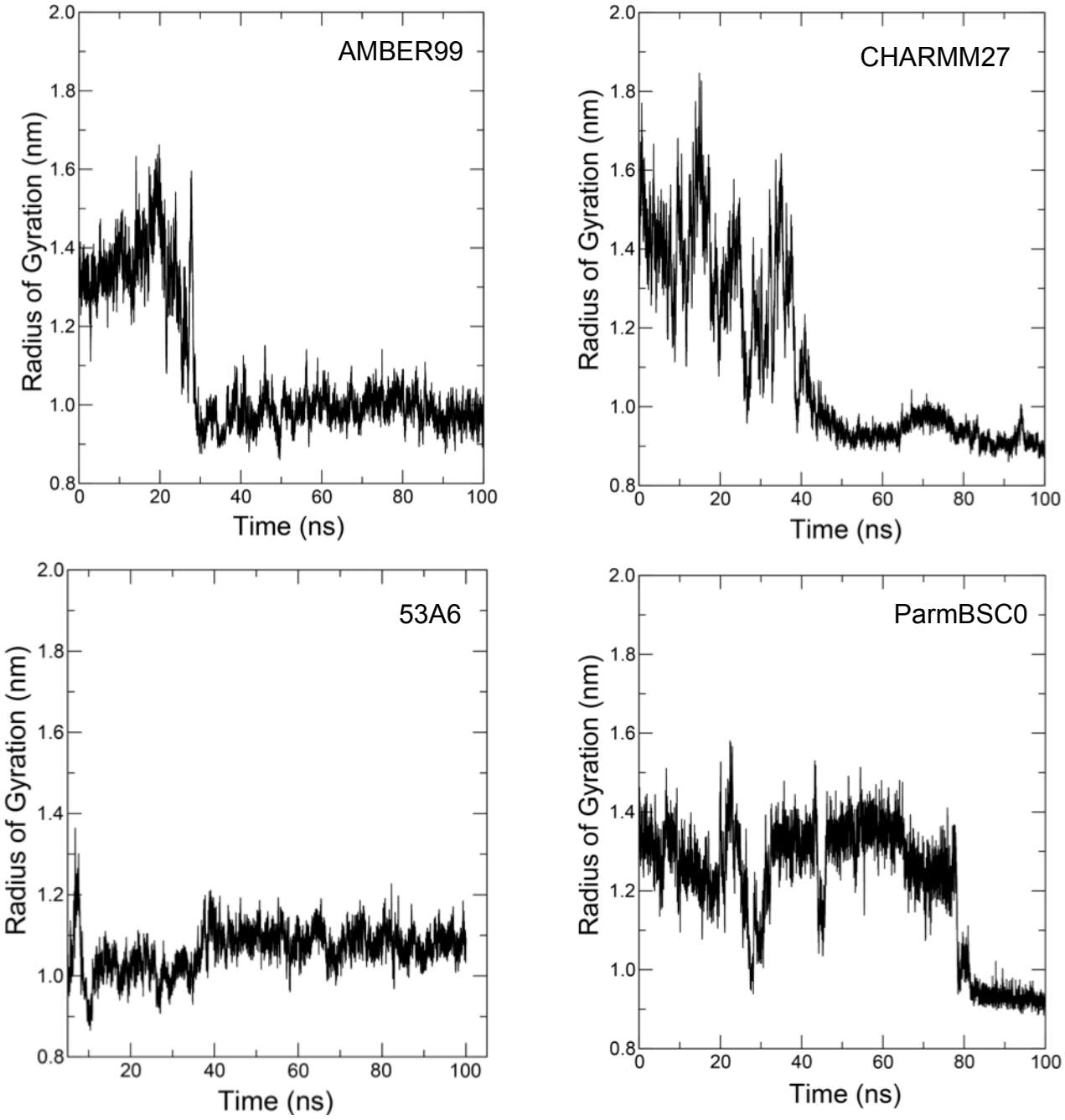

Figure 2. Radius of gyration of ssDNA across all force fields, at 0.1 M salt concentration. Top left: AMBER99, top right: CHARMM27, bottom left: GROMOS96 53A6, bottom right: ParmBSC0

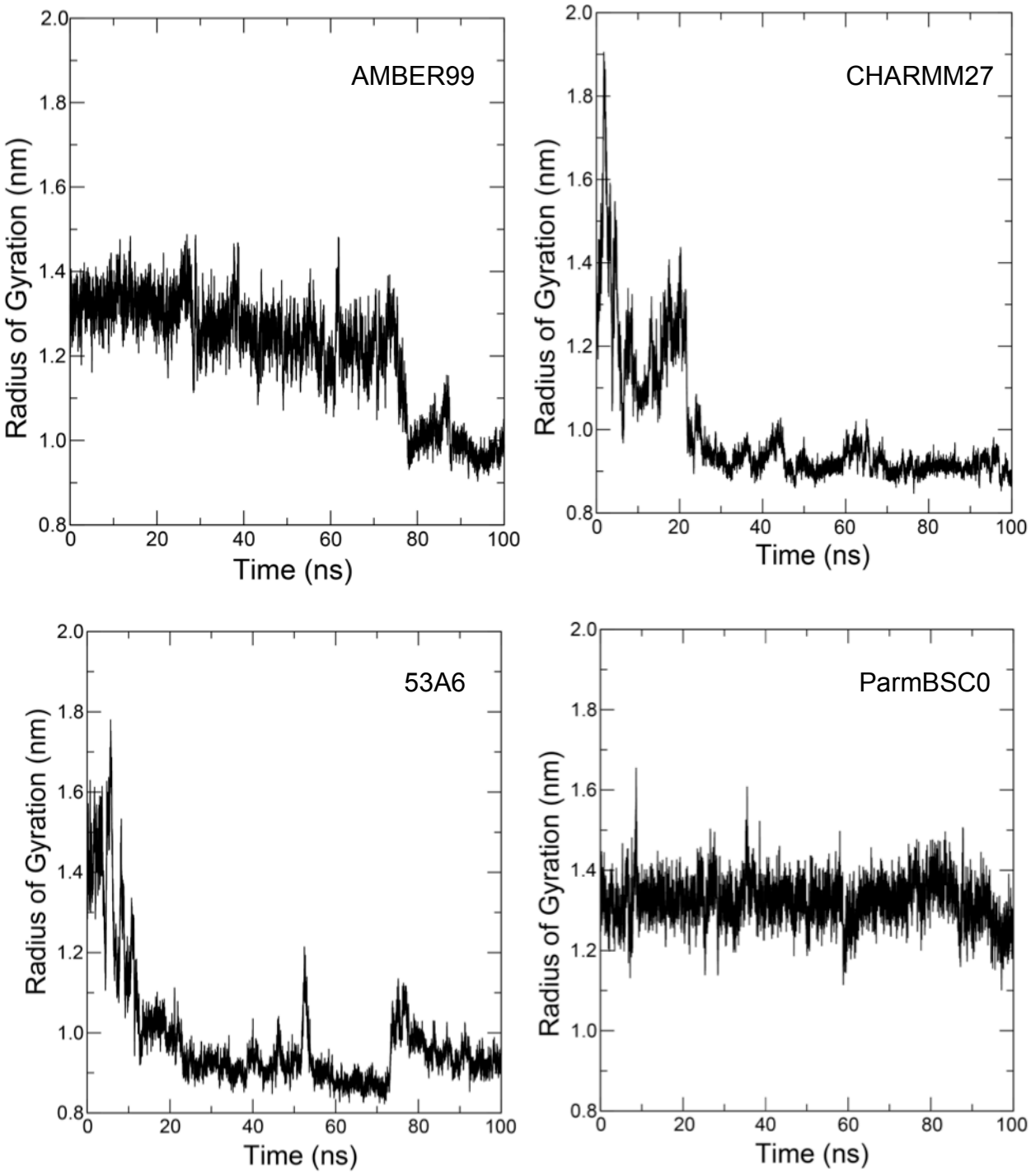

Figure 3. Radius of gyration of ssDNA across all force fields at 0.2 M salt concentration. Top left: AMBER99, top right: CHARMM27, bottom left: GROMOS96 53A6, bottom right: ParmBSC0

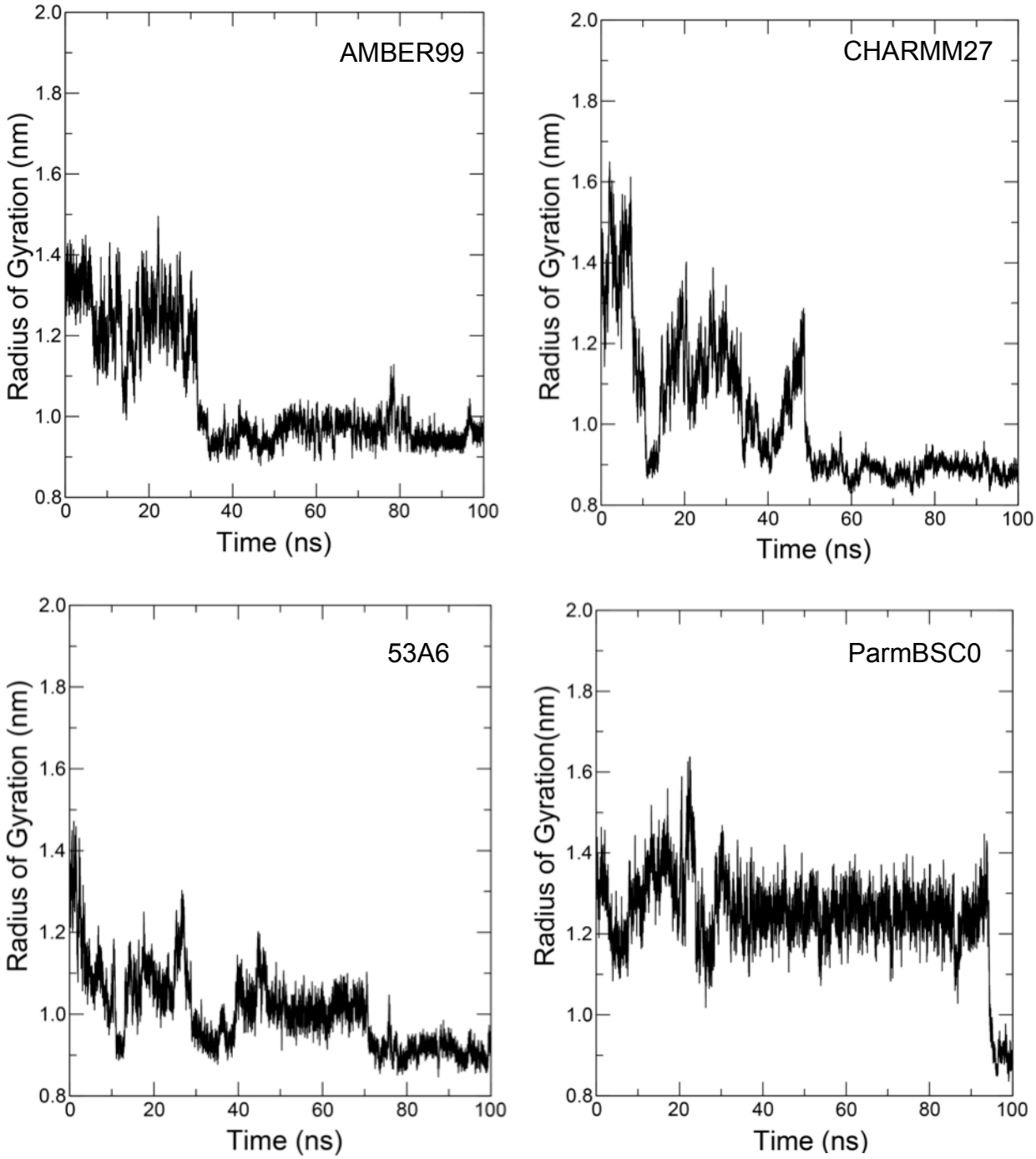

Figure 4. Radius of gyration of ssDNA across all force fields, at 0.5 M salt concentration. Top left: AMBER99, top right: CHARMM27, bottom left: GROMOS96 53A6, bottom right: ParmBSC0

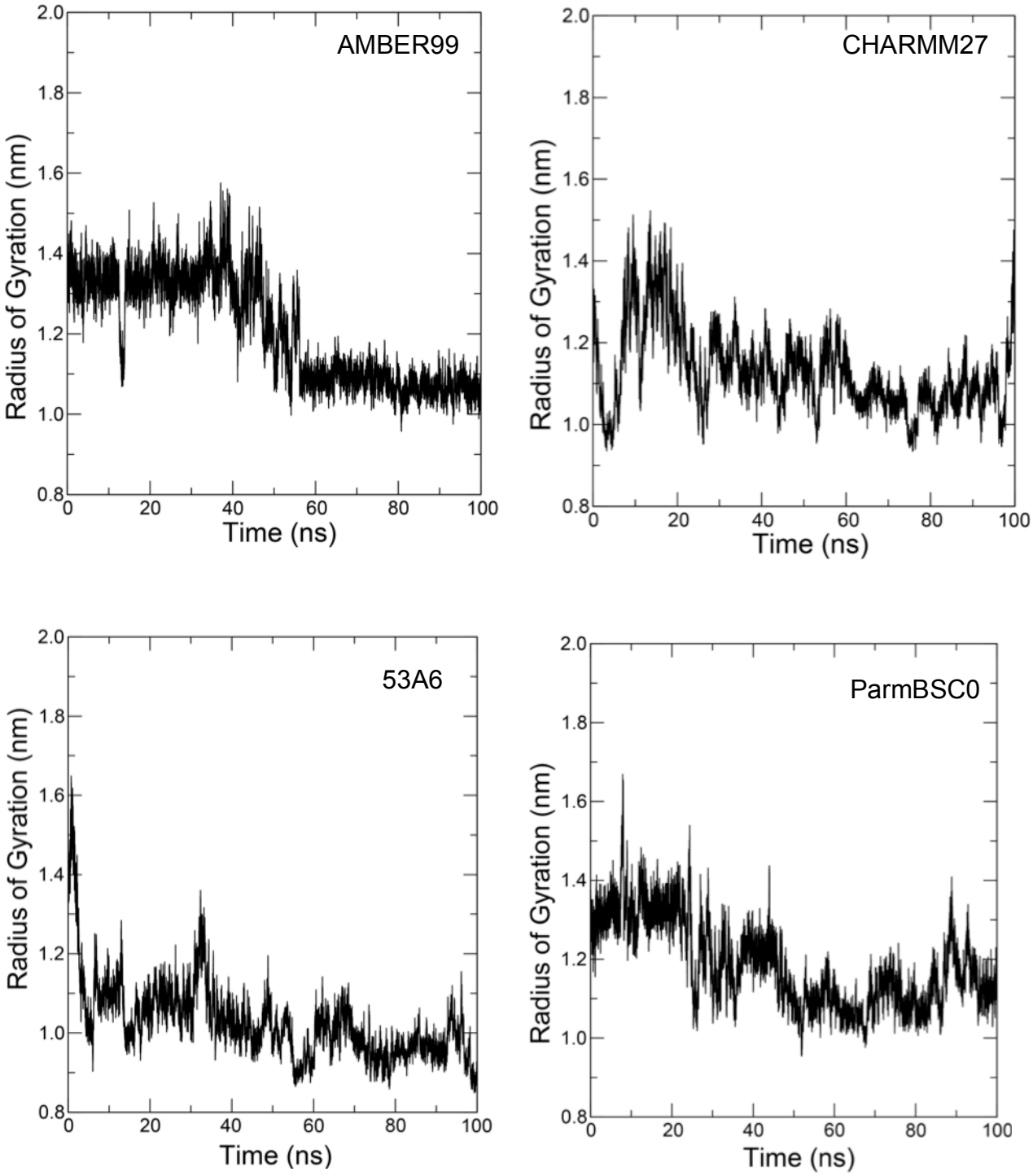

Figure 5. Radius of gyration of ssDNA across all force fields, at 1 M salt concentration. Top left: AMBER99, top right: CHARMM27, bottom left: GROMOS96 53A6, bottom right: ParmBSC0