UNIVERSITY OF Southampton

University of Southampton Research Repository
ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

http://eprints.soton.ac.uk

UNIVERSITY OF SOUTHAMPTON

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

SCHOOL OF CHEMISTRY

Free energy calculations of DNA translocation through protein nanopores and nanopore design for DNA sequencing

by

Richard Mathew Alexander Manara

Thesis for the degree of Doctor of Philosophy

December 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

Computational Chemistry

Richard Mathew Alexander Manara

DNA sequencing has vastly opened up the world of molecular biology, leading to new areas of interest, especially in medical research. Unfortunately the methods of DNA sequencing have only ever seen gradual improvements, as Sanger sequencing is still very much the norm despite its high cost and slow speed. Nanopores present an exciting opportunity for DNA sequencing, however, despite the concept being presented in 1996 several problems have prevented the creation of a publicly available sequencing device. The two main focuses of research into nanopores so far have been improving the resolution between bases and the slowing down of DNA translocation through the pore so modern ammeters can read the sequence accurately.

The simulation work presented in this thesis largely focuses on the energetics associated with DNA translocation. This is performed in several parts; an investigation into the probability of pore entry, study into the free energy of translocation for two proteins in addition to solvent contribution to this free energy, finally a theoretical project was undertaken to investigate bottom up nanopore design.

Table of Contents

List of Tables

List of Figures

DECLARATION OF AUTHORSHIP

I, Richard Mathew Alexander Manara, declare that this thesis entitled

Free energy calculations of DNA translocation through protein nanopores and nanopore design for DNA sequencing

and the works presented in it are my own and have been generated by myself as the result of my own original research.

I confirm that:

1. This work was done wholly in candidature for a research degree at this University;
2. 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;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. 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;
5. I have acknowledged all main sources of help;
6. 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;
7. Parts of this work have been published as:
8.

Nanomaterials *The nucleotide capture region of alpha hemolysin: insights into nanopore design for DNA sequencing from Molecular Dynamics simulations.* Manara, Tomasio, Khalid.

JCTC *Free-energy calculations reveal the subtle differences in the interactions of DNA bases with alpha-hemolysin.* Manara, Guy, Khalid.

Signed: ....

Date: ....

Acknowledgements

Firstly I would like to acknowledge everyone in the Khalid group, at all points of my PhD for their help. In particular Dr Andrew Guy for his extensive knowledge of the nanopore sequencing field and terrible humour. Also Jamie Parkin for his, frequently unique, discussions. Of course my supervisor Dr Syma Khalid deserves much thanks for her help, advice and guidance throughout my PhD.

Regarding funding, thanks to Oxford Nanopore Technologies for sponsoring this work. Thanks also to Jayne Wallace at ONT for helpful discussions over the PhD.

There are many friends I wish to thank as well. Dr Stephen Fox and Prof George Attard for advice both in and out of work, as well as Duncan Parker and Matt Parsons for their help keeping me sane during the writing period.

Definitions and Abbreviations

αHL: Alpha-Haemolysin

DNA: Deoxyribonucleic acid

dsDNA: Double-stranded DNA

ssDNA: Single-stranded DNA

OmpG: Outer membrane protein G

MD: Molecular Dynamics

MspA: *Mycobacterium smegmatis* protein A

PBC: Periodic boundary conditions

PDB: Protein Data Bank

PFT: Pore forming toxin

PME: Particle mesh Ewald

PMF: Potential of mean force

RNA: Ribonucleic acid

WHAM: Weighted histogram Analysis method

WT: Wild type

Chapter 1 Introduction

Gregor Mendel was the first to observe how the offspring of two flowers share traits from their parents [1], this discovery would lead to the field of genetics, the theory of evolution and eventually the discovery of the nucleus and nucleic acids. The structure of deoxyribonucleic acid, DNA, was characterised by Watson and Crick as the now iconic double helix [2]. Successful methods for sequencing DNA, such as Sanger sequencing [3] then opened up the field of molecular biology.

1.1 Chemical structure of DNA

DNA is comprised of nucleotides: negatively charged monomers which through a self condensation reaction form polymer chains, these chains can then form a double helix through base pairing, before binding to histones and chromosomes in the nucleus [4, 5].

1.1.1 DNA nucleobases

There are four naturally occurring DNA bases: adenine, thymine, cytosine and guanine. These are split into the purines and pyrimidines, as shown in Figure 1, the main difference between the two is that purines consist of fused heterocyclic rings, whereas pyrimidines are a single hetrocycle.

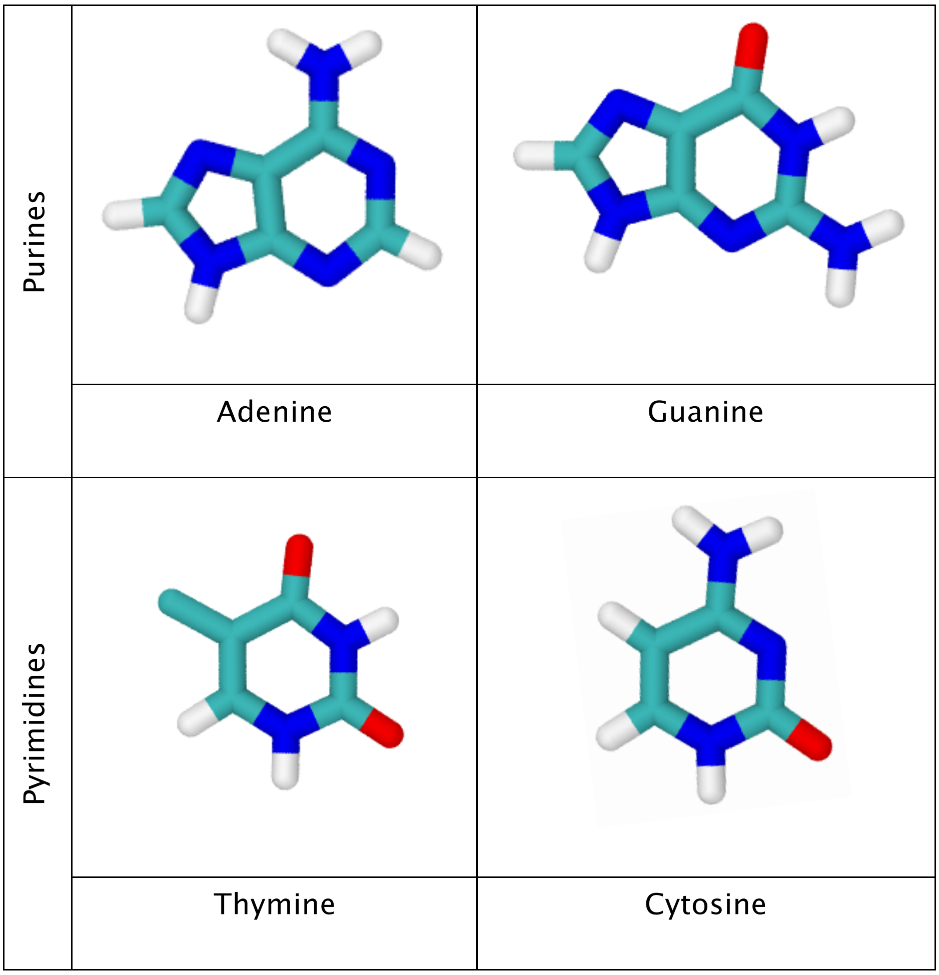


Figure 1: The four DNA bases. Where cyan represents carbon, blue, nitrogen, red, oxygen and white hydrogen.

These bases are attached to a phosphorylated ribose sugar, via replacement of the ring nitrogens' N-H bond with a N-C bond, to create a mononucleotide as shown in Figure 2. For the purines, this is the nitrogen on the five membered ring.

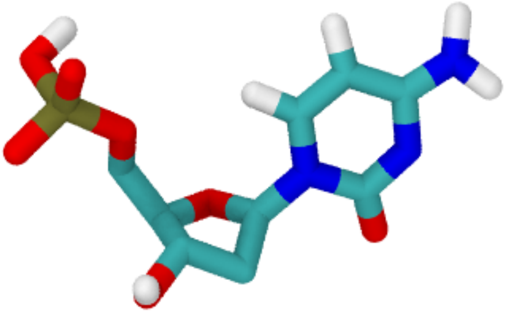


Figure 2: A cytosine DNA mononucleotide, colours as previously, with phosphorus in brown.

1.1.2 The double helix

DNA is largely stabilised by hydrogen bonds between base pairs, with the most stable base pairing occurring between a purine and pyrimidine. The most energetically favourable base pairs are adenine/thymine (A/T) and cytosine/guanine (C/G), proposed by Watson and Crick [2] along with the discovery of the double helix. It is worth noting that the CG base pair has 3 hydrogen bonds, whereas the AT pair has 2, shown in Figure 3; hence they are sometimes referred to as the ‘strong’ and ‘weak’ pairs. More recently, it has been suggested that the reason for this additional strength in DNA helices may be associated with pi-pi stacking interactions between base pairs in the double helix [6].

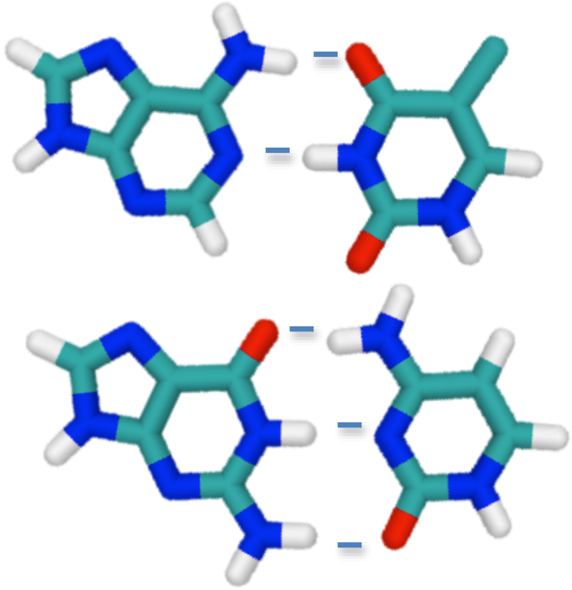


Figure 3: The AT (Top) and GC (Bottom) base pairs with hydrogen bonds shown as blue lines.

DNA double helices have been shown to exist in several forms in biology, see Figure 4, with A, B and Z helices being the three major forms. B-DNA, is the ‘traditional’ double-helix, proposed by Watson-Crick. In eukaryotic cells B-DNA wraps around histone proteins and then condenses into chromatin [7], whereas A and Z DNA, are less frequently seen, they are still naturally occurring. A-DNA is observed at low water concentrations and therefore is seen in some crystal structures, it can also form as part of a chimera DNA/RNA helix [8]. Z-DNA is very different to A and B-DNA, its biological role is uncertain, but believed to be associated with supercoiling [9, 10], which is discussed further in section 1.1.3.

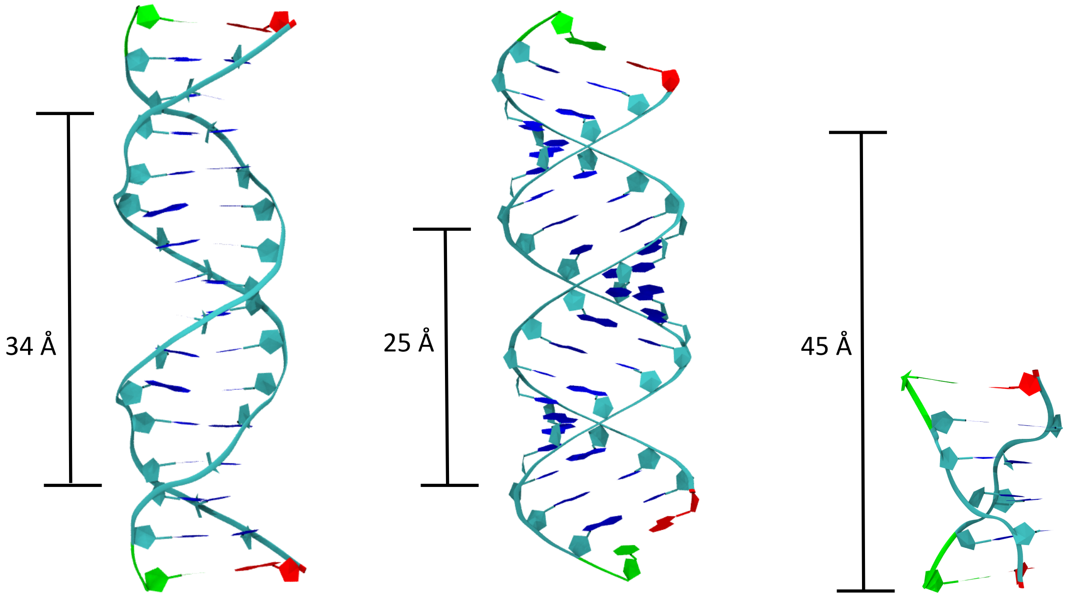


Figure 4: DNA helices, (Left) B-DNA, (Centre) A-DNA and (Right) Z-DNA. B-DNA taken from PDB 3BSE [11], A-DNA built online [12] and Z-DNA from PDB 2DCG [13]. The 5' terminal residues are shown in red and the 3' in green. Lines display the pitch per turn for each helix, note for Z-DNA the crystal structure does not complete a turn.

The primary methods for identifying the different DNA helices are: their turn handedness, the number of bases per turn, the pitch, and the relative sizes of the major and minor groove. As the dominant form of DNA, B-DNA is the standard, with a right-handed helix of 10 bases per turn and a pitch of 34 Å, with clearly identifiable major and minor grooves [2]. A-DNA is similar to B-DNA; it is again right-handed, with 10.7 bases per turn but a smaller pitch of 25 Å, which creates a wider helix, leading to the major and minor grooves becoming very similar in size. Z-DNA is the most unusual form, having a left handed helix with 12 bases per turn, a pitch of 45 Å and clearly identifiable major and minor groove [13].

1.1.3 DNA structure in a cell

Although short strands of DNA form helices in solution, it is not practical for cells to store DNA in its extended form. Indeed if the DNA of a single cell was randomly packed it would not be possible for the polymer to fit inside the cell [14], therefore higher levels of structure are used to reduce the space DNA occupies, similarly to coiling a garden hose.

In eukaryotic cells, DNA is organised as helices that are wrapped around an octomer of histone proteins, this complex is known as a nucleosome and has been previously crystallised [15]. Nucleosomes collectively coil to form chromatin fibres and chromosomes which are stored in the nucleus of the cell [16, 17].

In many bacteria, plasmids are also used to store some genetic information; these circular strands of DNA can have an external rotational force applied, which causes writhes and twists to form. Twisting in this sense is defined as the degree of helicity of the structure, i.e. the amount of times each strand wraps around each other. Twisting in DNA can be either positive or negative, causing the helix to contain either more, or less turns relative to its flat Watson-Crick helix. In comparison writhe is when the entire DNA helix wraps around itself, similarly to twisting a wire, which causes loops to form.

1.1.4 Non-standard base pairing

Despite Watson-Crick base pairs being the most energetically favourable, there are several alternative base pairing schemes, the most common of which are the ‘wobble’ base pairs and the Hoogsteen [18, 19]. Wobble base pairing is observed between alternative base pairs, in the Watson-Crick hydrogen bonding positions but with fewer or weaker hydrogen bonds, see Figure 5. The wobble base pairs are of a different width compared to the Watson-Crick pairs, which leads to a ‘wobble’ in the helix. Wobble base pairs often involve either an keto-enol tautomerisation or protonation of bases [18].

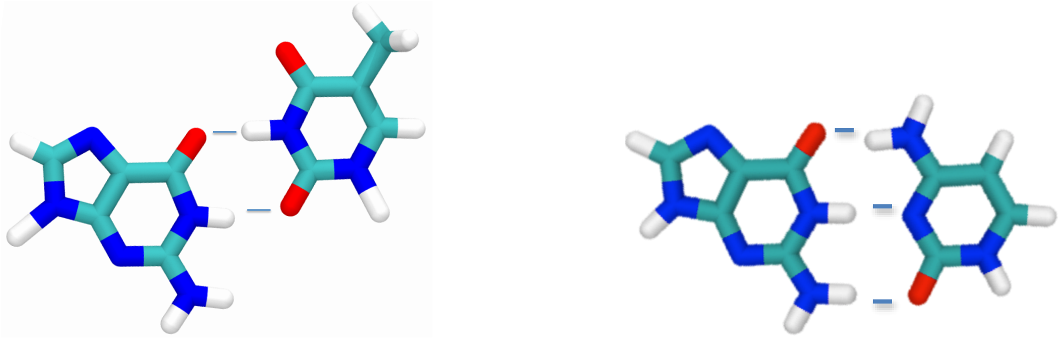


Figure 5: (Left) An example wobble base pair, GT, in comparison to (Right) the GC base pair.

Hoogsteen base pairs are formed when hydrogen bonds between base pairs occur in unusual locations [19]. They typically occur by rotation of the base/sugar dihedral, forcing the syn/anti bases to hydrogen bond outside the Watson-Crick positions. This can lead to unfamiliar structures, such as a triple helix or the G- quadruplex [20].

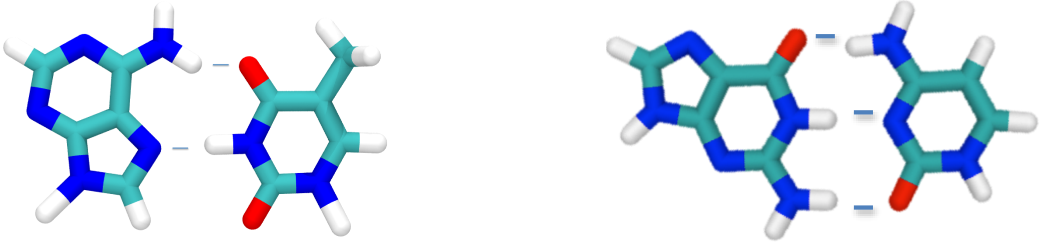


Figure 6: (Left) An example Hoogsteen base pair, GT, in comparison to (Right) the Watson-Crick GC base pair.

1.1.5 Differences between RNA and DNA

Ribonucleic acid, or RNA, as the name suggests, is an oxygenated form of DNA. The additional oxygen is present as an additional hydroxyl group on the ribose sugar in the 2’ position, as shown in Figure 7. The second hydroxyl group in RNA causes it to adopt a structure similar to A-DNA. RNA has many biological roles, notably as messenger RNA, ribosomal RNA, transfer RNA as well as carrying the generic information for viruses, in addition, there is also additional RNA in cells, the role of which is not entirely certain [21, 22].

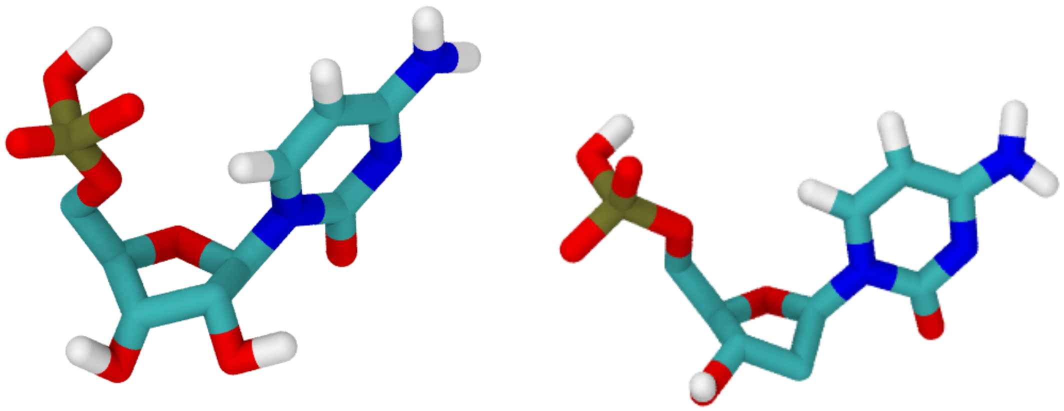


Figure 7: (Left) A cytosine RNA mononucleotide, in comparison to (Right) a cytosine DNA mononucleotide, note the additional hydroxyl group on the sugar.

In RNA, thymine is not present, instead being replaced with uracil, which is lacking the 5’ methyl group, otherwise the RNA bases are identical to those found in DNA, see Figure 8.

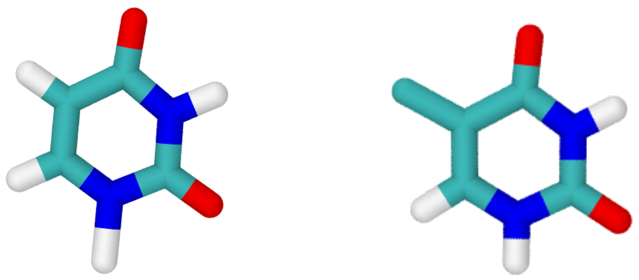


Figure 8: Comparison between (Left) uracil, and (Right) thymine. Note the missing methyl group on uracil.

1.1.6 Epigenetics and methylation

Epigenetics can be defined as “the study of changes in organisms caused by modification of gene expression rather than alteration of the genetic code itself” [23]. The molecular basis of epigenetics is carried out by several mechanisms, such as RNA interference [22, 24] or condensation of the helix into chromatin, preventing access to those genes [25] The most commonly referred to epigenetic modification that alters the molecular structure of DNA the methylation of cytosine into 5-methylcytosine [23], this is important as the additional bulk is detectable by nanopore sequencing [26].

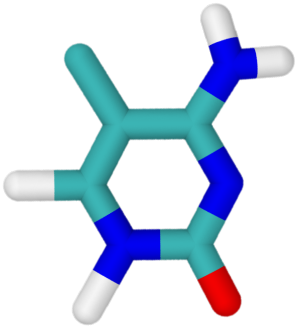


Figure 9: 5-Methylcytosine.

1.2 DNA sequencing

Initial sequencing techniques were used for RNA and pioneered by Ray Wu, the first DNA methods were slow, primer based methods that were sequence specific [27]. The first gene sequenced was from the bacteriophage MS2 [28], which was followed by the full MS2 genome four years later [29].

Several new methods were designed, notably Maxam-Gilbert sequencing [30] along with Sanger sequencing [3]. Maxam-Gilbert sequencing works by radioactive labelling of nucleotides. The modified DNA is then fragmented, and separated by gel electrophoresis, creating the recognisable separation of lines associated with DNA sequencing. The Sanger sequencing method relies on labelled di-deoxynucleotides, which lack the 3' hydroxyl group resulting in chain termination, followed again by electrophoresis for sequencing. Although Maxam-Gilbert sequencing was designed later than the Sanger method, it quickly gained popularity due to the ability to use purified DNA as opposed to requiring the cloning steps in Sanger sequencing. Unfortunately due to the large number of hazardous chemicals and difficulty in scaling, Maxam-Gilbert sequencing eventually fell out of favour and was replaced with Sanger based methods such as shotgun sequencing [31]. These improved methods allowed the mapping of the entirety of our DNA; the human genome project was started in 1990, and finished in 2003 [32, 33].

1.2.1 Uses of DNA sequencing

DNA sequencing has become routine practice in modern society. In this section we cover several ways sequencing technologies have been exploited, as well as how nanopore sequencing could improve the usage. For example, molecular biology is the study of genomes, proteins and genes [34-37]. DNA sequencing enables study of this field, whilst nanopore sequencing could also allow study of individual tissue's epigenetics, due to methyl cytosine, at a given time [38]. However improved sequencing would not have a large impact on evolutionary biology, a subset of molecular biology, which studies the historic relationships between organisms [39, 40], aside from reducing costs.

Another use of DNA sequencing is in the field of metagenomics, which is the detection of species in a sample [41], allowing study of organisms that do not grow in lab environments [42, 43]. This can also be used for detection of biological warfare agents [44]. Nanopore sequencing may improve the speed of this, as detection of specific genes has previously been demonstrated [45].

There are also several medical uses for DNA sequencing, which can involve detection of genetic diseases [46], research into how a person's genetics can influence health [47, 48] and parental testing. Cheaper sequencing would allow an increased number of genomes to be sequenced, which would increase the amount of research that would be possible, finding links between various genes and effects.

1.2.2 The Archon genomics X PRIZE

In 2006 Archon genomics offered a $10 million prize to the first group that could demonstrate a technique able to sequence 100 human genomes to high accuracy in less than 30 days for an average cost of $1,000. This became known as the $1,000 genome project and led, as intended, to an increased amount of targeted research to achieve this goal [49].

Several companies and researchers attempted to meet this target using novel sequencing methods. Notably Illumina met the target in January 2014, using a Sanger based method of chain termination, with the addition of fluorescent labels. Their method is highly parallel and uses reversible terminators, allowing for extremely rapid sequencing [50], indeed the implemented method of their sequencer can process approximately 20,000 genome a year [51].

1.3 Nanopore sequencing

One method born of the $1,000 genome race was nanopore sequencing [52], which will be the method investigated in this thesis. One of the main attractions of this method is the simplicity; the DNA is simply purified and sequenced, there are no amplification steps required and therefore the theoretical minimum cost is reduced.

Nanopore sequencing, works by setting up a system where a nanoscale aperture separates two chambers containing salt solution. When a voltage is applied across the aperture an ionic current begins to flow. The current that flows between the two compartments is proportional to several factors, one of which is the size of the aperture between them. If a substrate causes a blockage in the aperture the ionic current reduces, which leads to the conclusion that nanopores can be used as detectors for entities that can fit in the pore [53-56], as shown in Figure 10. As each nucleobase is a subtly different size, they would provide unique current blockages, giving the potential for nanopores to act as a sequencer of DNA, through inspection of ionic currents produced [57-63]. It is worth noting that the substrates have to be charged, else they are unlikely to be captured by the nanopore.

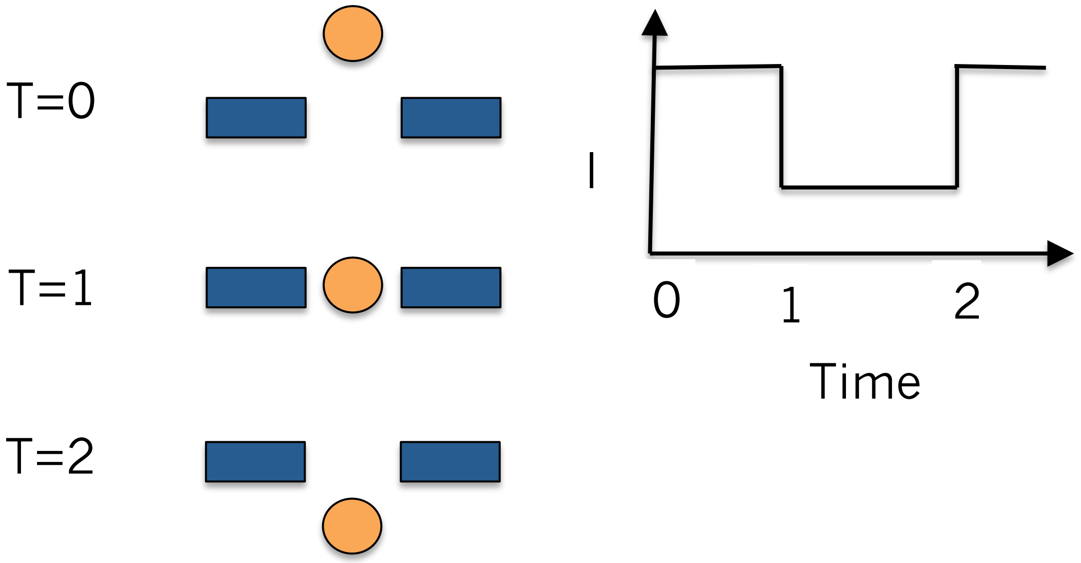


Figure 10: How nanopores can act as detectors of small, charged analytes.

This technique is very sensitive, it is even possible to detect enantiomers of amino acids or drug molecules [64, 65], with an engineered, chiral pore.

1.3.1 Manufactured nanopores

Nanoscale apertures can be man made; commonly out of silicon constructs or graphene layers. Benefits of using synthetic nanopores include their structural stability and that they lend themselves towards large scale manufacture. By structural stability we mean an increased stability over ranges of pH and temperature compared to protein pores [60, 66-68].

1.3.1.1 Silicon nanopores

Silicon nanopores are commonly made from silicon nitride or silicon dioxide. The process starts with a large pore being created by a high intensity beam. The resulting pore is then filled by movement of local silicon atoms into the pore to produce the required width [61].

Silicon nanopores have been shown to be capable of differentiating between short homopolymers of DNA, however the pores here are elliptical and it is unclear how this would influence sequencing [69]. In work reminiscent of early protein nanopore work, the length of DNA can also be assessed [70]. Studies that are more recent have demonstrated the ability to detect methylation in a double helix, although this is through recognition of a protein binding to the methylated DNA, as opposed to direct detection [71].

1.3.1.2 Graphene nanopores

Graphene based nanopores exist in two forms, either single or multiple layered [72, 73]. It is possible to alter the properties of graphene pores by the addition of atoms of varying electronegativity to the pore entrance, such as hydrogen or fluorine [67]. For graphene layers, surface deposition can also affect the properties of translocation [74].

As sequencers, graphene nanopores appear to have greater noise levels in the ionic current than those observed for silicon nitride pores [74, 75]. Simulations have shown that for single atom thick graphene pores not all bases are read, as translocation can occur in steps greater than 1 base at a time [73]. In contrast, other groups have reported a strong ability to control DNA translocation through use of varying the electric charge on the graphene, in this case, they were able to demonstrate “stop-and-go” translocation, suggesting that there is still potential for the successful use of graphene based DNA sequencers [76].

Of course there is both computational and experimental research emerging into new materials, such as MoS_2 as an analogue to graphene sheets [77, 78]. Indeed it would be surprising if as new nanomaterials are introduced they did not get tested as nanopore detectors.

1.3.2 Protein nanopores

Proteins, or polypeptides, are organic polymers made up of amino acids connected by amide bonds; the sequence of amino acids determines the final folding pattern and therefore function of the final protein. There are 20 naturally occurring amino acids, which vary in charge, steric bulk and polarity.

The structure of proteins is split into several levels, primary to quaternary. The primary structure is simply the order of the amino acids in the peptide chain. Protein secondary structure refers to alpha helices and beta sheets [79], which are covered separately below. The tertiary structure is the final arrangement of the monomer in space and the quaternary is how monomers interact together to carry out their function e.g. alpha-hemolysin is made up of 7 monomers that collectively form a pore [80].

Protein secondary structure, as previously described, consists of alpha helices and beta sheets both of which are stabilised by internal hydrogen bonding. Alpha helices are right handed, containing 3.6 residues per turn and each turn being 5.4 Å long. The stabilisation between beta sheets occurs from adjacent beta strands, which can run either parallel or anti-parallel. Beta sheets are more prone to developing further structures in proteins, such as beta-barrels, or propellers [81].

Biological nanopores have proven hopeful for sequencing DNA [62, 82-84], with those detailed in this section being subject to much scientific scrutiny. The primary advantage is the easy modification via mutagenesis or chemical modification [53, 64, 85, 86], while disadvantages are those expected from using proteins, i.e. the range of temperature and pH is limited.

1.3.2.1 Alpha-Hemolysin

Alpha-hemolysin, or αHL, shown in Figure 11, is a 240 kDa mushroom shaped, homo-heptemeric protein from *Staphylococcus aureus* [80] excreted naturally as a toxin, the mechanism of which is to act as a large pore through which water, ions and biological molecules, such as ATP, can pass through; leading to cell lysis and death. The protein consists of 2 domains, a large cap region with an internal vestibule and a 5 nm transmembrane beta-barrel, with a narrowest point of 1.4 nm.

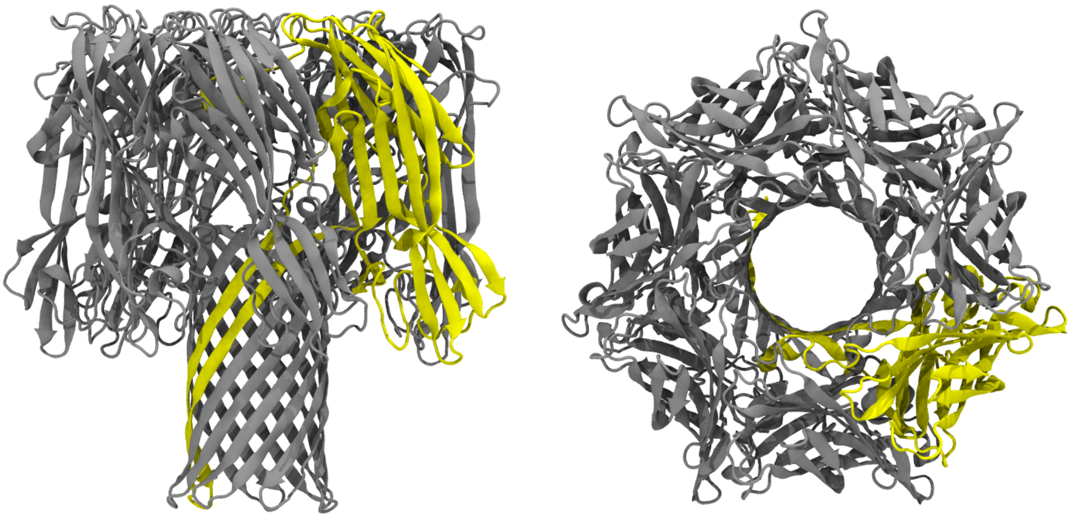


Figure 11: The Alpha-haemolysin nanopore with the monomer shown in yellow. Taken from PDB 7AHL [80].

Alpha-Haemolysin has been the subject of much scientific study as it was the first reported nucleotide detector [87]. Further experiments on αHL primarily focused on its ability to sequence DNA, detect small molecules and their chirality [65]. As such there is a large amount of data available on its ability to sequence DNA [26, 86, 87].

1.3.2.2 Outer membrane protein 'G'- OmpG

OmpG, is a 14 stranded barrel found in the outer membrane of gram-negative bacteria [88]. It is speculated that OmpG is a non-specific porin for oligosaccharides and unlike most outer membrane proteins, exists as a monomer [88, 89].

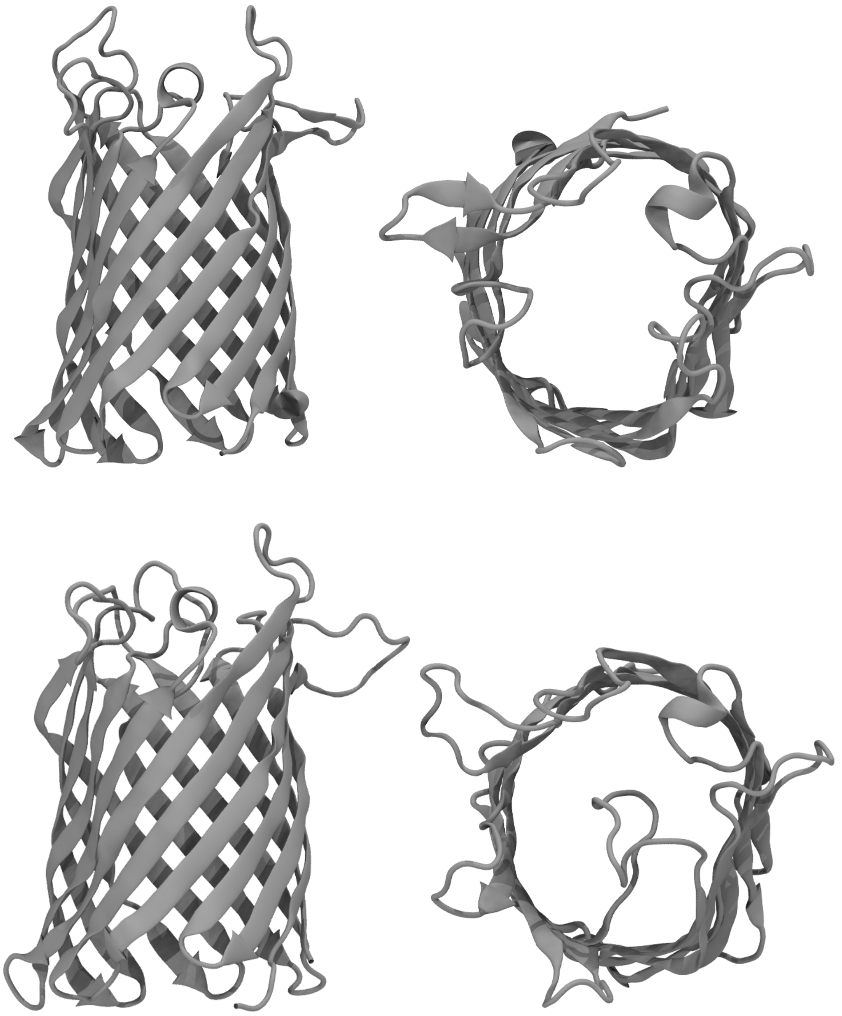


Figure 12: OmpG, with the gating loop in the open (top) and closed (bottom) position. Taken from PDBs 2IWW and 2IWV [90].

OmpG is comparable to the transmembrane domain of αHL, as it is a fourteen stranded beta barrel with similar dimensions [80, 88]. As OmpG is a monomer, this would allow for finer tuning via mutagenesis than αHL.

One unfortunate property of OmpG for its use as a stochastic detector is that a flexible loop is present, which gates the pore preventing translocation. This loop is both pH and voltage dependent see Figure 12, closing at 100 mV in either direction and below pH seven. In addition, spontaneous gating is also observed. The number of gating events can be reduced through protein engineering; indeed studies have been successful in elimination of up to 90% of gating events, although it has not yet been possible to completely remove them [91].

1.3.2.3 Mycobacterium smegmatis main porin A, MspA

MspA, Figure 13, is a 10 nm long, goblet shaped octomeric porin from *Mycobacterium smegmatis* [92], whose function is to allow transport of hydrophilic nutrients into the cell [93]. The transmembrane domain consists of two beta barrels of sixteen beta sheets each, these barrels are separated by a proline residue, which introduces a sharp kink, disrupting the hydrogen bonding required to maintain the beta sheets. The lower beta barrel is the narrowest point observed across the pore at 1.2 nm [92].

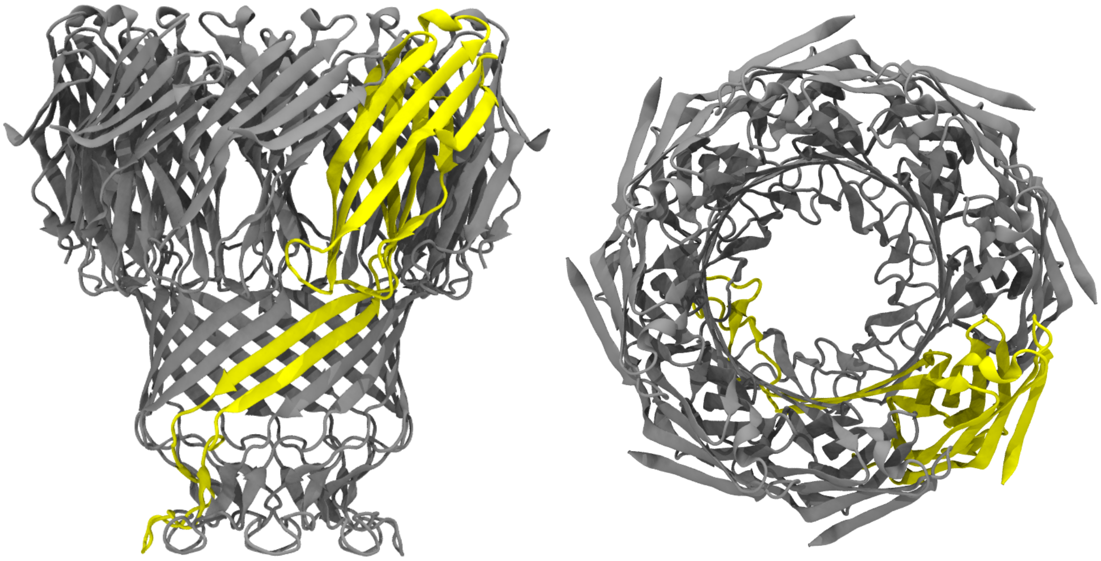


Figure 13: MspA, the yellow region is the monomer. Taken from PDB 1UUN [92].

Like αHL, MspA has proven to be suitable for nanopore sequencing [85, 94-96]. However, the wildtype protein is unsuitable, due to electrostatic repulsion between aspartic acid residues D90, D91 and D93, located in the lower beta barrel, and the DNA phosphate backbone. Mutant MspA proteins have been used successfully, commonly mutating the previously mentioned aspartic acids to asparagines [84].

1.3.3 Oxford Nanopore Technologies

Oxford Nanopore Technologies, ONT, is a company formed in 2005 based on research performed by Prof H. Bayley at the University of Oxford. They are the industrial sponsors of this work.

ONT has 3 sequencing platforms they plan on utilizing; MinION, PromethION and GridION. MinION is a small sequencing device, which is designed to be single use. PromethION is a bench top instrument, with a modular design enabling experiments to be tuned based on the system; from cheaper testing on small samples to parallel running of several. Finally GridION is designed as a modular instrument, either working individually or like a server, cross communicating to conduct sequencing [97].

In October 2013, the company launched the MinION access program. This early access program enables researchers to test this new technology in order to improve the system for the future.

1.3.4 Challenges with nanopore sequencing

The main issue holding back protein-based nanopores appears to be that the rate of DNA translocation is too rapid, which prevents accurate reading of the ionic currents produced by each monomer. As such one of the main focuses is on slowing down the DNA, without reducing the ionic current [82, 85]. This is a challenge as many traditional methods, such as increasing viscosity or reducing temperature, also result in a reduced ionic current [83].

Mutations to the proteins are often attempted in order to reduce the speed of translocation and intuitively this normally includes the addition of positive charges, to interact with the negatively charged phosphate backbone of DNA, however it has been shown that this can influence the order DNA strands exit the pore [98].

Enzymes can be covalently bound to the nanopore, in order to slow DNA translocation. One such enzyme is phi29 DNA polymerase [85]. This process is sometimes referred to as “ratcheting” the DNA through the nanopore [99].

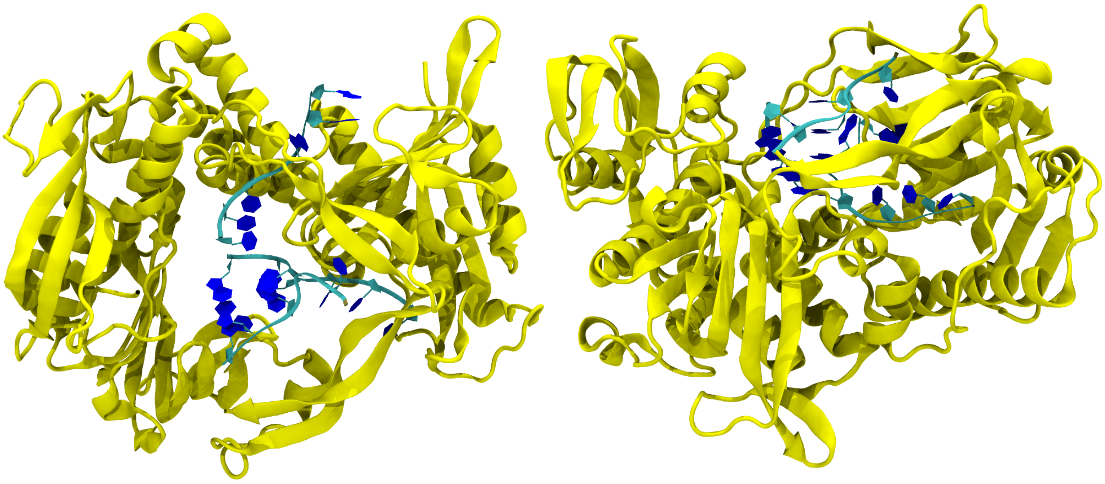


Figure 14: Phi29 DNA polymerase, from 2 orthogonal views. Taken from PDB 2PY5 [100].

Cyclodextrin rings have also been used to reduce the rate of translocation [101, 102], this often involves mutation of the pore to allow covalent linkage of the cyclodextrin to the protein, which also allows for control of the orientation of the ring [103].

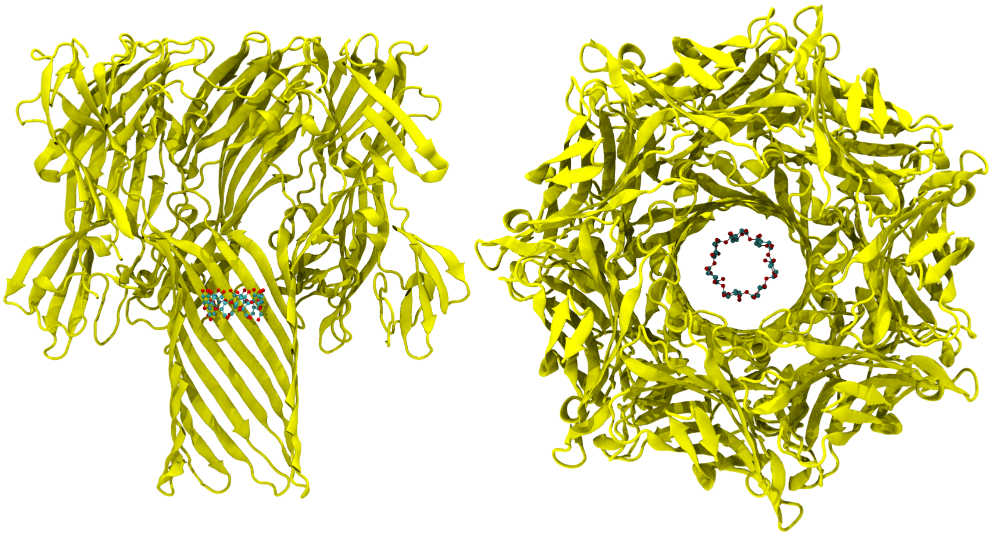


Figure 15: M113N αHL mutant, with β-cyclodextrin. Taken from PDB 3M4E [102].

Therefore, we will use atomistic MD simulations to investigate why the rate of translocation is so high, methods of reducing this rate and alternative nanopore based approaches.

1.4 DNA simulations

Prior to our investigation, we will discuss where simulations have been successful, especially in the context of DNA-protein interactions.

Although simulations of proteins have generally been well parameterised, those of nucleic acids lagged behind in performance in the 1990's. This is due to the following reasons: difficulties caused by the charged phosphate group, the lack of explicit hydrogen bonding terms in molecular dynamics simulations and short timescales [104, 105].

The charged phosphate group, combined with poor methodologies for electrostatics led to issues with initial simulations of DNA. Due to the large number of highly charged regions close together in DNA, electrostatic interactions must be calculated with high accuracy, to obtain realistic results. Unfortunately, limitations with simulation software in the late 90's [104] and hardware, meant algorithms had very short range cut-offs and no long range summations, so these early simulations encountered problems, even going as far as helices unwinding, with disruption between base pairs [106].

Hydrogen bonding in MD is only taken as explicit electrostatics and therefore is normally underestimated. It is expected for DNA, as the hydrogen bonding is so important that this will lead to artefacts, in early simulation work artificial restraints had to be added in order to maintain Watson-Crick bonding [107]. The terminal residues also are of importance, as even the recent, 2010, GROMOS 53A6 parameter set, or forcefield, does not contain these definitions and again, this causes significant distortion to the helix [108]. Even with terminal definitions, the GROMOS 53A6 forcefield adopts a conformation between A and B DNA structures [109].

Unfortunately, due to short timescales many of these issues were not discovered until hardware and software had improved, however now we are in a situation where simulations frequently are of the order of 10 ns to 1 µs, allowing forcefields to be much better parameterised.

1.4.1Current DNA forcefield comparison

The AMBER series of forcefields [110], have demonstrated that they are stable for B-DNA over short time periods (less than 10 ns) [111]. The A-BDNA consortium [112], formed in 2001, is a group of researchers who agreed to co-operate in order to improve the B-DNA parameters, which as previously described, were somewhat lacking. They have improved the AMBER parm99 forcefield, correcting it to the parmBSC0 forcefield which included alterations to the α/γ phosphate backbone torsions in 2007 [111].

GROMOS 53A6 is much more flexible than AMBER, frequently sampling space away from B-DNA and with a low energy barrier associated with base separation [113]. In general, GROMOS 53A6 is a proven protein forcefield but needs re-parameterisation of the α/γ phosphate backbone torsions before it is used for accurate dsDNA simulation [108].

In simulations of B-DNA, the CHARMM22 forcefield adopts the A form of DNA after only 600 ps, while the CHARMM27 forcefield remains as B-DNA [114]. Somewhat contradictorily, another study found that CHARMM27 samples the space between the A and B forms of DNA [115], suggesting that there may be either a velocity or sequence dependence to the flexibility of this forcefield. Therefore, the literature suggests that for study into dsDNA, the PARMBSC0 forcefield is the most suitable.

1.4.2 DNA simulations in solution

There are many studies that detail the flexibility of the DNA polymer and to this end all possible junctions between the 4 bases have been extensively studied [116, 117]. Simulations have also enabled study into how the solvation layer interacts with the grooves of DNA [118]. It appears that ions predominantly interact with electronegative sites along the strand, e.g. the phosphate group, but largely do not sample with the bases in the grooves. MD has also been used to show how temperature can cause melting of dsDNA [119].

Comparatively little work has been performed on the alternative forms of DNA. This is partially to do with the increased difficulty of working with non B-DNA experimentally, but also is due to the dominance of B-DNA in nature. Overall this has resulted in little interest in simulations of alternative DNA structures, apart from a few ‘exotic’ structures, such at triple helices, Z-DNA and G-quadruplexes [20, 120, 121].

As previously described in 1.1.3, DNA is stored as a compressed version of the helix in cells. There has been much simulation work into how external forces applied to helices and plasmids [122, 123], can cause twist and writhe of the DNA. This has revealed that GC regions are less prone to the effects of supercoiling in comparison to AT regions [124] and extreme torsions can cause localised deformation for the DNA helix [125].

ssDNA is more difficult to simulate as there is a lack of data from experiments to compare to and no crystal structures in the protein data bank exclusively of ssDNA, making it difficult to know if forcefields are accurately representing reality. As the forcefields are parameterised for dsDNA, they generally are more rigid than might be expected. In the case of PARMBSC0 and AMBER, DNA can retain its initial helical structure for more than 40 ns, however, it has been shown that GROMOS and CHARMM forcefields are much more flexible [126]. An issue with this is that we cannot compare this data to experimental values and therefore cannot know which is more accurate. Unfortunately, due to the requirement of using the TIP3P water model, simulations require much more computational resource to run with the CHARMM27 forcefield. Therefore, for simulations of ssDNA, GROMOS53a6 appears to be the most suitable, due to the expected flexibility of the polymer, which allow for more rapid exploration of phase space.

1.4.3 DNA interactions with proteins

DNA is frequently found in complexes with proteins, for example during DNA replication or transcription as well as binding to histones for dense DNA storage in the nucleus. Simulations have frequently studied these complexes, in general, investigating how the protein causes distortion of the DNA away from the canonical structures, in addition to how the protein re-organises itself in order to bind more favourably [127-129].

There is much dependence on sequence for many DNA-protein interactions. P53 for example binds to a sequence of “two RRRCWWGYYY decamers separated by a small number of bases” [130], where A, T, C, G are the 4 standard nucleic acids, R is guanine or adenine, W is thymine or adenine and Y is cytosine or thymine. If even single point mutations are introduced to proteins, they will not recognise their target, further emphasising their specificity [131].

1.5 Nanopore simulation

1.5.1 Solid-state nanopores

Simulations on solid-state nanopores, for silicon and graphene pores, have enabled detailed study of translocation [60, 73]. Despite having some success, an accurate forcefield has not been created for silicon nitride that also includes DNA parameters, this led to parameters being combined from separate forcefields [60], which is not expected to produce accurate results. Despite these limitations, simulations have managed to show how DNA behaves in the pore for both ssDNA and dsDNA, highlighting how the differences between the two regarding mechanical properties allows translocation through various pores [132].

1.5.2 Protein nanopores

Protein nanopores have been the subject of much investigation through simulation methods [61, 83]. Like the experimental works, simulations have covered αHL [98, 126, 133] and MspA [134].

Free energy calculations have been used in order to calculate the free energy change associated with translocation [135], yet this study was at low resolution; only 1 barrier to translocation was observed and due to the length of DNA used (20 bases), it was not apparent as to which residues are responsible for the gain in energy.

Mutations are frequently introduced and the results compared to the WT pore [98, 134], enabling simulations to be used predictively as well as validating against previous experimental works. Simulations allow us to explain on a molecular basis why events occur, which allows us to justify or reject the use of mutations; for example previous work has shown that the G199R mutant in αHL slow DNA, one of the properties mutant pores attempt to introduce, but cause the bases to exit in an incorrect order, potentially leading to issues regarding sequencing [126].

Frequently, truncated pores are used to effectively reduce the size of the proteins [98, 126, 133]. For αHL, a system for the full protein is approximately 20,000 atoms, whilst with a model pore a similar system will contain only 2,000 atoms and therefore run in a tenth of the time.

Simulations have also investigated the effect of the solution on the DNA, alongside conformational dynamics of the ssDNA. Markosyan *et al* suggest that water clustering largely is drowned out by noise in αHL [136].

1.6 Considerations

Although nanopores have been subject to much scientific scrutiny regarding their ability to sequence DNA, there are still obstacles to fully understanding these systems, most notably the rate of translocation. Therefore we will investigate methods of reducing the rate ssDNA translocates and alternative means that circumvent this issue.

One mystery in nanopore sequencing is that the energetics associated with translocation are largely unknown. Whilst experimental work can observe rates and ionic currents as DNA translocates nanopores, simulations can explain why this is the case. It will also be possible to explain how mutations influence these energetics and therefore guide future experiment design.

1.7 Aims

- To investigate sequencing methods that circumvent the issues involved with ssDNA.
- To explore the energetics of DNA translocation through both αHL and MspA.
- To investigate how charged mutations influence DNA translocation.

Chapter 2 Methods

2.1 Computational chemistry

Computational chemistry is now a frequently used tool for scientists to gain an understanding of molecular systems and events they would not otherwise be able to observe. It allows the study of chemistry on the nanometre and nanosecond scale. Under experimental conditions it is not possible to gain an insight into how an individual atom centre, or amino acids residue, can affect a biological system, however they can be of great importance for explaining how a ligand binds to a pocket, for example showing how displacement of individual water molecules or hydrogen bonds can stabilise a structure [137].

The most intuitive way to model chemical systems is to simulate the nucleus and electrons of each atom, leading to a quantum mechanical model (QM) which explicitly shows each individual electron and nucleus, however this involves approximating the Schrödinger equation for each atomic centre in the system. Even with modern computational resources QM simulations are currently limited to small systems over small timescales e.g. 100's of atoms over picoseconds [138, 139]. There are several methods available to maintain a high level of electronic accuracy, however due to limited timescales these are unable to model many systems for a sufficient length of time for many biological systems, therefore the simulation time must be increased and accuracy sacrificed, a trade off that is seen frequently throughout computational chemistry.

2.2 Molecular mechanics

When performing molecular simulations of biological systems it is common practice to assume all atoms are in the electronic ground state and treat each atom as a soft sphere. This approximation is known as molecular mechanics (MM). MM methods are able to model much larger systems for longer timescales, e.g. millions of atoms or microseconds [140]. However as electrons are not explicitly modelled, any study of electronic movement, including bond formation, charge transfer and spectroscopic measurements are no longer possible. Therefore they are frequently used to study molecular systems in which non-bonding forces dominate, such as cell membranes [141].

There are two main methods of sampling QM and MM systems, Monte Carlo (MC) and Molecular Dynamics (MD). MC calculations work by doing “random moves”, which will either be accepted or rejected based on the simulation parameters, normally depending on the change in energy. It is worth noting that the MC method is independent of time [142]. MD in comparison uses Newton’s laws of motion to update the positions of particles from a starting structure with velocities. MD indicates how the system develops over time and was the method used in this thesis.

2.2.1 Levels of accuracy

There are various models that exist whilst performing MM simulations, shown in Figure 16; they increase the speed of the simulation by reducing the number of particles simulated.

The most intuitive MM model is one at atomistic resolution, where every atom is simulated as a separate particle, but at the cost of speed [143]. The united atom model merges non-polar hydrogen atoms in the system with their parent bonded atom in order increase efficiency, e.g. CH_3 would become a single particle. This model dramatically increases the speed of systems where long carbon chains are present, such as lipids, without compromising on accuracy [144, 145].

The merging of atoms used in the united atom model can be carried out on a larger scale, at the cost of accuracy, by grouping several heavy atoms with similar properties into one 'bead'. These beads represent their constituent atoms with appropriate charges and polarities, they can also be non-spherical, in some force fields, elliptical beads are used to more accurately represent the system [146, 147]. Coarse-grained (CG) models do have drawbacks, compared to united atom and atomistic simulation, due to the smoothness of the long-range energy landscape diffusion takes approximately a quarter of the time compared to atomistic simulation, as there are fewer micro-states, in addition individual atomic motions are lost [148].

Several levels of accuracy can be introduced to the same simulation, such as the highly mobile membrane-mimetic (HMMM) model [149], where united atom and coarse-grain representations are present in the same simulation box. Merging of methods has also been done with QM and MM, allowing electronic study in a region of particular interest, such as an enzyme active site [150].

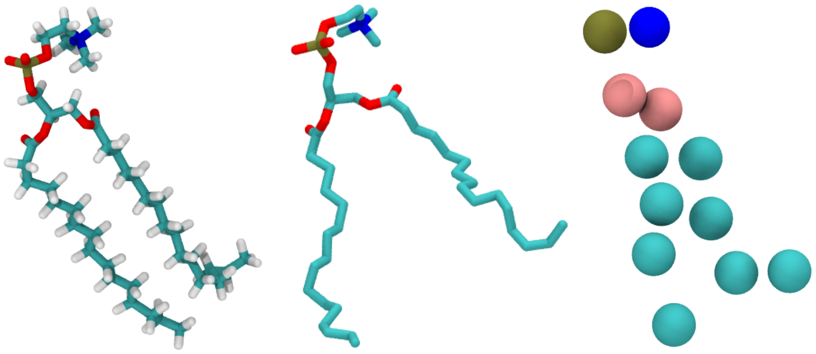


Figure 16: A DPPC lipid shown from left to right as atomistic, united atom and coarse grain models. Note the reduction in atom count from left to right.

These methods not only increase simulation speed by reduction of atom count, but also through increasing the timestep, this is discussed more in section 2.4.2. Therefore for a comparable sized system moving from coarse grain to atomistic may reduce the atom count to a quarter, but the speed up in simulation time should be far in excess of four times.

2.2.2 Energy minimisation

Energy minimisation is a process by which the particles in a system are moved to a lower potential, this is done in order to reduce steric clashes and bring the system to a local potential energy minimum, thereby increasing the stability of the system before MD is performed.

The methods of steepest descents and conjugate gradient are very similar. The main difference between the two methods is that conjugate gradient takes into account the previous position as well as the current gradient, whereas the steepest descents method will move the structure orthogonally to the previous step. This is due to steepest descents being a first derivative method, whereas conjugate gradient is a second derivative method. Ultimately conjugate gradient is more efficient, but will require more computational resource than steepest descent.

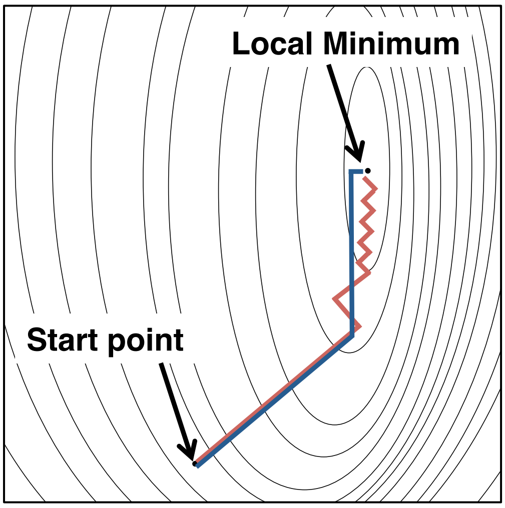


Figure 17: A comparison between steepest descents, in red, and conjugate gradients, blue. The steepest descents method simply makes moves orthogonally to the last step whilst for the comparison conjugate

gradient method, the previous steps are taken into account when calculating the new direction.

2.2.3 MD methodology

Simulations only require an initial structure and a set of parameters describing how to perform the simulation. For protein simulations the starting structure will be from a crystal structure or a homology model [151, 152], that has undergone some pre-simulation processing in order to add atoms not crystallised [153] and to introduce the structure into an appropriate environment e.g. solvated in KCl solution, or inserted in a lipid bilayer [154].

With the initial co-ordinates we apply a forcefield, described in section 2.3, which gives us the potential energy of the system. The negative gradient of the potential energy yields the resultant force on the particle, shown in Equation 1.

$$\vec{F} = -\nabla \left(k_{\boldsymbol{\theta}} \left[1 + \cos\left(n\{\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{eq}}\}\right)^{2} \right] + \tfrac{1}{2} k_{\boldsymbol{\omega}} \left[\boldsymbol{\omega} - \boldsymbol{\omega}_{eq} \right]^{2} + \tfrac{1}{2} k_{\mathbf{b}} \left[\mathbf{b} - \mathbf{b}_{eq} \right]^{2} + 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] + \frac{1}{4\pi\varepsilon_0} \frac{Q_1 Q_2}{r} \right)$$

Equation 1: The relationship between force and potential energy. Where *F* is the force vector and the remaining terms, sum to the potential energy.

The acceleration is calculated from the force using Newtons' second law, Equation 2.

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

Equation 2: Newtons' second law, *F* is the force vector, *m*, mass and *a* is the acceleration vector

Combining Equations 1 and 2, along with the definitions of acceleration, it is possible to show how the potential of a particle relates to its change in position, Equation 3. The acceleration is applied for a finite duration, the timestep, by the integrator, which then leads us to updated co-ordinates and a new set of potential energies allowing the cycle to repeat from calculation of potential, as shown in Figure 18 below. Under normal circumstances, the co-ordinates are saved periodically, which leads to a trajectory, showing how the system changes over time.

$$-\nabla PE = m\vec{a} = m\frac{\delta\vec{v}}{\delta t} = m\frac{\delta^2 r}{\delta t^2}$$

Equation 3: How potential energy is related to the position. Where *v* is final velocity, *t* is time and *r* is position.

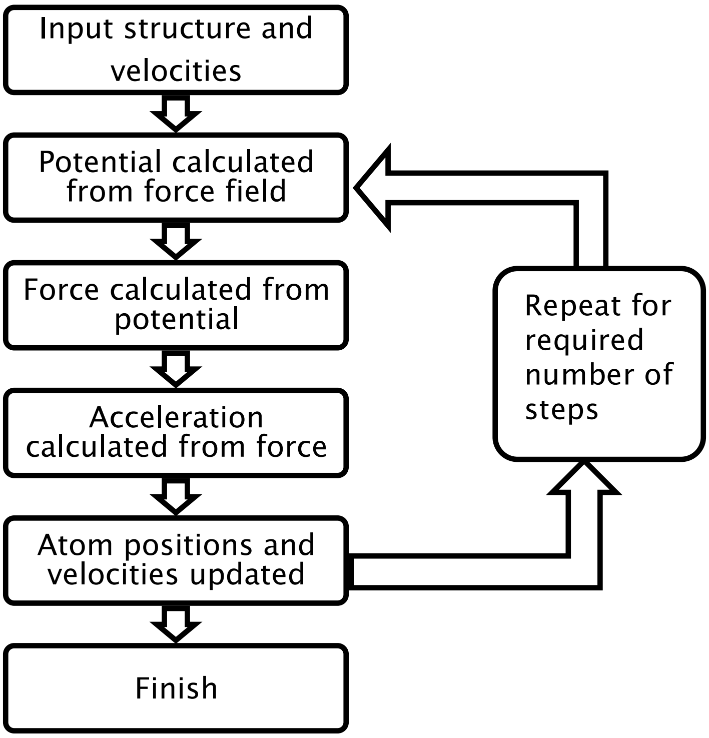

Input structure and velocities
Potential calculated from force field
Force calculated from potential
Acceleration calculated from force
Atom positions and velocities updated
Finish
Repeat for required number of steps

Figure 18: A diagram explaining the flow of MD simulations.

2.3 Forcefields

The forcefield is how the simulation software can calculate the potential of each atom; as such it is a summation of bonded and non-bonded terms. This explanation is for the GROMOS set of forcefields [109, 155-158].

2.3.1 Bonded terms

The bonded terms in a forcefield are made up of four parts: the ‘proper’ and ‘improper’ dihedrals, bond angles and bond lengths.

‘Proper’ dihedrals are potentials due to groups of 4 atoms rotating around a bond and as such are modelled using a cosine potential. The relationship between potential energy and angle is shown in Figure 19, Figure 20 demonstrates rotation of a dihedral and Equation 4 illustrates the functional form.

$$PE = k_{\boldsymbol{\theta}}\left[1 + \cos\left(n\left\{\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{eq}}\right\}\right)\right]$$

Equation 4: Potential energy relationship to dihedral angle, PE is the potential energy, *k* is a constant, *n* is an integer, $\boldsymbol{\theta}$ is the angle and $\boldsymbol{\theta}_{\mathbf{eq}}$ is the equilibrium angle.

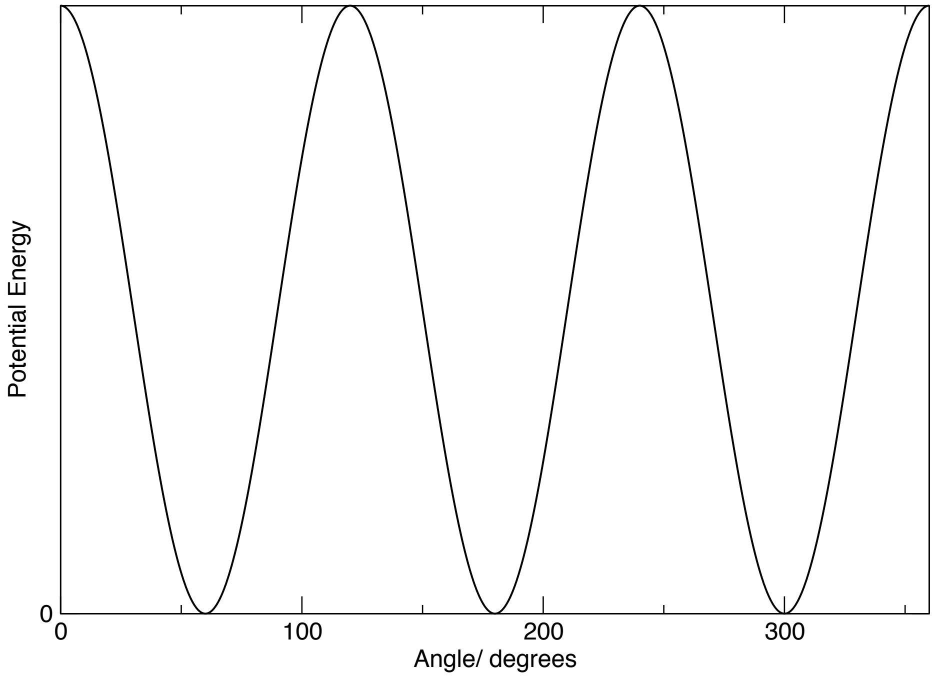


Figure 19: Potential energy relationship to a simple dihedral angle

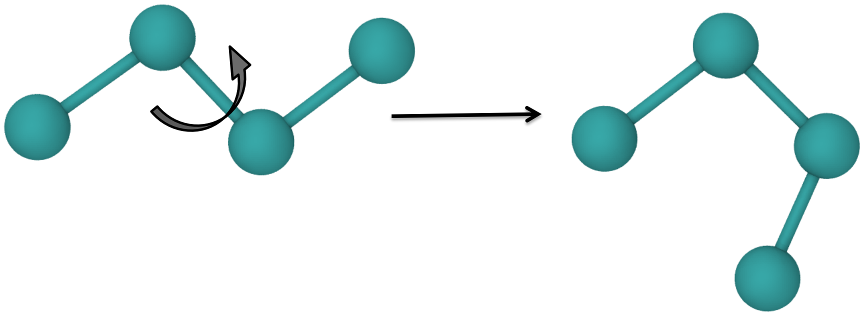


Figure 20: Rotation of a proper dihedral

Improper dihedrals also model 4 atom motion, however their role is opposite to proper dihedrals, instead of being used to allow rotation around a bond improper dihedrals exist to prevent motion. This is useful for systems where the stereochemistry is important, for example planar ring systems or chiral centres, shown in Figure 22. The potential is a harmonic as shown in Equation 5 and Figure 21.

$$PE = \frac{1}{2}k_{\omega}\left[\boldsymbol{\omega} - \boldsymbol{\omega}_{eq}\right]^{2}$$

Equation 5: Harmonic potential for improper dihedrals, ω is the out of plane angle

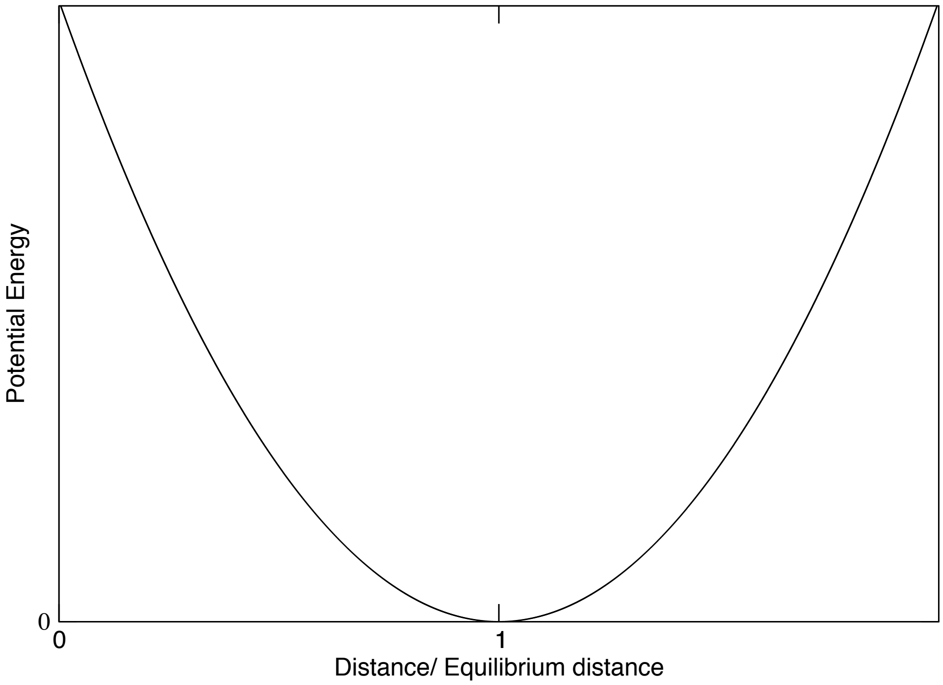


Figure 21: Harmonic potential. Minimum value is when the distance is equal to the equilibrium value.

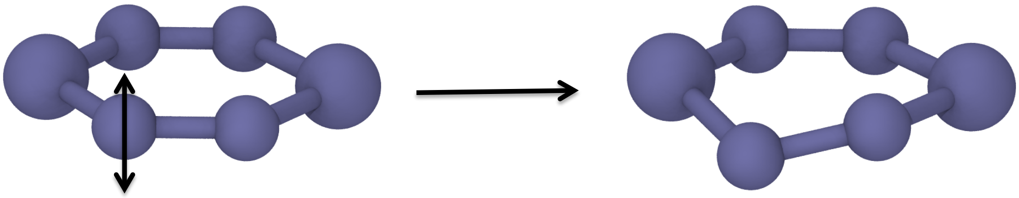


Figure 22: Movement of an improper dihedral for a planar ring system

Bond stretching, shown in Figure 23, like improper dihedrals, follows a harmonic potential and therefore is represented with a similar equation, shown in Equation 6. The timestep of simulations must be used in order to capture the highest frequency vibration, determined by the harmonic constant for bonds, discussed in section 2.4.3.

$$PE = \frac{1}{2}k_{\mathbf{b}}\left[\mathbf{b} - \mathbf{b}_{eq}\right]^2$$

Equation 6: Harmonic potential for bond stretching, **b** is bond length

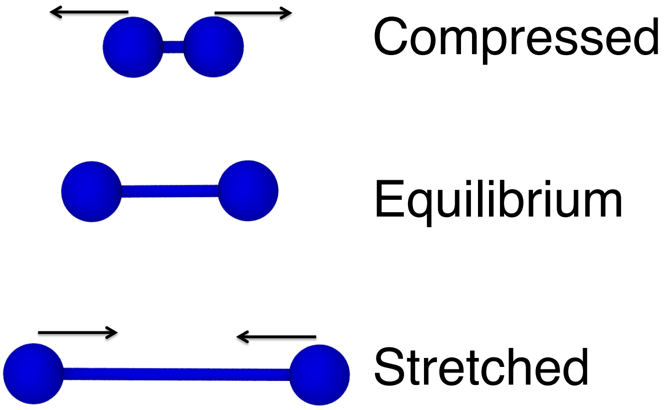


Figure 23: Bond stretching and compression, with resulting forces, for a simple diatomic molecule

Angle bending occurs between 3 atoms, shown in Figure 24, and again is modelled by a harmonic potential, shown in Equation 7 and plotted in Figure 25.

$$PE = \frac{1}{2}k_{\boldsymbol{\theta}}\left[\boldsymbol{\theta} - \boldsymbol{\theta}_{eq}\right]$$

Equation 7: Harmonic potential for bond angles

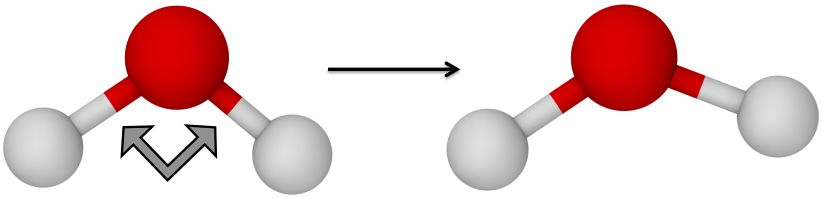


Figure 24: Angle bending for water, note the bond lengths are unchanged but the distance between the hydrogen atoms has increased.

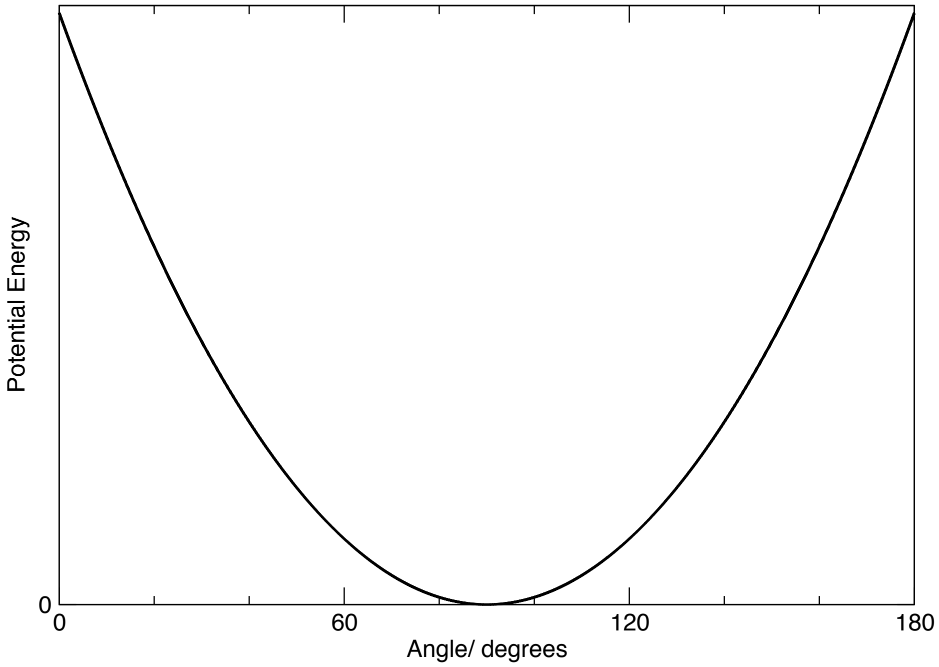


Figure 25: Harmonic potential, for an angular change, where the equilibrium angle is 90 degrees.

2.3.2 Non-Bonded terms

The non-bonded terms in a forcefield cover 3 forces, Pauli repulsion, Van der Waals forces and electrostatics. Pauli repulsion and Van der Waals forces, shown in Figure 26, are modelled by the Lennard-Jones potential, given by Equation 8 and plotted in Figure 27.

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

Equation 8: The Lennard-Jones Equation, where r is distance between particles, σ is the distance of zero potential and ε is the minimum potential.

Figure 26: Van der Waals attraction, shown on the left, is the ubiquitous weak force present between all atoms at all times. Pauli repulsion, on the right, is the strong repulsive force caused by electrons attempting to enter already occupied orbitals.

The Lennard-Jones equation models Pauli repulsion and Van der Waals forces, by the r^{-12} and r^{-6} terms respectively, as seen on the plot seen below. The energy is infinite at zero distance, dropping to zero when the distance equals σ, continuing to reduce in energy until after it passes the minimum potential, ε, beyond which the curve tends towards zero. This behaviour can be explained by the combination of forces being modelled; at distances close to zero Pauli repulsion is the major constituent, modelled by the r^{-12} term, but this rapidly decreases as the atoms pass beyond their Van der Waals radius, when the r^{-6} term starts to dominate, leading to an attractive potential, weakening as the inter-atomic distance increases.

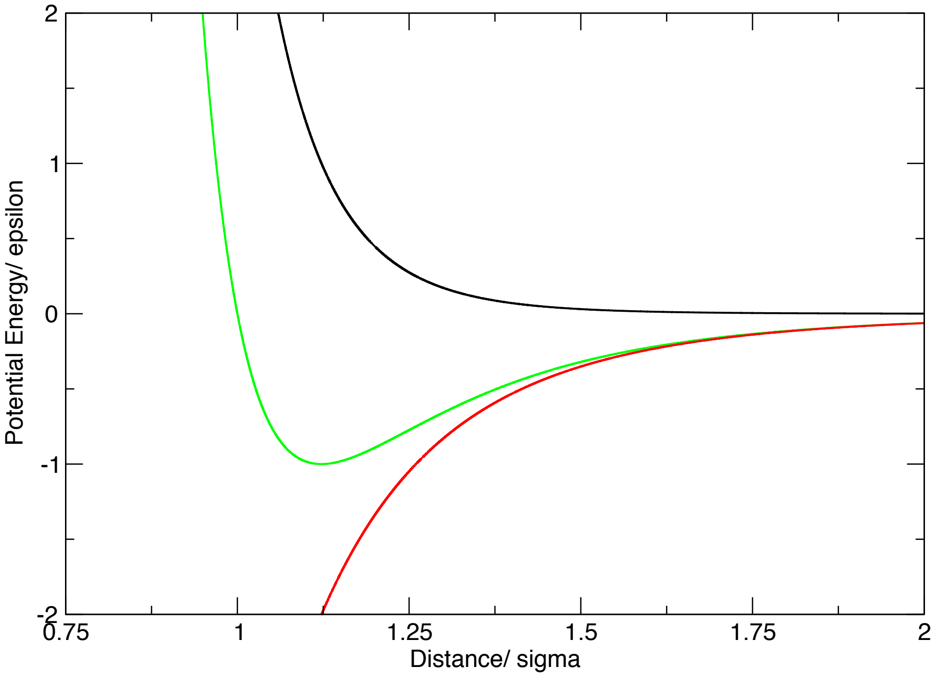


Figure 27: The Lennard-jones potential in green, note the composite parts, the r^{-12} term in black and the r^{-6} term in red.

The final non-bonded term is the Coulomb potential, shown in Equation 9, which is used to calculate the potential between 2 charged particles. Unlike most equations in MD, is not a simplification from reality and is the real potential. Figure 28 plots how the Coulomb potential varies for several charge combinations. The Coulomb potential decays rapidly, from infinite at zero distance, to zero at infinite distance. Coulombs' potential can be either attractive or repulsive, dictated by the signs of the charges, as similar charges repel and opposites attract.

$$V = \frac{1}{4\pi\varepsilon_0}\frac{Q_1Q_2}{r}$$

Equation 9: The Coulomb potential, where ε_0 is the permittivity of free space, Q_1 and Q_2 are the charges on particles 1 and 2 and r is the distance between them.

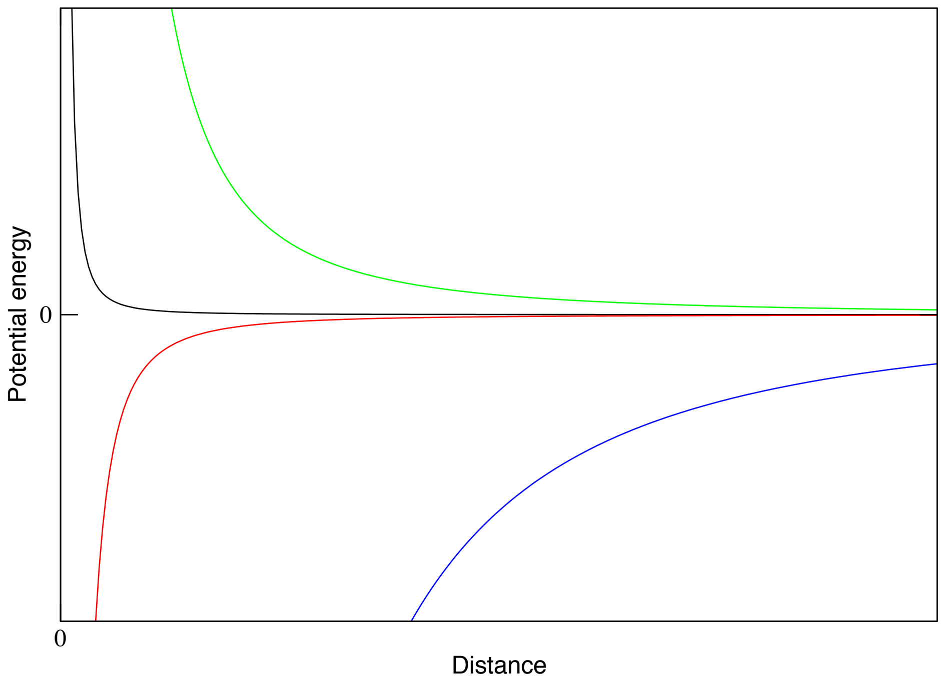


Figure 28: The Coulomb potential, for several charge combinations (Q_1Q_2), for the black line Q_1Q_2 is 1 elementary charge squared (e^2), for red Q_1Q_2 is -10 e^2, for green Q_1Q_2 is 100 e^2 and for blue Q_1Q_2 is -1000 e^2.

2.3.3 Forcefield parameterisation and usage

A forcefield is comprised of the terms described above for each atom, in each residue described; for example a carbon in a lipid chain would be represented differently to one in a benzene ring, these are known as atomtypes. The parameters for forcefields are derived from highly accurate QM calculations and experimental values. Each atomtype will be assigned charges, mass and Van der Waals terms. The strength of the interactions between a pair of atomtypes can be then calculated from these terms.

Forcefields are generally parameterised for a particular purpose, this is due to them being designed to replicate certain physical properties, as such forcefields are more accurately parameterised for certain interactions than others, leading to them working better with some systems than others e.g. the MARTINI coarse grain forcefield [148] was originally designed for use with lipid bilayers and as such the original model does not represent other lipid phases as well. This also leads to forcefields being non-interchangeable; parameters for DNA from one set cannot be used with protein parameters from another.

2.4 The Integrator

The integrator uses the potential of the particles in a system in order to update the position of those particles. It is not possible to predict the position of more than two particles interacting at an arbitrary time in the future, this is known as the three-body problem. Small iterative steps are therefore taken to estimate the positions of the particles over time.

2.4.1 Leap frog

The leapfrog integrator is the default MD integrator in GROMACS. MD integrators use potential energy to calculate the acceleration at time *t*, calculates the accelerations and updates the position at time *t* + *dt*, where *dt* is the timestep.

Two separate calculations are performed; one for velocities and a second for positions. Velocities are calculated at half timesteps and positions at whole timesteps, this allows for explicit calculation of velocities and therefore temperature. A schematic is shown in Figure 29.

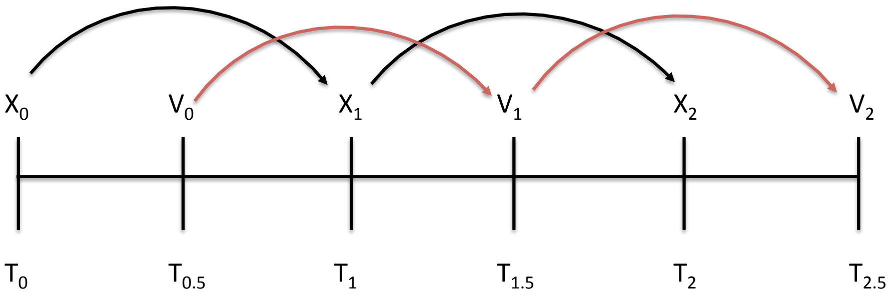


Figure 29: Leap frog integrator schematic, where *X* is the positions of the particles, *V* is the velocities and *T* is the time. The name arises from the velocity and position calculations 'leap-frogging' past each other.

The leap-frog integrator was used in all simulations.

2.4.2 The timestep

The timestep is, as previously described, the length of time the acceleration is applied for. Upon first inspection the ideal scenario appears to be having a large timestep as this would result in fewer steps and therefore less computational power to get to the required duration. Unfortunately large timesteps frequently leads to unstable systems due to atoms overlapping and a system "explosion" due to gaining a large amount of energy from Pauli repulsion. Conversely too small a timestep results in a waste of computational resource and poor sampling of conformations available. The appropriate timestep results in the simulation covering the majority of conformations, minimal computational power and realistic collisions.

Normally the timestep is chosen in order to observe the highest frequency vibration in the system. In an all-atom forcefield it is normal to have a timestep of 1 femtoseconds (fs), in order to observe 10 points along a hydrogen (atom) bond, whereas united atom forcefields can use a larger timestep of 2 fs and still capture 10 points along all remaining, non-hydrogen bonds and coarse grain systems can use timesteps of up to 40 fs as many high frequency motions are not represented.

It is possible to test the suitability of a timestep by running a simulation using the NVE ensemble, explained in more detail in section 2.6, and monitoring the total energy of the system. An appropriate timestep will result in minimal change in the energy of the system, however simulation artefacts mean that it is unlikely that energy will be completely conserved.

2.5 Periodic boundary conditions

Simulations are of limited size and even for the largest of MD simulations there is a significant percentage of atoms that are at the edge of the simulation box, it is undesirable to allow these to experience a vacuum and therefore periodic boundary conditions (PBC) are used. PBC are when an atom leaving one edge of the box appears on the opposite face, this eliminates any particles from experiencing vacuum, but does enforce the condition that simulation boxes, they must tessellate perfectly, as shown in Figure 30. Commonly used box shapes are cubic, rectangular and triclinic, however any tesselatable box would be suitable.

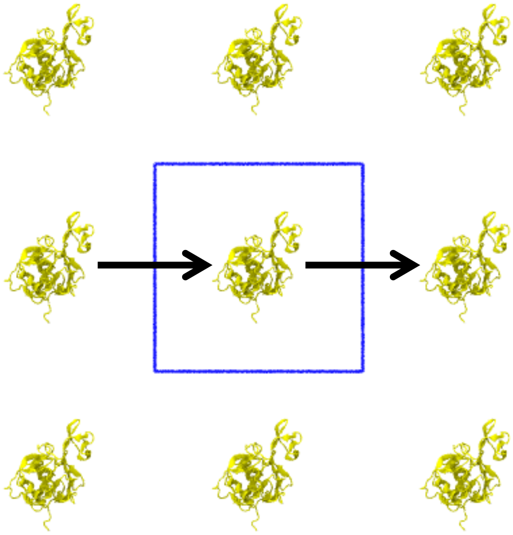


Figure 30: Periodic boundary conditions and tessellating boxes, a single periodic image is shown in blue. The arrows show movement in one periodic image occurs in all others identically, therefore as one body moves out of a periodic image it is replaced on the opposing face.

2.5.1 Cut-offs

One problem with periodic boundary conditions is the inter-molecular potentials have an infinite distance term and therefore there are an infinite number of calculations to perform. The solution is to limit the distance potentials are calculated to, beyond which they are deemed insignificant.

The simplest solution is the plain cut-off; this simply means that past a certain distance the potential is not calculated. More accurate methods are the switch and shift cut-offs. The shift cut-off moves the entire potential curve upwards, so that at the cut-off distance the potential is zero. Whereas for the switch scheme, past the cut-off the potential is linearly moved to zero and hence requires two cut-off distances, the first to stop using the normal potential and the second to set the potential to zero.

For Lennard-Jones cut-offs are used, whilst for electrostatics, despite the interactions past the cut-off being small, they can summate to a meaningful amount, in DNA simulations for example the charges of the phosphate group play an important role and cannot be ignored, even at long range. A solution is to use particle mesh ewald (PME) for charges beyond the cut-off [159]. PME uses a grid method to approximate the potential beyond the cut-off distance and has been shown to have a significant effect on the simulation, especially for DNA as the repeating monomers are charged.

2.6 Ensembles

Ensembles are the thermodynamic system the simulations are run in. The three commonly used ensembles for MD simulations are the microcanonical ensemble, the canonical ensemble and the isothermal-isobaric ensemble. These are otherwise referred to as the NVE, NVT and NPT ensembles respectively, where N is the number of particles, V, volume, E, energy, P, pressure and T, temperature.

These different ensembles are used for various reasons, for example it is normal to use NVT and NPT ensembles prior to running a final MD simulation in order to confirm the temperature and pressure are stabilised before the production run is performed, this process is called equilibration. Some free energy calculations also require the NVT ensemble as the theory only holds true under a constant volume.

2.6.1 Thermodynamic properties

Simulations are performed on small systems, even the largest simulation has only a few million atoms, as such it is difficult to represent properties on the mesoscale that are easily defined on the macroscale e.g. temperature.

The velocities of the system define temperature. On the macroscale a Boltzmann distribution of velocities exists for a given temperature, however on the mesoscale there are not enough atoms to represent such a distribution. In simulations, the temperature and velocities of the system fluctuate in order to both represent the mesoscale more accurately as well as allowing variation of the kinetic energy of the system.

2.6.2 Temperature and pressure coupling

There are several methods available to represent the correct temperature and pressure; these are referred to as different thermostats and barostats respectively. Thermostats change the velocities of the particles in the system, whereas barostats alter the box size.

The Berendsen thermostat is a weak-coupling method, which treats the temperature as though it was coupled to an external bath [160]; this causes exponential decay towards the target temperature. Unfortunately it has been suggested that the Berendsen method does not correctly represent the thermodynamic ensemble, therefore it is not suitable for production runs but is useful for equilibrations as it allows for larger fluctuations in temperature than other thermostats.

A modification to the Berendsen thermostat, which causes it to correctly create the NVT ensemble, is called the velocity-rescaling (V-rescale) thermostat [161]. V-rescale adds a stochastic term to the Berendsen thermostat, which causes fluctuations in temperature, which are expected of the mesoscale. The Nosé - Hoover thermostat allows variations in temperature by having a heat bath in the Hamiltonian with an additional degree of freedom, essentially creating a frictional term [162, 163].

The Berendsen barostat instantaneously scales the box size every timestep, leading to the suggestion that it does not create the true NPT ensemble. Like the Berendsen thermostat, it is very useful for equilibration purposes. The Parrinello-Rahman barostat alters the box vectors over time, which suggests it better represents the true NPT ensemble [164, 165].

Pressure coupling can be applied in several ways, isotropic pressure coupling, semi-isotropic or anisotropic. Isotropic coupling is when all axes (x, y, z) change by the same amount, semi-isotropic is when the x and y axes are linked whilst z is coupled separately whereas anisotropic pressure coupling has all three axes independently coupled. Isotropic coupling is commonly used in simulations in bulk water, whilst semi-isotropic is logical for systems that are expected to have different a compressibility in the x/y plane in comparison to z e.g. lipid bilayers in solution. Anisotropic has been suggested to encourage the formation of bilayers but can be unstable, resulting in unusual changes to box size [166].

Chapter 3 Sequencing DNA through αHL using an exonuclease enzyme

Abstract

Protein nanopores currently appear to be the most likely method through which nanopore sequencing can be made into a working system in the near future. Indeed Oxford Nanopore Technologies have released a sequencer based on αHL that is currently in a public beta test. DNA sequencing using an exonuclease enzyme, is where strands of DNA are sequentially split into monomers and sequenced, has been shown to be a viable technology, but a working device based on this technology as yet, does not exist. Key benefits of this technology in comparison to strand sequencing are that the secondary structure of the DNA is no longer relevant and that monomers translocate slower than strands. We therefore performed simulations in order to investigate how the location of mononucleotide release corresponds to successful capture by the protein.

We show that probability of entry into the αHL vestibule is highly dependent on distance of release, with a rapid decrease in probability of capture as distance increases. We also show how translation of the monomer away from the centre of the vestibule entrance results in decreased probability of capture, although at a slower rate than increasing distance. Notably, small translations of the mononucleotide increase the rate of capture, by reducing the distance of the nucleotide away from the vestibule lining residues.

Chapter 3

Notes

The αHL protein in the DMPC membrane was created by Dr Tomasio, she also ran several preliminary simulations, none of which are included here.

This chapter has been published in Nanomaterials under the title “The nucleotide capture region of alpha hemolysin: insights into nanopore design for DNA sequencing from Molecular Dynamics simulations”.

3.1 Introduction

Exonuclease sequencing is a potential technology, that works by cleavage of the DNA strand into individual monomer units, which are then sequenced as shown in Figure 31. The main focus in this investigation is on the distance required for adequate capture of nucleotides and entry into the vestibule, which is the solvated space inside the αHL cap region, as it has previously been shown that the technique is able to accurately differentiate between the 4 mononucleotides [167].

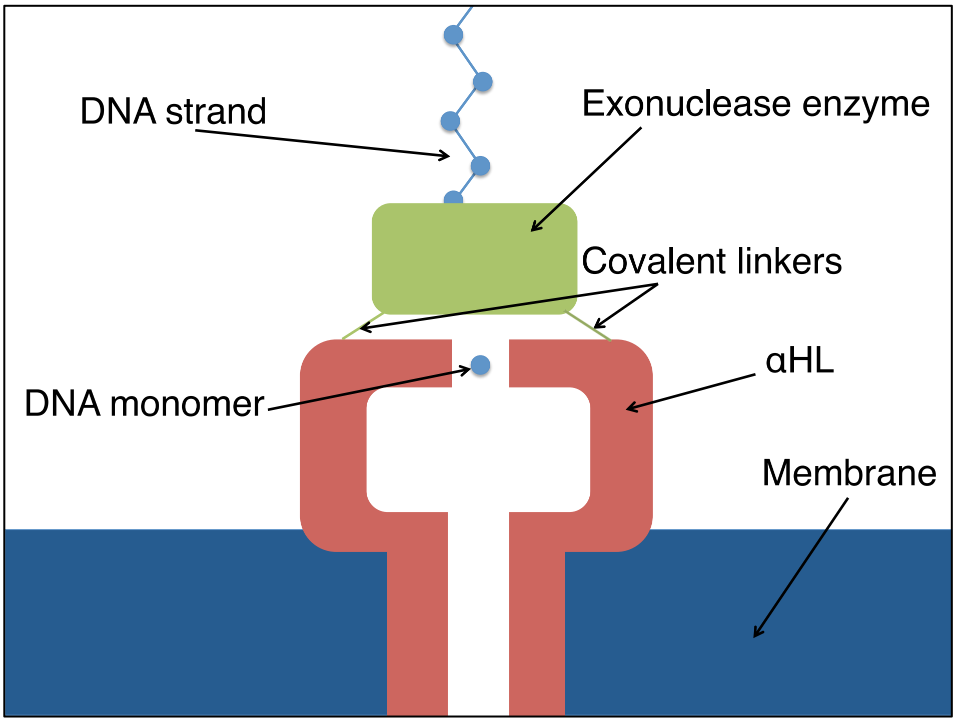


Figure 31: A schematic overview of exonuclease sequencing.

Many exonuclease enzymes are known, previous studies into DNA sequencing used Exonuclease I from *E. Coli* [167] shown in Figure 32. This enzyme progresses along the DNA strand from 3’ to 5’ and liberates 275 mononucleotides a second, or approximately one every four milliseconds.

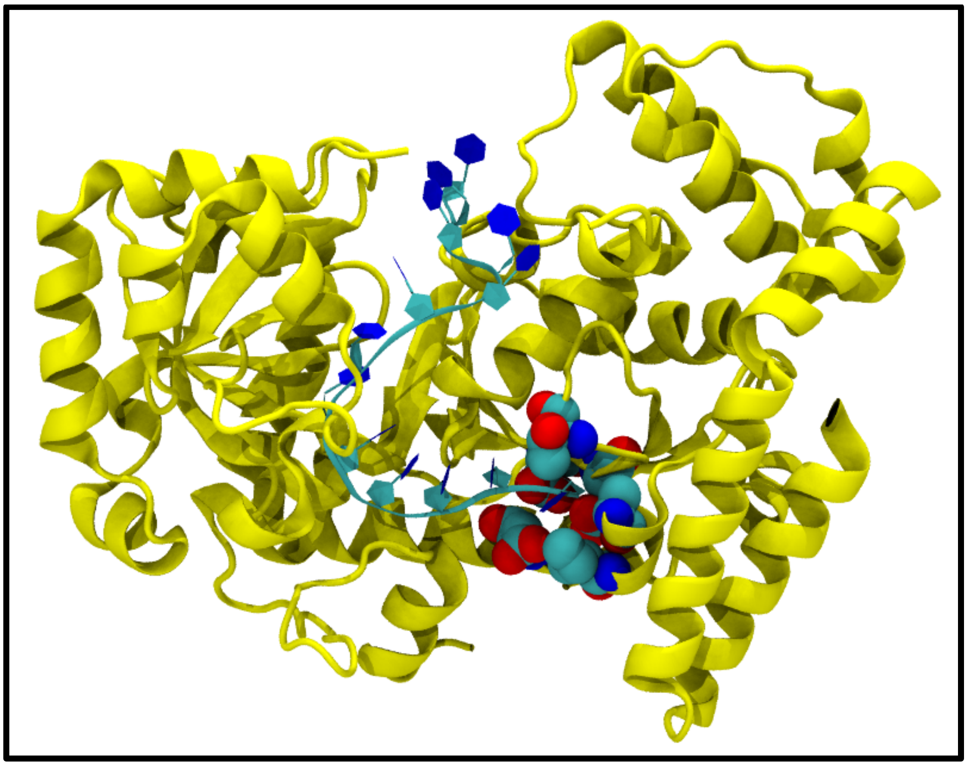


Figure 32: Exonuclease I from *E. Coli* in complex with a DNA strand taken from PDB 4JRP [168]. The residues in the active site are shown as Van der Waals spheres.

There are several differences in the results that would be observed between exonuclease and strand based nanopore sequencers. Importantly, exonuclease enzymes allow for individual mononucleotides to enter the αHL pore, whereas a strand sequencer has approximately ten nucleotides in the beta-barrel at a given time [63]. This allows exonuclease based systems to have an increased resolution of individual nucleotides compared to strand system as only one base is in the pore at a time, however no re-reading of nucleotides can occur at a secondary constriction [169], so if a base is misread, it cannot be identified. As mononucleotides are used in exonuclease sequencing, it is not possible for secondary structures to form due to charged residues, as is seen with DNA strands [126], this allows mutations to be used which may be unsuitable for strand sequencing.

3.2 Methods

3.2.1 DNA parameters

The DNA parameters used are based on those from the GROMOS3a6 forcefield [158] as it is better parameterised for ssDNA than others [109, 126], however this forcefield does not include any parameters for the terminal ends. As such it was decided to extend the changes by Guy et al [126] where the terminal end parameters were based on those in RNA from the same forcefield, to produce mononucleotides, as this also will maintain consistency with previous computational work done on αHL.

3.2.2 Exonuclease setup

The full αHL protein was used in place of the model pores previously used within the group [98, 126] since the primary focus of these simulations is to observe entry into αHL, and therefore key interactions will be with the entrance to the vestibule.

The initial system of αHL in the 1,2-dimyristoyl-sn-glycero-3-phosphocholine (DMPC) lipid membrane was setup and equilibrated by Dr Tomasio. In all simulations, beyond the equilibration of αHL in the membrane, the phosphate of the lipid head group was restrained using a harmonic potential of 1000 kJ.nm^{-1} to prevent movement of the protein.

The cytosine mononucleotide was released at distances of 1.0, 1.5, 2.0, 3.0 and 4.0 nm from the pore lumen. The pore lumen is defined as the average K8 Cα z coordinate from the centre of mass of the mononucleotide, shown below in Figure 33. Further simulations were setup in order to investigate the influence of translation of the release point away from the centre of the pore. Simulations were setup at a distance of 1.5 Å and then translated by 0.5, 1.0, 1.5, and 2.0 nm.

For each of these points, the mononucleotide was released in 2 orientations, the phosphate group pointing either towards or away from αHL, referred to as the “down” or “up” orientations respectively to investigate if, like transport through other proteins, the orientation of the substrate is important [170, 171]. Twenty repeats simulations of each location and orientation were performed, meaning each simulation accounts for 5% of the data.

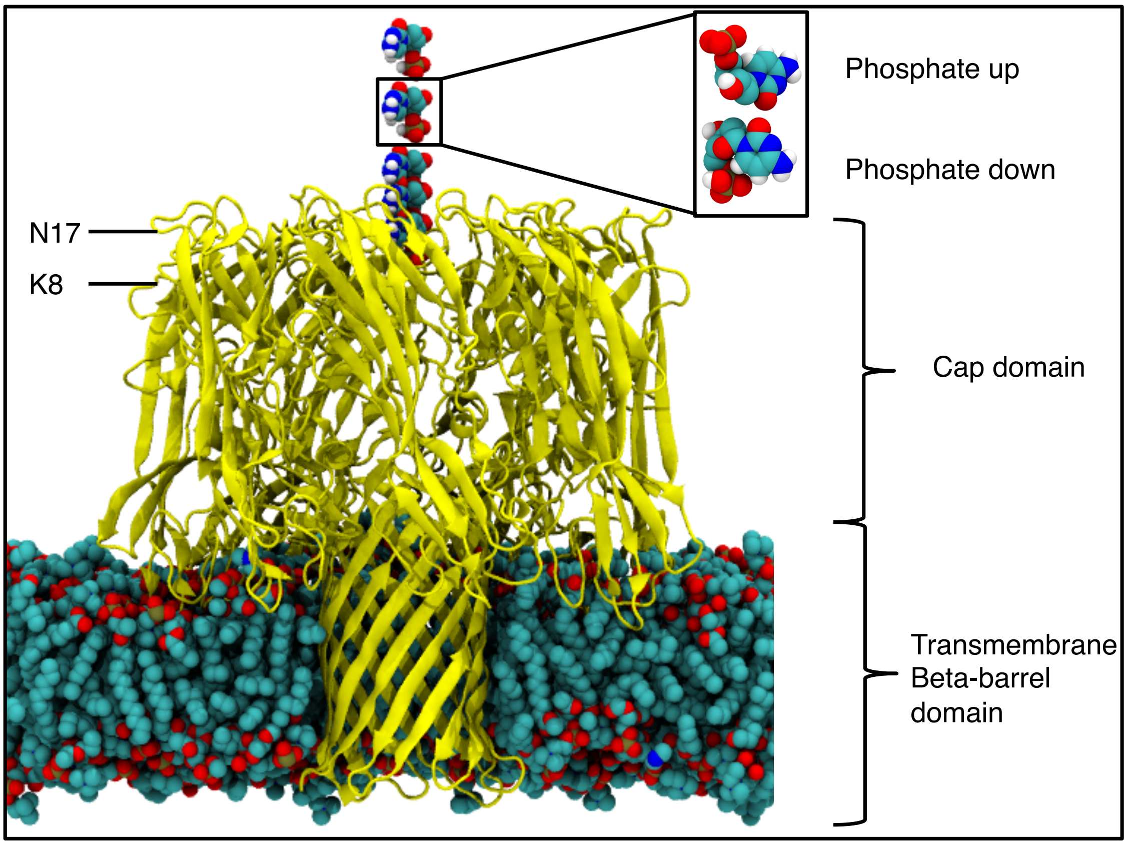


Figure 33: Schematic of the simulation setup, with the distance that the mononucleotide was released at varied, for the phosphate in the "up" orientation. Each mononucleotide is a separate simulation, waters and ions omitted for clarity.

Systems were equilibrated under the NPT ensemble for 5 ns, then with an electric field of 0.1 V.nm^{-1}, equivalent to 350 mV across the membrane, for 5 ns prior to the production run of 10 ns, with 20 repeats per distance and phosphate orientation. The mononucleotide was restrained using a harmonic potential of 1000 kJ.nm^{-2} during equilibration. All systems were run using the Parrinello-Rahman barostat [165] and the v-rescale thermostat [161] at 1 bar and 310 K. Brownian dynamics was not used, as this reduces the influence of starting coordinates, which was the subject of this investigation.

3.3 Results

Three terms are defined for the results of simulations: capture, possible capture and failed capture. Capture of the cytosine mononucleotide is when all atoms of the mononucleotide are below the ring of N17 C_α atoms as over all simulations no mononucleotides were observed exiting the pore from this region. Possible capture is when parts of the mononucleotide are above the ring and others below. Failed capture is when the entirety of the mononucleotide is above N17.

The system was treated as a binomial distribution; simulations either succeeded in nucleotide capture or failed. Error bars were calculated using the standard deviations, calculated from Equation 10.

$$\sigma = \sqrt{np(1-p)}$$

Equation 10: The standard deviation of a binomial sample, where σ is the standard deviation, n is the number of samples and p is the probability of success.

3.3.1 Distance

The results of varying the distance from K8 are plotted in Figure 34. As expected, the further from the entrance to the vestibule the lower the chance of entry. Increase of the mononucleotide release distance results in a near linear decrease in entry into the protein vestibule. The orientation does not appear to influence the probability of pore entry, as the curves largely remain within error.

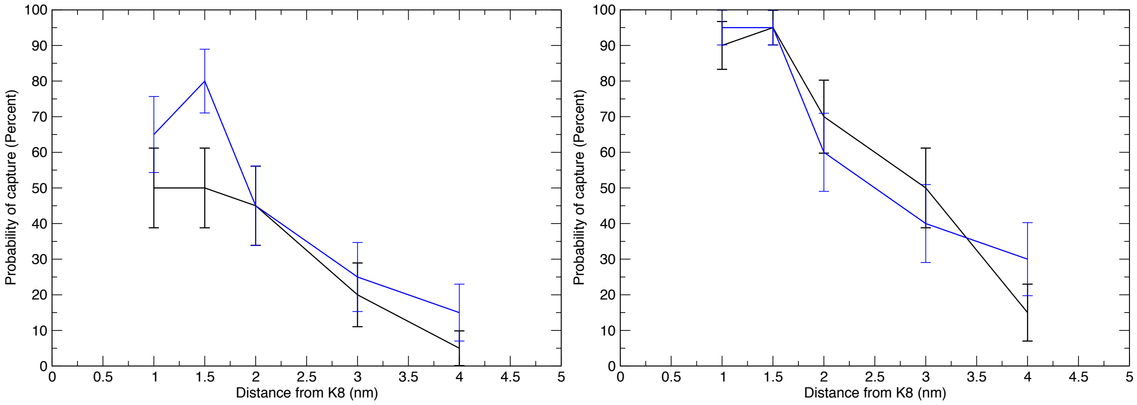


Figure 34: The relationship between distance and (left) capture of the mononucleotide and (right) possible capture of the mononucleotide, with the black line presenting the phosphate up orientation and the blue line the phosphate down orientation.

3.3.2 Translation

As expected, the further the mononucleotide is translated from the centre of the pore, the lower the probability of entry, as shown in Figure 35. It appears that small translations, e.g. 0.5 nm, result in an increase in probability of entry. We theorise this is due to a decrease in the required lateral diffusion to interact with the edge of the pore, as the mononucleotide is now released above the edge of the vestibule entrance instead of over the centre of the pore.

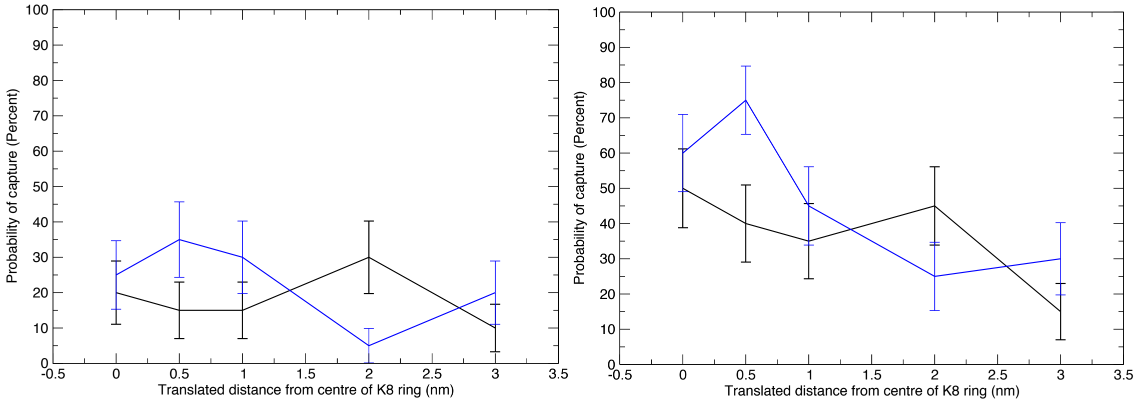


Figure 35: The relationship between translation and (Left) capture of the mononucleotide and (Right) possible capture of the mononucleotide, with the black line presenting the phosphate up orientation and the blue line the phosphate down orientation.

3.3.2.1 Interactions with αHL

Unlike OccD1 [170] or OprP [171] the orientation of the mononucleotide does not noticeably influence the observed interactions with the protein; there are binding modes observed for the nucleobase, phosphate group and the hydroxyl group on the sugar. Using an in-house script and the visualisation program VMD [172] the frequency and duration of Van der Waals contacts between the mononucleotide and protein residues was calculated.

For simulations where capture was observed, the mononucleotide had extended interactions, over 5 % across all simulations, with the following residues: D2, S3, K8, T9, D13 and N293. Mostly these residues are located in the vestibule. In contrast, for simulations where possible capture was observed these residues were A1, N6, K8, T9, G10, T11, D13, G15, S16, D17, T18, T19, V20, D45, K46 and N47 all of which are located near the entrance to the vestibule. For simulations where capture was not observed the mononucleotide either did not interact with the protein at all or with residues K8, I16, N17, T18, K46 and D47. These residues are summarised in Figure 36. It appears that D45, K46 and N47 reduce the probability of capture, by providing favourable interactions outside of the vestibule.

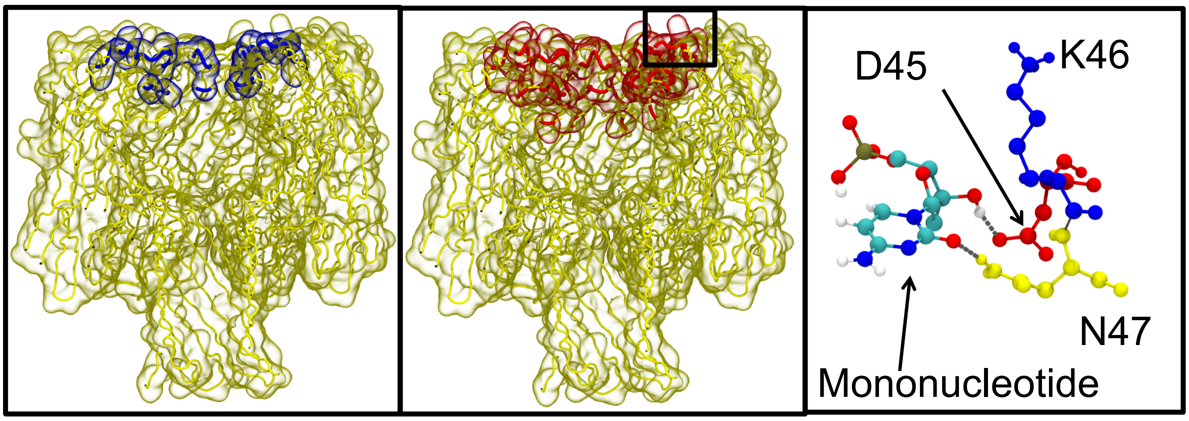


Figure 36: (Left) The blue highlighted residues associated with failed capture are predominately on the edge or surface of the vestibule. (Centre) The red highlighted residues associated with possible capture are predominantly on the edge of the entrance to the vestibule. The box in the centre image corresponds to the location of D45, K46 and N47. (Right) The most frequently observed binding mode for the mononucleotide to these residues, with dashed lines indicating hydrogen bonds, is from the sugar hydroxyl to the sidechain on D45 and the nucleobase to the sidechain of N47.

3.3.2.2 D45A K46A N47A mutant

Based on these interactions, it was decided to run a set of 20 simulations at 2 nm from K8 for the D45A, K46A, N47A mutant to observe how removal of this interaction would effect capture. The mononucleotide was released in the phosphate up orientation. Alanine mutations were chosen to maintain chirality of the amino acids whilst removing polar sidechains. The mutant was created via MODELLER [173] followed by a 5 ns equilibration.

For the WT protein, we observe 45 % of mononucleotides released in this location are captured, with an additional 15 % that are possible captures. The mutant shows a similar ability to capture mononucleotides, with 45 % being captured and 15 % undergoing possible capture. The binding site for possible capture has now moved, with residues N17, A47 and E289 forming the pocket. N17 and A47 form hydrogen bonds to the phosphate moiety, whilst E289 hydrogen bonds to the nucleobase as shown in Figure 37.

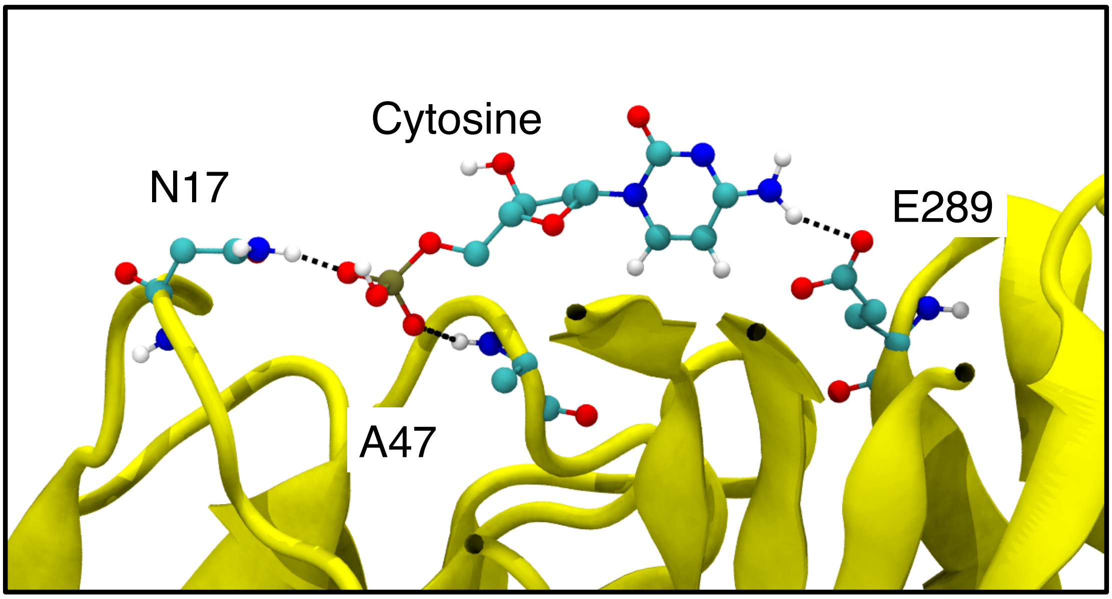


Figure 37: Hydrogen bonding between the mononucleotide and N17, A47 and E289.

3.3.3 Sum effects to mononucleotide entry into the αHL vestibule

Using root mean squares linear regression, the equations of the lines for the relationship between distance and capture probability was produced for both orientations, shown in Equation 11.

$$Phosphate\ up, capture\ probability = \ 72\ \% - (16.5\ \%.nm^{-1})$$

$$Phosphate\ down, capture\ probability = 93\ \% - (20.5\ \%.nm^{-1})$$

Equation 11: The different relationships for capture probability given (top) the phosphate up orientation and (below) the phosphate down orientation.

It is observed that even if released at K8, the location defined as zero distance, the probability of capture is 72-93 %. This suggests that nucleotide diffusion is rapid, with some mononucleotides released below N17 still unable to be captured. To illustrate the distances involved these calculated probabilities were then plotted with αHL in Figure 38.

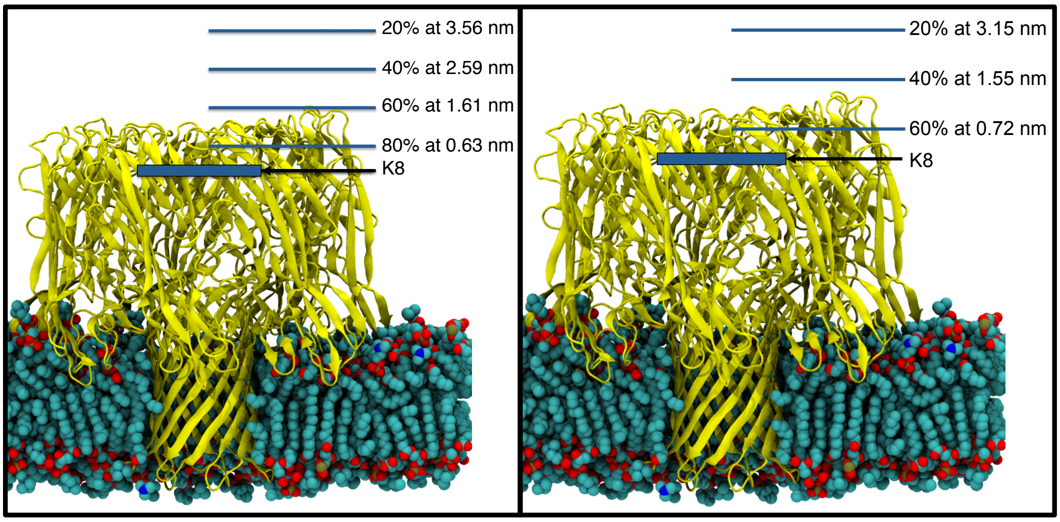


Figure 38: The probability of capture of mononucleotides by αHL is dependent on the orientation of the mononucleotide release. (Left) The phosphate down orientation and (Right) the phosphate up orientation.

For the translated simulations, the same linear regression was performed, however the probability at zero translation was set to intercept the predicted probability for the released distance, 15 Å, form Equation 11. Therefore

Equation 12 displays how the sum effects of location affect probability of capture of a mononucleotide into the αHL vestibule.

$$\begin{aligned} &Phosphate\ up, capture\ probability \\ &= 72\ \% - (16.5\ \%. nm^{-1}) - (2.3\ \%. nm\ translated^{-1}) \end{aligned}$$

$$\begin{aligned} &Phosphate\ down, capture\ probability \\ &= 93\ \% - (20.5\ \%. nm^{-1}) - (6.2\ \%. nm\ translated^{-1}) \end{aligned}$$

Equation 12: The sum effects of location on capture probability for (top) the phosphate up orientation and (below) the phosphate down orientation.

3.4 Discussion and Conclusions

As the distance of mononucleotide release increases from the entrance of αHL there is a rapid decrease in probability of pore entry. With a 1 nm increase of distance, there is a corrosponding decrease in pore entry probability of up to 20.5 %. Also, as the mononucleotide is increasingly translated away from the centre of the pore, the probability of pore entry decreases, but to a lesser extent than with distance, only 6.2 % per nm.

Translation of the mononucleotide away from the centre is beneficial when the distance is less than 1 nm, as it decreases the distance required to interact with the protein. However increasing translation beyond this minor distance results in a decreased probability of entry due to the mononucleotide preferentially binding to the surface of the cap region and not entering the vestibule.

Study of the D45A, K46A, N47A mutant pore does not show any improvement of the ability of αHL to capture mononucleotides. For further studies we would suggest mutation of E289 as well as these residues, to attempt to eliminate all binding that can occur outside the vestibule.

Based on these results, we would recommend that for the highest probability of pore entry, the mononucleotide should be released no further than 1.5 nm from the pore entrance with the phosphate group of the mononucleotide closest to the pore. We would also recommend that the exonuclease enzyme must be bound tightly to cap region, as even minor motions relative to the protein would significantly decrease capture of the mononucleotide.

This technology has problems that need to be overcome prior to widespread usage. Conventional, chemical sequencing technologies have a high accuracy (greater than 99 %) whereas so far, under the best circumstances we have observed a comparatively large failure rate of five percent, which also assumes perfect sequencing of captured nucleotides. It is possible that with appropriate amplification steps that this technology could be used, as errors could be reduced through extensive repeats.

3.5 Future work

We have not considered the effect of the exonuclease enzyme. Future simulations should investigate how exonuclease enzymes and the DNA strand associated with it influence the pore entry probability. For example, how the location of the protein influences pore entry and the ionic current. Comparisons of experimental rates of these enzymes could be used in order to select a variety of enzymes suitable, choosing only those with a slower rate of reaction than the average translocation rates of the nucleotides. It is worth noting here that due to the size of the system QM/MM is likely to be too computationally expensive and therefore the enzymatic site of the selected proteins will be unable to break bonds; therefore the simulations will have to be set up with the strand and mononucleotide already separated.

Investigations into alternative proteins could also be performed. An obvious problem with αHL is the size of the entrance to the protein, relative to the rest of the pore. MspA upon initial inspection might be better suited to capturing mononucleotides, due to the large, goblet shaped pore entrance [92] it is also likely translational effects will be less important due to the increased size of the protein entrance.

Chapter 4 Free energy of translocation through αHL

Abstract

Previous computational studies have shown that binding locations in protein nanopore can influence the secondary structure of ssDNA as it translocates [126]. Therefore in order to fully understand DNA sequencing using αHL it is important to consider the free energy changes that occur during translocation. It is known that each base has a different residence time in the αHL pore [167], leading us to assume that each DNA bases binds with different strengths to the protein. Free energy calculations have been performed to understand the molecular origins for these differences.

We show that there are several specific regions of the pore that are crucial to the interactions with DNA. Therefore we have presented logical targets for future mutations. We also show that the energetics of translocation are highly complex, with electrostatics, hydrogen bonding and hydrophobic interactions all playing key roles. These interactions can be water mediated and therefore are incredibly difficult to predict.

Notes

This chapter was produced in collaboration with Dr Guy from the Khalid group at the University of Southampton. Dr Guy started the simulations of the adenine, cytosine and phosphate potentials of mean force (PMFs), however he was not able to complete them due to limits of computational resources and time. Dr Guy's PMFs and histograms are presented in the condition they were received alongside the final results, in appendix A, to ensure that credit for the work is shown fairly, prior to the final results presentation.

This chapter has been published in The Journal of Chemical Theory and Computation, JCTC, under the title "The molecular origins of DNA base discrimination by alpha-hemolysin revealed by free-energy calculations".

4.1 Introduction

Umbrella sampling and the weighted histogram analysis method (WHAM) [174] are methods to calculate potentials of mean force (PMFs), which measure the free energy over a reaction co-ordinate, for example a distance through a membrane [175]. Before these are discussed further, the concept of free energy will be discussed and derived.

4.1.1 Free energy

The free energy of a system is defined as the amount of work the system can perform on its surroundings; therefore it is the energy ‘free’ to do work. Free energy dictates how “favourable” a process or state is and is a measure of the likelihood of that process or state being observed, with more negative energies being more favourable. Note that a favourable free energy for a reaction does not provide information on the rate of that reaction, merely if it occurs spontaneously.

The internal energy of a system consists of its potential and its kinetic energies. This is the energy required to create a system and therefore is the energy contained within it. This is trivial to calculate in MD, as it is simply taken by summing all the terms calculated by the forcefield.

$$U_{internal} = U_{potential} + U_{kinetic}$$

Equation 13: Internal energy can be split into potential and kinetic terms, *U* is energy.

The total energy of a system is due to internal energy and the energy required to displace the surroundings. The total energy is called the systems enthalpy.

$$H = U + PV$$

Equation 14: Enthalpy, *H* is enthalpy, *P* is pressure and *V* is volume.

The law of conservation of energy states energy cannot be created or destroyed, leading to the 1st law of thermodynamics, which states; *"The change in internal energy of a closed system is the difference between the heat added to the system and the work done by the system to its surroundings"* shown as:

$$\Delta U = Q - W$$

Equation 15: *Q* is heat, *W* is work.

Therefore for the internal energy to change, work must be provided from an external source. For isolated systems the internal energy is constant, as work cannot be done to the system.

The first law of thermodynamics does not consider that processes have directionality, e.g. energy spontaneously flows down a temperature gradient, whereas the reverse does not occur. We cannot explain this using the first law, as the internal energy of the system remains constant; therefore we introduce the concept of entropy. The second law of thermodynamics defines entropy.

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

Equation 16: Entropy change, *S* is entropy, *Q* is the heat to drive the process, and *T* is temperature.

The entropy of a closed system never decreases, it either remains constant if a system is at equilibrium or increases with time. Entropy is a measure of the amount of disorder in a system, e.g. a crystal is low entropy, due to being highly ordered, whilst a salt solution is a high entropy system. Another definition of entropy is the amount of energy unavailable for work; instead being associated with thermal noise. Entropy is also a measure of the number of equivalent states of the system.

As entropy is the energy unavailable for work, the free energy can be defined as the difference between the total energy and the entropy. Two definitions are commonly used, the Helmholtz free energy and the Gibbs free energy. The Helmholtz free energy is the creation of the system, in a vacuum and therefore in absence of pressure or volume changes. The Helmholtz free energy treats the total energy as the internal energy *U*. The *TS* term shows the energy available using heat transfer from the environment.

$$F = U - TS$$

Equation 17: The Helmholtz Free energy, *F*

In contrast the Gibbs free energy of a system is the energy required for the creation of the system and any energy required for the displacement of volume required by the system.

$$G = U + PV - TS$$

Equation 18: The Gibbs free energy, *G.*

However, as we can see from the definition on enthalpy, Equation 14, this can be re-written, replacing "*U +PV*" with enthalpy.

$$G = H - TS$$

Equation 19: Another form of the Gibbs free energy.

This definition is the absolute free energy for the creation of a system. However, the change in free energy is often more useful than the absolute energies as it provides a measure of energy gained or lost moving between two states.

$$\Delta G = \Delta H - T\Delta S$$

Equation 20: The change in Gibbs free energy.

A negative change in free energy indicates that a reaction occurs spontaneously, whereas a positive change indicates that the reaction requires energy to occur. Again, note that a reaction being favoured energetically does not indicate that it occurs on meaningful timescales, as the kinetics may be slow.

4.1.2 Potential of Mean Force, umbrella sampling and the Weighted Histogram Analysis Method

Potentials of mean force (PMFs) measure free energy changes of a substrate across a reaction coordinate, calculated from the force across a well-sampled system. The reaction coordinate is often a distance and a well-sampled system is one that has gone through all possible configurations. PMFs have allowed study of a wide variety of systems. For example where the reaction co-ordinate is a direction, the energetics of permeation across membranes [175, 176], ligand binding [177] or dissociation of a DNA-protein complex [178].

Umbrella sampling is a PMF method that uses multiple simulation windows across a reaction coordinate, shown in Figure 39. By restraining the substrate using a harmonic potential to one section of the reaction co-ordinate, the substrate is forced to sample all the degrees of freedom. The force required to maintain the position is recorded, allowing the relative favourability of windows to be calculated. The harmonic potential does introduce a bias into the potential, which is removed in the final step in the construction of the PMF.

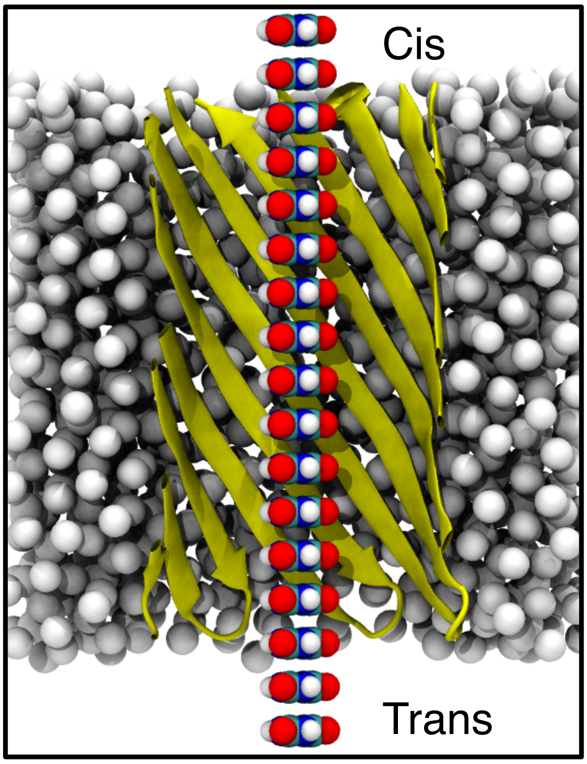


Figure 39: Thymine umbrella sampling windows showing how the position of the base is varied along the principal (z dimension) axis of the protein barrel. Only a selection of windows are shown for clarity, with water and ions are omitted. The methane slab is shown as white spheres and yellow ribbons represent the protein backbone.

The PMF of phosphate and the chloride ion through the ion channel OprP has been investigated using umbrella sampling and GROMACS [171]. The protein OprP from *Pseudomonas aeruginosa* is a transmembrane beta-barrel consisting of 16 strands, which is responsible for phosphate uptake into cells. This study was able to identify important regions for phosphate binding, as well as rationalise energetic differences between the chloride and phosphate ions. This lays the foundations for our study, using umbrella sampling and GROMACS to study the 14 stranded beta-barrel αHL.

PMFs were constructed using the Weighted histogram analysis method (WHAM) [174]. When the simulations run, they sample an area of the reaction co-ordinate, this area is shown by histograms, however the different frames may sample the reaction coordinate inconsistently, resulting in the forces not representing all areas of the reaction coordinate equally. The WHAM calculation removes this error, by weighting each frame inversely proportionally to the autocorrelation times of the energies, i.e. simulations with longer autocorrelation times are weighted less.

Sampling leads to an issue when calculating PMFs as the simulation time is always limited; incomplete sampling and data that is not converged leads to errors in the produced PMF [179]. However there are several methods for checking the system produces converged results.

Inspection of the histograms and the bulk water levels are computationally inexpensive methods of checking the convergence of PMFs [174]. Firstly, if the histograms do not cover the entire reaction co-ordinate, the system requires additional simulation in the regions of poor coverage. If there is insufficient sampling in a region then the WHAM calculation cannot produce an accurate PMF, normally seen as large, sharp changes in energy. For a system such as a membrane or nanopore with water on both sides, it is expected that the energy of the bulk water will be identical on either side of the membrane/ nanopore; therefore discrepancies reveal a lack of convergence in the PMF.

More rigorous methods to inspect convergence are inspection of the autocorrelation times, bootstrapping and block analysis [180]. The simulation length, per frame, should be several times longer than the autocorrelation time, to ensure adequate sampling. Bootstrap analysis is when a limited amount of the available data is taken and a profile constructed. Bootstrapping is repeated numerous times, which allows for statistical measures such as standard deviation and mean to be calculated. These statistics can then be used to estimate the error in the PMF. Block analysis is when the data is split into sections of time and compared, for example the first quarter, the first half and the first three quarters of the simulation, these are then analysed separately. When the WHAM calculation is performed on these blocks, several PMFs are produced; these are then compared in order to confirm the substrate is not trapped in a conformation for an extended amount of time.

4.2 Method

Due to the long expected autocorrelation times of the nucleic acids it was decided to split the DNA into fragments, the phosphate and the nucleobases, to allow converged PMFs within realistic timescales with current computational resources. The phosphate and the nucleobases were simulated as the phosphates charge will likely be the only moiety affected by charges, whereas the base will provide the unique PMF profile for each mononucleotide. We chose not to simulate the ribose sugar, as doing so would involve either repeating atoms previously simulated, such as the oxygen in the phosphate, leading to the conclusion that the fragments would not be additive, or alternatively those atoms would not be repeated, leading to a poor representation of the sugar fragment.

The fragments chosen were represented as $H_2PO_4^-$ and nucleobases with an additional hydrogen, where the bond to the ribose sugar would normally be. The phosphoric acid form chosen was the $H_2PO_4^-$ structure, as this is the dominant form at neutral pH, with approximately 70% taking this structure due to the pKa of the second hydrogen being 7.2. Dr Guy created the structures by using the nucleotide parameters as a reference and adjusting the hydrogen charge, so that the total charge was zero, or minus one for the phosphate.

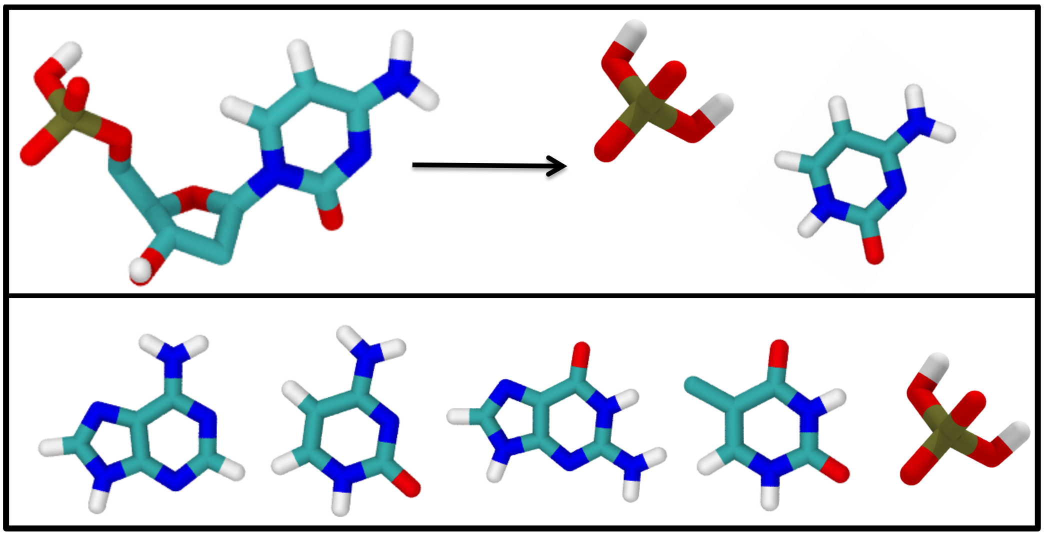


Figure 40: The nucleotide fragments studied. (Top) A cytosine mononucleotide is fragmented into in the phosphate fragment and the cytosine nucleoside. (Bottom) from left to right Adenine, Cytosine, Guanine, Thymine and Phosphate fragments.

In order to reduce the complexity of the system and increase the available sampling, the model pore system described Bond et al by was used to represent αHL [98]. In this model pore, only the beta barrel is simulated, and the lipids are replaced with methane, in order to reduce the atom count and allow access to longer timescales without significant loss of accuracy. Figure 41 details the inwards pointing residues within the αHL beta-barrel and the z-coordinates they correspond to in the model pore. Note that at the top of the pore, E111, M113 and K147 form the narrowest region in the pore. The forcefield used was GROMOS 53A6 [109, 158]. The WHAM calculation was carried out using g_WHAM [181].

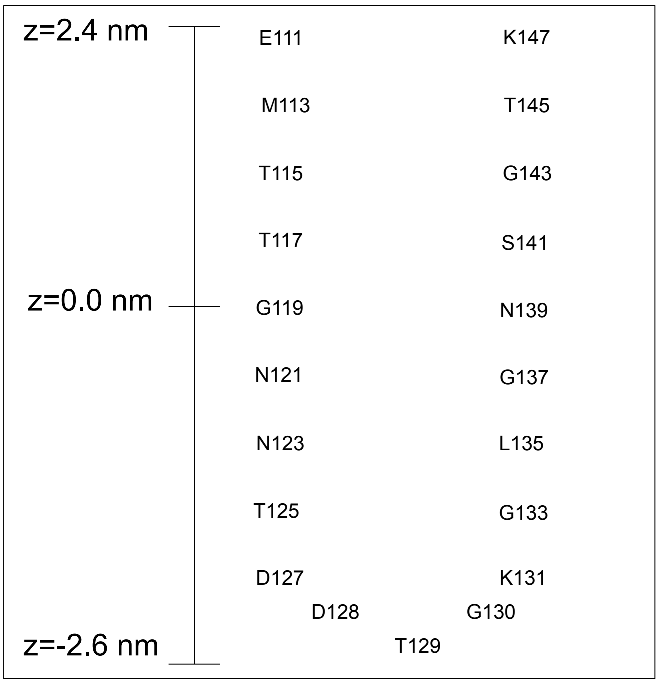


Figure 41: The sidechains of importance within the αHL pore, residues pointing outside the pore were simulated as glycine and are not shown here for clarity.

Each frame for the αHL PMF was spaced 0.05 nm apart. Each frame was run for 250 ns. The harmonic potential was applied with a force constant of 1000 kJ.mol^{-1}.nm^{-2} restraining the centre of mass of the nucleic fragment in the z-axis, whilst allowing free sampling of the xy plane.

During the project, IRIDIS4, the University of Southampton’s new supercomputer was built; this allowed access to a much-improved computational resource. It was decided to produce a PMF of a mononucleotide to compare to the fragment PMFs. This system was setup identically, except with a spacing of 0.025 nm between windows and a force constant of 2000 kJ.mol^{-1}.nm^{-2}. The thymine PMF windows were also set up using this method.

In the WHAM calculation, the profile was calculated from the pull force, using bins equal to the number of PMF windows and with the first nanosecond discarded for temperature and pressure equilibration, where bins are the amount of histograms the data is split into. The simulations were run for 250 ns at 310 K and 1 Bar. The pressure and temperature were kept constant using the Parrinello-Rahman barostat [164, 165] and Nosé-Hoover thermostat [163].

4.3 Results

4.3.1 Convergence analysis

As previously discussed, it is important to confirm the systems are sufficiently sampled prior to analysis. Here the results of the autocorrelation times, histograms and bootstrap analysis are presented. Note here only guanine is shown, which is representative of all simulations. For the analysis of the remainder see appendix A.

The histograms, shown in Figure 42, show the minimum sampling at any region is 10,000 counts and the average is in excess of 20,000. Each count signifies a snapshot where the guanine was in that region, therefore the histograms show sufficient sampling of the principle axis.

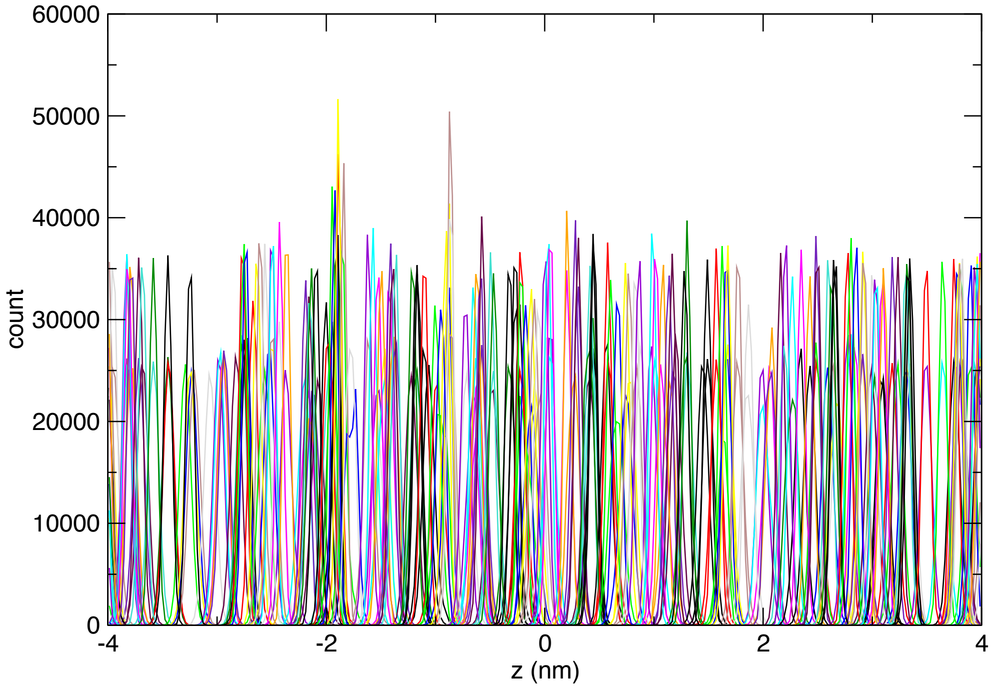


Figure 42: Histograms for the guanine base PMF through αHL.

Figure 43 shows that the autocorrelation times of the free energy, as calculated by g_WHAM, are all below 12.5 ns, with only 7 frames where it is above 5 ns. As the simulation length is 250 ns all the autocorrelation times have all been covered extensively. The largest observed autocorrelation times are generally near the pore cis entrance and trans exits, where the fragment has an interaction with the protein but is free to sample the bulk water.

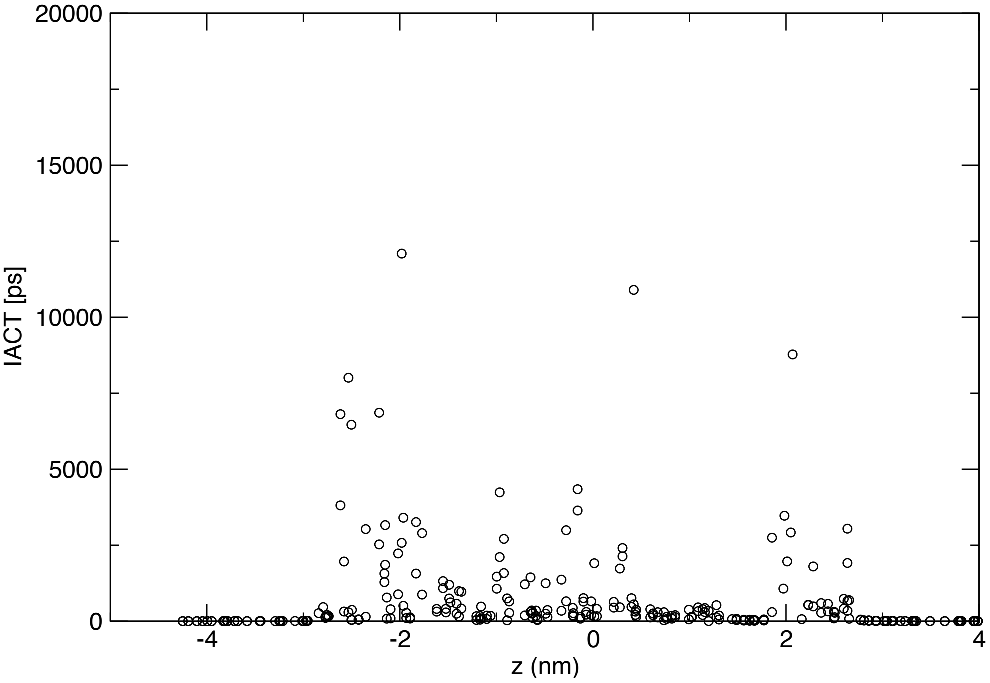


Figure 43: The autocorrelation times for the guanine base PMF.

The bootstrap analysis is shown in Figure 44, as calculated by g_WHAM. There are negligible errors in the bulk water, and errors of ±1.05 kJ.mol^{-1} in the protein. As thermal error ($k_B T$) is 2.68 kJ.mol^{-1} at 310 K the simulation is sufficiently converged.

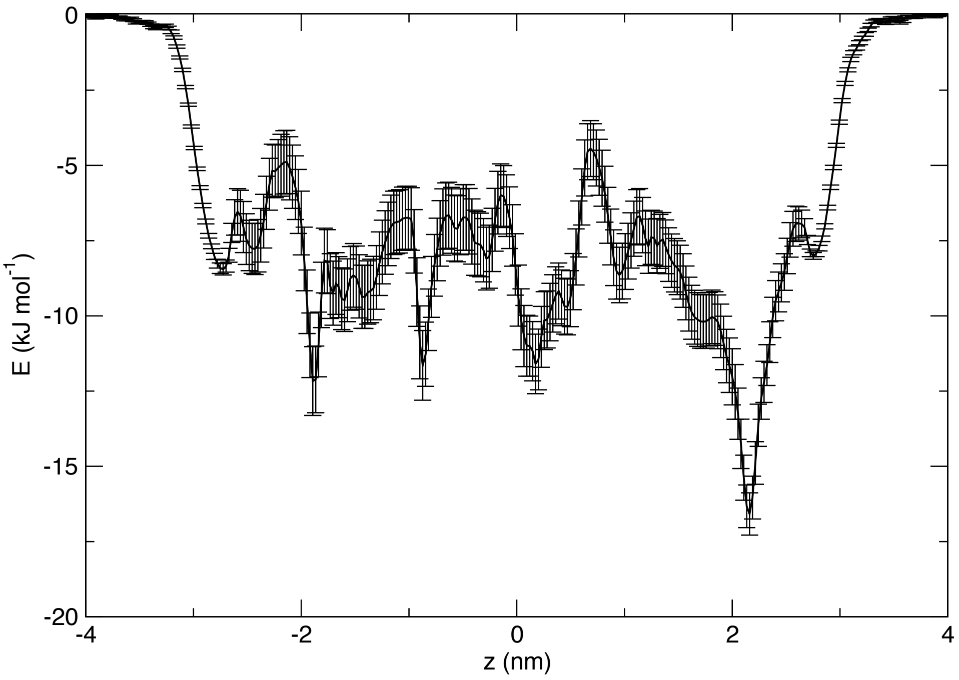


Figure 44: The bootstrap analysis for 500 bootstraps, of the guanine PMF.

This analysis shows that the PMFs have been run for a sufficient duration to be considered converged. Therefore further study can be performed to explain the origins of the features of these PMFs.

4.3.2 αHL Guanine nucleobase

The potential of mean force for the guanine nucleobase through αHL is shown in Figure 45. The details of the guanine PMF will be discussed first, then compared to the other nucleobases. The fragment always produces a favourable interaction with the pore, especially at the narrowest point, suggesting the most important interactions experienced are hydrophobic in nature.

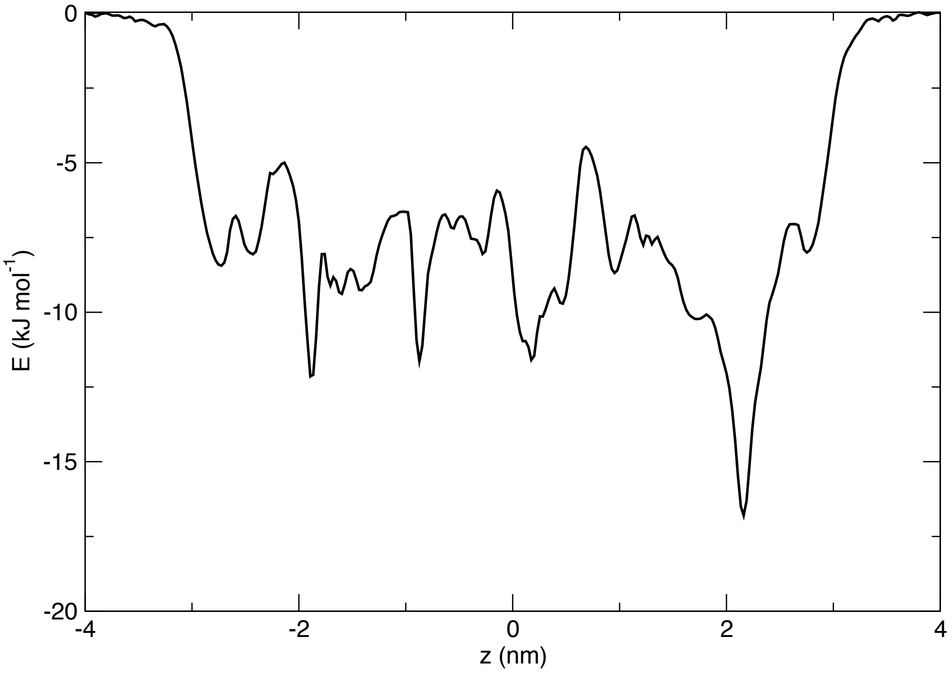


Figure 45: Guanine PMF though αHL.

As hydrophobicity appears to be important, a comparison between pore radius and energy was made, shown in Figure 46. The minimum pore radius was measured using the program HOLE [182]. HOLE measures the maximum size sphere that can fit in the pore, without overlapping with the Van der Waals radius of amino acid sidechains, at points along the principle axis of the protein. It is worth mentioning that the HOLE analysis does not wholly answer the question of solvation, as there are pockets that reduce the fragments exposure to water, as well as hydrophilic or hydrophobic residues altering the local solvation.

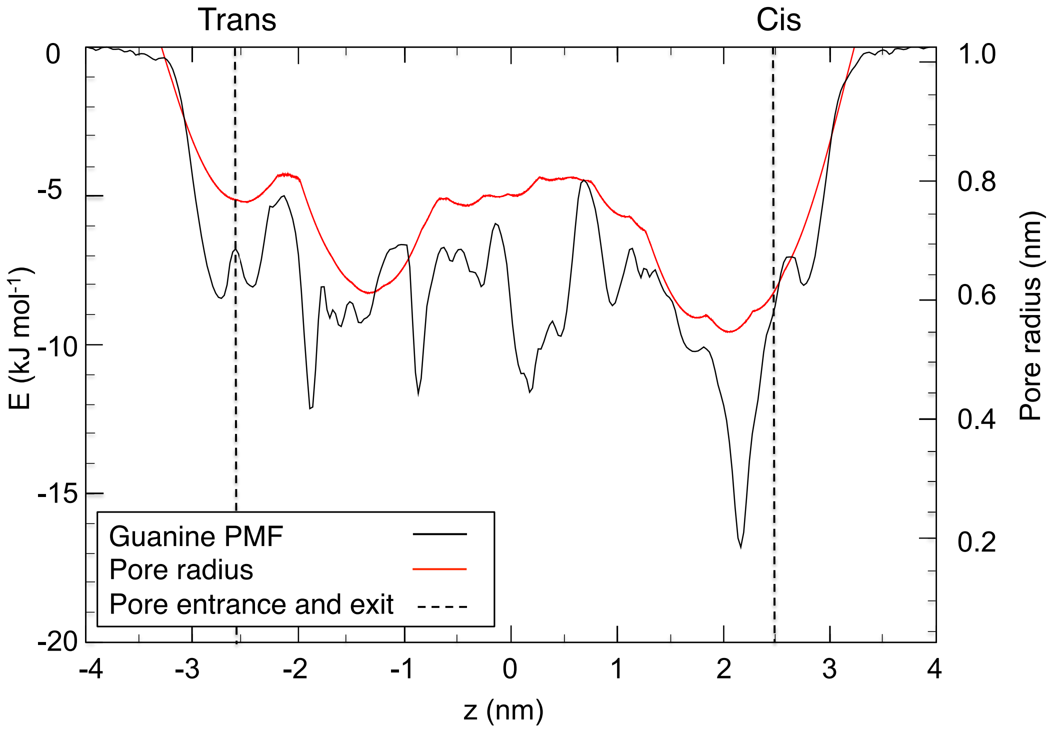


Figure 46: Pore radius compared to guanine PMF energy. The black line is the guanine PMF and the red line is the pore width as measured by HOLE.

Comparison of the pore radius analysis against the PMF, shown in Figure 46, reveals some correlation between the two plots; reduction in pore radius corresponds to a decrease in free energy, this is most noticeable when approaching and exiting the pore, as well as at the narrowest point in the pore, z=2.4 nm. Similarly, when the pore radius increases there is a corresponding increase in free energy, this is apparent at z=0.6 nm, where the pore is at its widest. As shown in Figure 41 this region corresponds to G143 and T115 residues and therefore is less favourable partially due to a reduction in dispersion interactions compared to the remainder of the pore.

However, there are many regions in the PMF profile with highly favourable energetics that do not correspond to a decrease in pore width: z=-2 nm, z=-0.9 nm, z=0 nm to z=0.5 nm, z=1 nm and z=2.4 nm for example. These regions are discussed in more detail below.

The trough at z=-2 nm is due to the interactions with the pore exit residues: T125, D127, D128 and K131. There are two interactions observed for the aspartic acid residue D127 here, it either forms a salt bridge with K131 or hydrogen bonds to guanine. The nucleobase interacts with D127 or D128 for 75% of the 250 ns simulation, an interaction is defined as a distance of less than 0.3 nm. Each hydrogen bond is short lived, with durations of less than 15 ns, whilst one to four hydrogen bonds exist at any given time. These interactions are summarised in Figure 47. The water in this region is mobile, freely alternating between hydrogen bonding with the nucleobase, the protein and itself. There are on average 20 water molecules within 0.4 nm of the nucleobase. Therefore, in this region, the energetic favourability is explained by hydrogen bonding with D127, D128 and water.

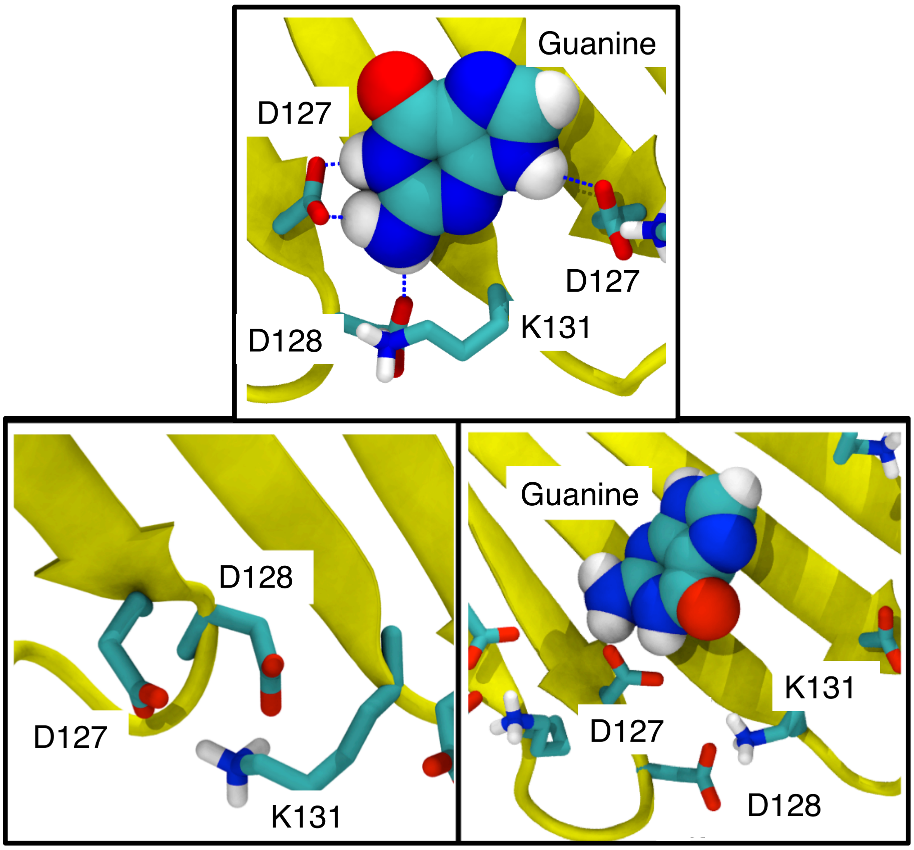


Figure 47: The available binding modes at z=-2 nm. (Top) Four simultaneous hydrogen bonds between guanine, D127 and D128 are shown as dashed lines. (Bottom left) A salt bridge formed between D127, D128 and K131. (Bottom right) D127 rotating away from the exit in order to hydrogen bond to the guanine nucleobase.

The reduction in energy at z=-0.9 nm is due to N123, N121 and L135 residues. In this region they form a pocket, which fits the 5 membered ring of the purine. The stabilisation in this pocket appears to result from hydrophobic and dispersion interactions, due to the shape complementarity of the pocket and the nucleobase as shown in Figure 48. Commonly whilst in the pocket no hydrogen bonds are observed between the nucleobase and either the water or the protein. We predict due to the location of this interaction on the nucleobase, that this interaction would not be observed on a mononucleotide due to the steric clash of the ribose sugar with the sidechains.

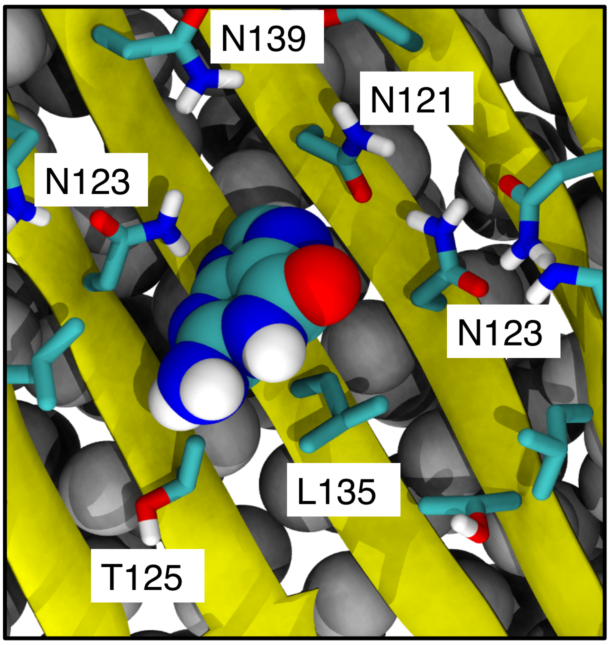


Figure 48: How in the region z=-0.9 nm shape complementarity allows guanine to fit in the pocket formed by N121, N123 and L135.

For z=0.0 nm, N121 and N139 from hydrogen bonds to the nucleoside. We also observe water mediated hydrogen bonding to N121 and N123 as shown in Figure 49. The base will preferentially reside between the sidechains of N139, parallel to the backbone of the beta barrel with the oxygen pointing into the centre of the pore. Despite this, in this region the nucleobase freely alternates between the subunits of the heptamer as shown in Figure 49, again this further demonstrates that the simulations show sufficient sampling.

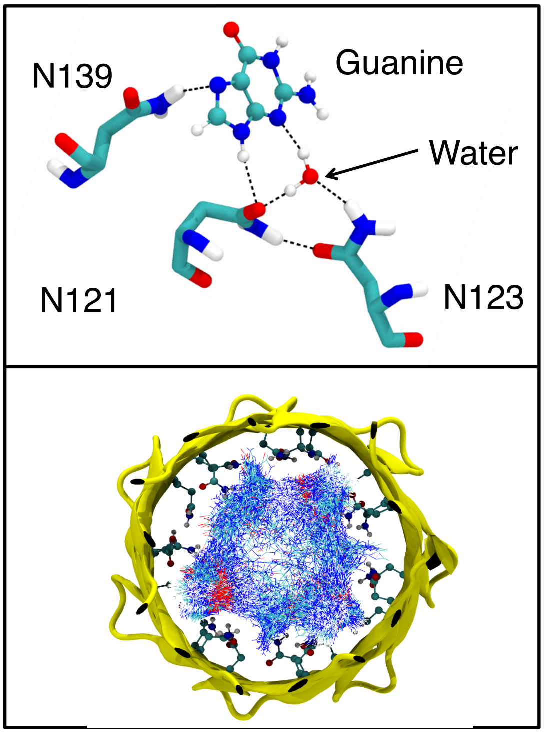


Figure 49: At z=0.0 nm, (Top) how N121 and N139 directly hydrogen bond to the guanine, whilst hydrogen bonding via water occurs to N123 and N121. (Bottom) Every simulation frame shown simultaneously from the cis compartment demonstrates good sampling of the heptamer with N121, N123 and N139 sidechains shown.

The strongest interaction is at z=2.4 nm, where K147, M113 and E111 form a constriction in the pore. It has been shown previously that this is the αHL read site [169], due to favourable hydrogen bonding and electrostatics caused by the ion pair, K147 and E111 [126]. Studies have shown that the E111N/K147N protein has a weakened recognition of bases compared to the WT protein [169]. Whilst mutation of M113 can result in increased resolution between bases [183], which emphasises the importance of this region for sequencing. Hydrogen bonds are observed between the nucleoside and K147, E111 in a similar style to Watson-Crick base pairing, as shown in Figure 50. It is possible that for a mononucleotide K147 would preferentially form a salt bridge with the phosphate moiety, however, due to the heptemeric nature of the protein it is likely both would occur simultaneously. Other binding modes were also identified, for example the amine group of the guanine forming two hydrogen bonds to E111.

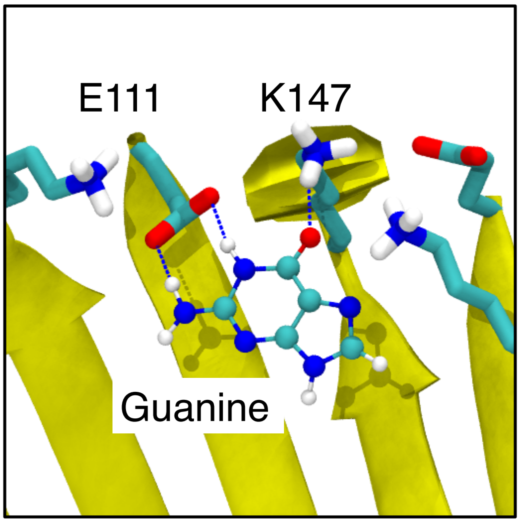


Figure 50: The hydrogen bonding at z=2.4 nm often occurs in the Watson-Crick positions of the guanine. Other sidechains are ignored for clarity.

4.3.3 Comparison of αHL nucleobase PMFs

The nucleobase fragment PMFs through αHL are shown in Figure 51. Initial inspection reveals that the shape of each nucleobase profile is very similar. Excluding guanine all profiles generally are within thermal error, 2.6 kJ.mol^{-1} at 310 K, of each other throughout the protein. As expected, the purines profiles are comparable in shape, as are the pyrimidines. The most notable difference between the purine and pyrimidines is the trough at z=-0.9 nm, which as previously discussed arises due to the five membered ring of the purines accessing a pocket formed by N139, N121 and S141. This feature is likely not observed for the pyrimidines simply due to their lack of a five membered ring. Each profile is now compared to guanine in turn.

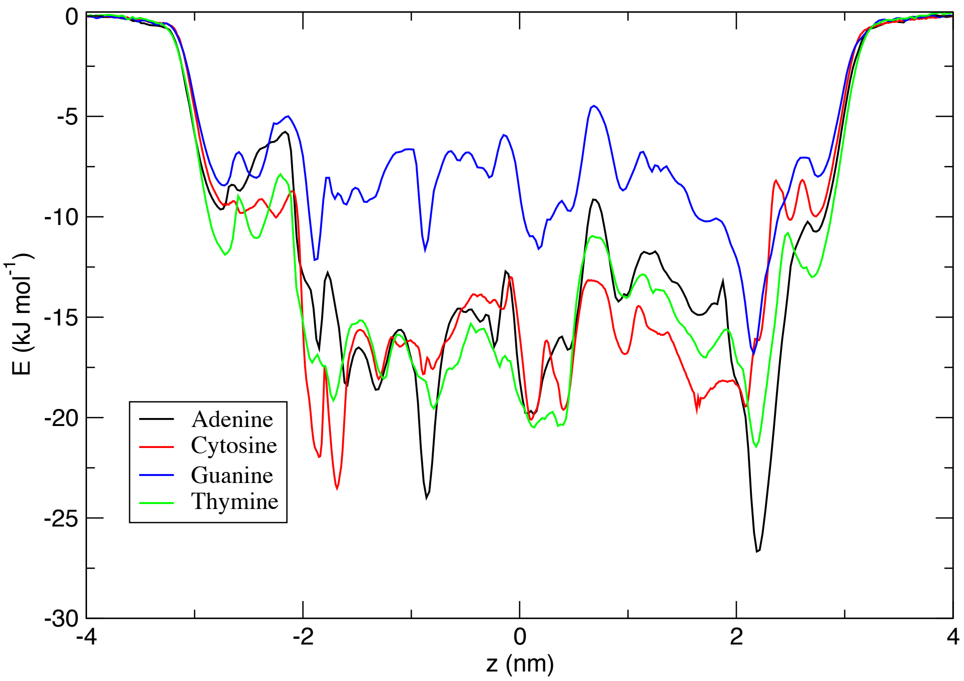


Figure 51: The PMFs of the nucleobases through αHL, with the trans exit of the pore at z=-2.6 nm and the cis entrance at z=2.4 nm.

All regions observed with guanine are present in the adenine PMF. We note adenine has the strongest interaction with the pore at the constriction, z=2.4 nm, which would explain the molecular origin of poly(A) DNA strands translocating slower than poly(C) [184]. The only major disparity between guanine and adenine is the 7.5 kJ.mol^{-1} higher energy for guanine throughout the pore. Both adenine and guanine are purines, however guanine is part of the strong base pair and has an additional hydrogen bond donor group, whilst adenine is part of the weak. It is not immediately obvious as to why its additional ability to contribute to hydrogen bonding should result in a higher energy.

For thymine the regions discussed for guanine appear to be conserved, for example there are still relative drops in energy at z=-1.9 nm, z=0 nm and z=2.2 nm. In contrast, for cytosine the regions observed for guanine are present but in general larger barriers are observed for cytosine than the other bases. We theorise that the increased barriers to cytosine are due to the harmonic restraint being applied to the centre of mass of the base, as it is the smallest base it has the least ability to manoeuvre to find favourable interactions whilst satisfying the harmonic potential. Cytosine appears to be unique in that the strongest binding site is observed at the trans exit, as opposed to the cis entrance like the other nucleobases. The reason behind this change in preferred binding site is unclear, as the hydrogen bonding patterns seem similar to those observed for the other nucleobases.

4.3.4 αHL Phosphate

The phosphate fragment, $H_2PO_4^-$, PMF though αHL is shown below in Figure 52. As expected for a hydrophilic fragment, the reduction in solvation results in unfavourable energetics across the pore.

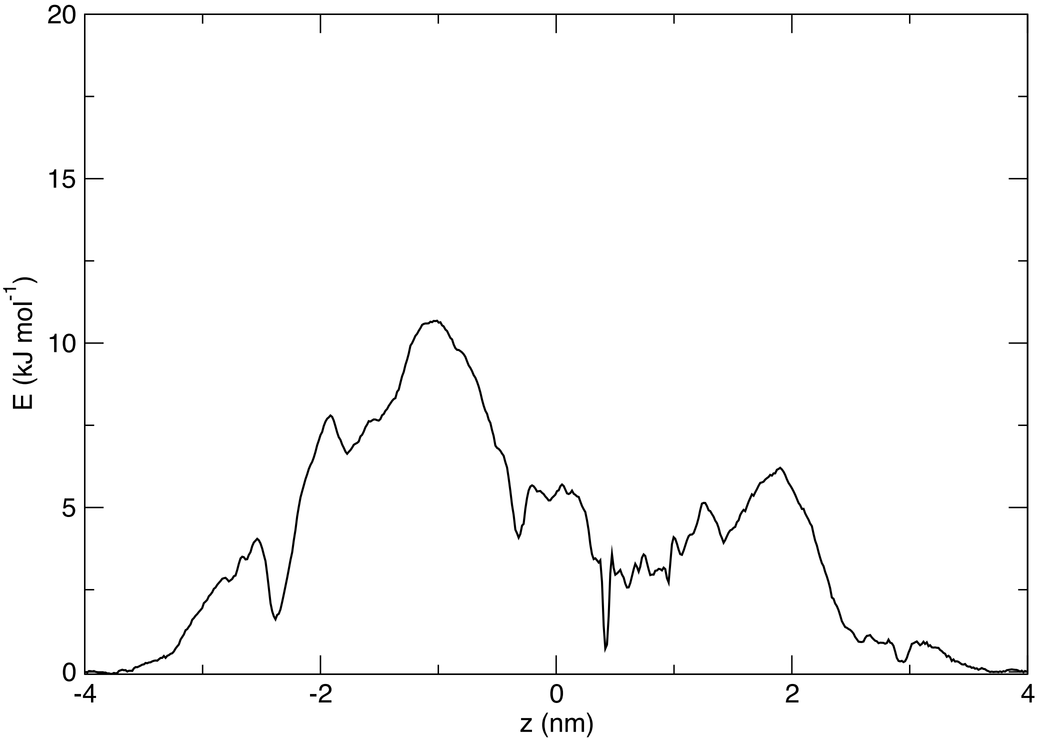


Figure 52: Phosphate PMF through αHL.

The peak in the profile at z=-2.0 nm is due to electrostatic repulsion from D127 and D128 residues. The most unfavourable region occurs at z=-1.0 nm, corresponding to L135 and N123, with G133 and G137 adjacent. In this region, aside from N123, there is a lack of hydrogen bonding or electrostatic interactions with the protein due to the surrounding hydrophobic residues, which results in destabilisation.

The most favourable region in the pore is hydrogen bonding in a pocket formed by T115, T117 and S141 at z=0.4 nm. This pocket is only accessed during one simulation, but for the entirety of the 250 ns, suggesting that the harmonic restraint for that PMF window encouraged sampling of this region, but it is not accessible during the timescale of the remaining simulations.

At the narrowest point in the pore, z=2.4 nm, there are several energetic considerations: the electrostatic effects of K147 and E111 cancel each other, whilst hydrogen bonding to K147 stabilises and a reduction in solvation destabilises the phosphate. The sum consequences are close to zero, there is no sizeable well or barrier in this region.

4.3.5 αHL nucleotide fragment PMFs

It appears from the fragment PMFs, shown in Figure 53, that if we added the phosphate PMFs to the nucleosides, all the resulting energies, excluding that of guanine, would remain at a favourable energy throughout the pore. Of course, this is not the same as simulating a mononucleotide: this is ignoring the energetics of the ribose sugar, as well as the different binding modes that would be observed to the altered sterics. Despite this, as the only differences would be on the nucleoside, energetic differences between the mononucleotides are expected to be conserved.

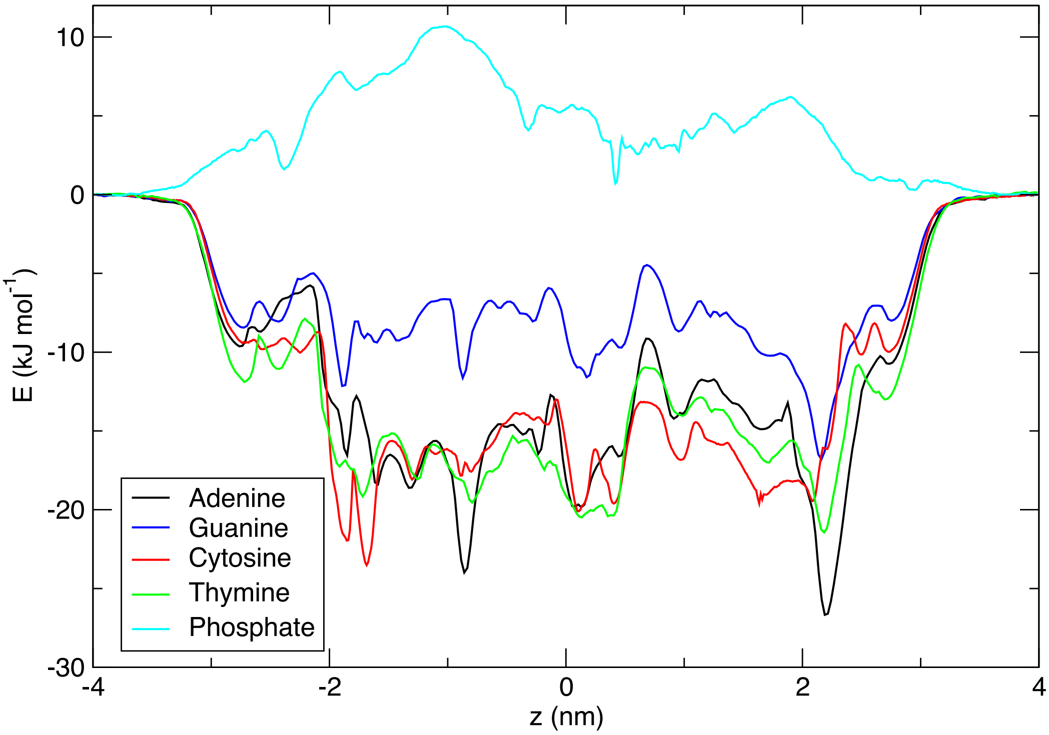


Figure 53: The PMFs of all the nucleotide fragments through αHL.

4.3.6 αHL Cytosine mononucleotide

As discussed, with the arrival of IRIDIS 4, a PMF of a cytosine mononucleotide was performed in order to compare results to those of the fragment PMFs.

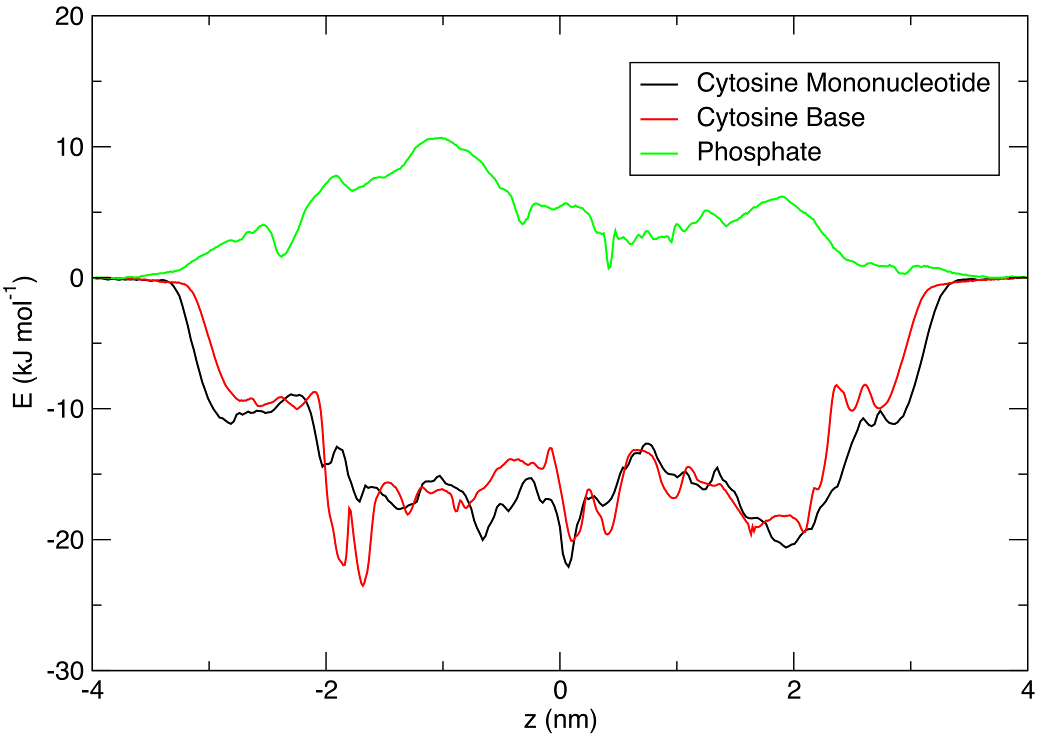


Figure 54: Cytosine mononucleotide translocating through αHL in comparison to the nucleoside and phosphate.

Of course, as this is a new system the convergence must be re-analysed. The histograms, autocorrelation times and bootstrap analysis all suggest the system is converged, as shown in appendix A, whilst visualisation of the simulations demonstrate good sampling of the heptemeric arrangement, as shown in Figure 55.

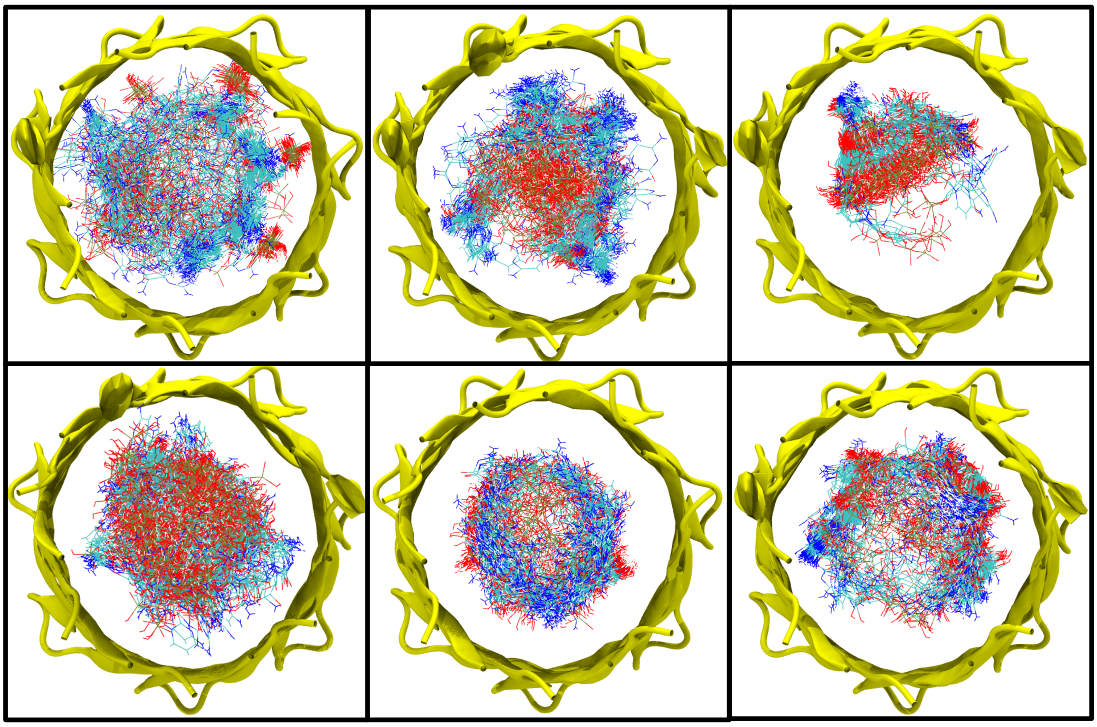


Figure 55: The position of every simulation frame superimposed, for various cytosine mononucleotide PMF windows as viewed from the cis entrance of αHL. This demonstrates good sampling of the heptamer, clockwise from top left to bottom left, z=-2.5 nm, z=-1.5 nm, z=-0.5 nm, z=0.5 nm, z=1.5 nm z=2.5 nm.

The profiles for the mononucleotide are very similar to those seen with the cytosine base, with the phosphate profile only providing minor contributions. The profile is slightly wider due to the increased size of the molecule, the harmonic is still applied to the centre of mass but the molecule is able to interact from further away, this also explains why some features of the profile have been shifted by a small amount.

The most interesting feature is the increase in energy for the mononucleotide at z=-2. This appears to simply be caused by electrostatic repulsion of the phosphate due to the aspartic acid residues. We believe that the phosphate peak at z=-1 nm is not present as the cytosine base interacts with sidechains in this region whilst the phosphate group is now in the water of the pore and free to hydrogen bond.

4.3.7 αHL Guanine mononucleotide

As the autocorrelation times, and therefore required simulation times, were much smaller than expected for the cytosine mononucleotide, it was decided that it was possible to perform an additional PMF to confirm that the trough observed for the purines at z=-0.9 nm was not possible for the mononucleotides. Therefore a PMF was setup for the guanine mononucleotide.

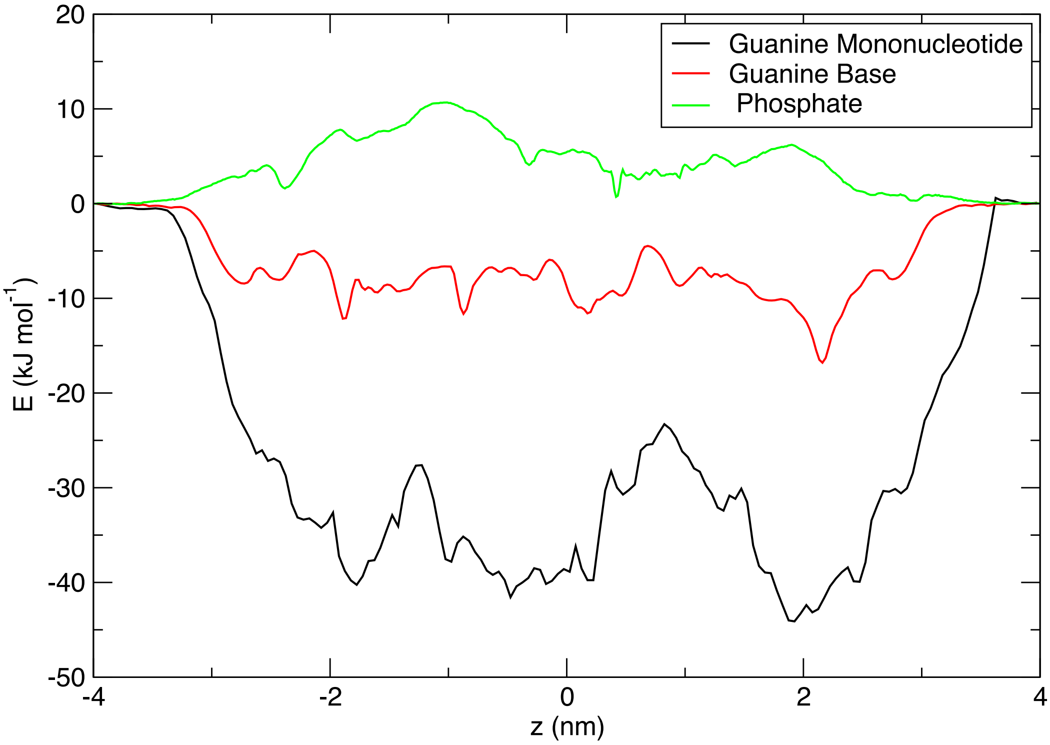


Figure 56: Guanine mononucleotide translocating through αHL in comparison to the nucleoside and phosphate.

The energetics of the guanine mononucleotide are very different to that of the base. The guanine nucleoside is at 7.5 kJ.mol^{-1} higher energy than the other nucleosides, whilst the guanine mononucleotide is 20 kJ.mol^{-1} more favourable than the cytosine mononucleotide. For the cytosine mononucleotide we observed energetics that were very comparable to the corresponding nucleoside, whereas this is not the case for guanine. Comparison between the guanine nucleoside and mononucleotide show different energetic landscapes:

the nucleotide observes a barrier at z=-1.2 nm that is not seen for the nucleoside and the minor gains in energy at z=0.4 nm and z=1.4 nm are much more pronounced.

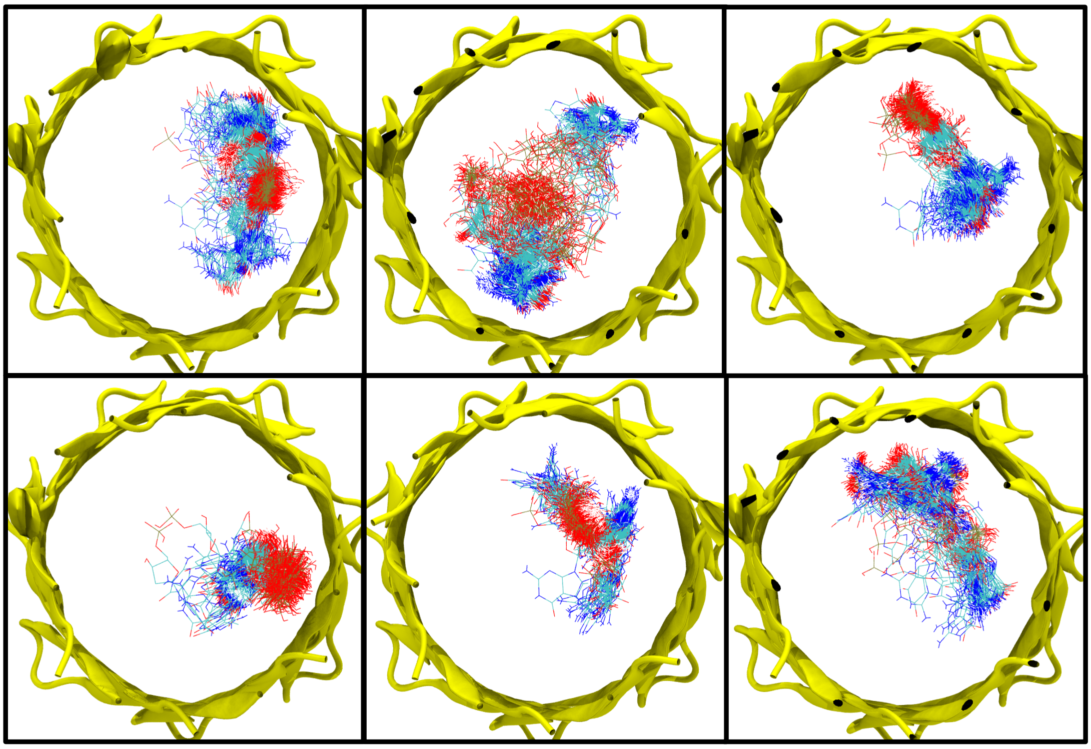


Figure 57: The position of every simulation frame superimposed, for various guanine mononucleotide PMF windows as viewed from the cis entrance of αHL. This demonstrates poor sampling of the heptamer, clockwise from top left to bottom left, z=-2.5 nm, z=-1.5 nm, z=-0.5 nm, z=0.5 nm, z=1.5 nm z=2.5 nm.

As the results presented for the guanine mononucleotide are so drastically different to those expected, the convergence was inspected. The sampling observed for these simulations is poor; the guanine frequently does not sample the heptemeric arrangement of the protein, instead finding one or two binding modes and remaining in those for the duration of the simulation as shown in Figure 57. Therefore the PMF has not sampled all space and cannot be considered converged, in order to produce converged results these simulations must either be extended significantly or more advanced techniques used. Also block analysis, shown in Figure 58, presents significantly different features over time, proving the results are far from convergence.

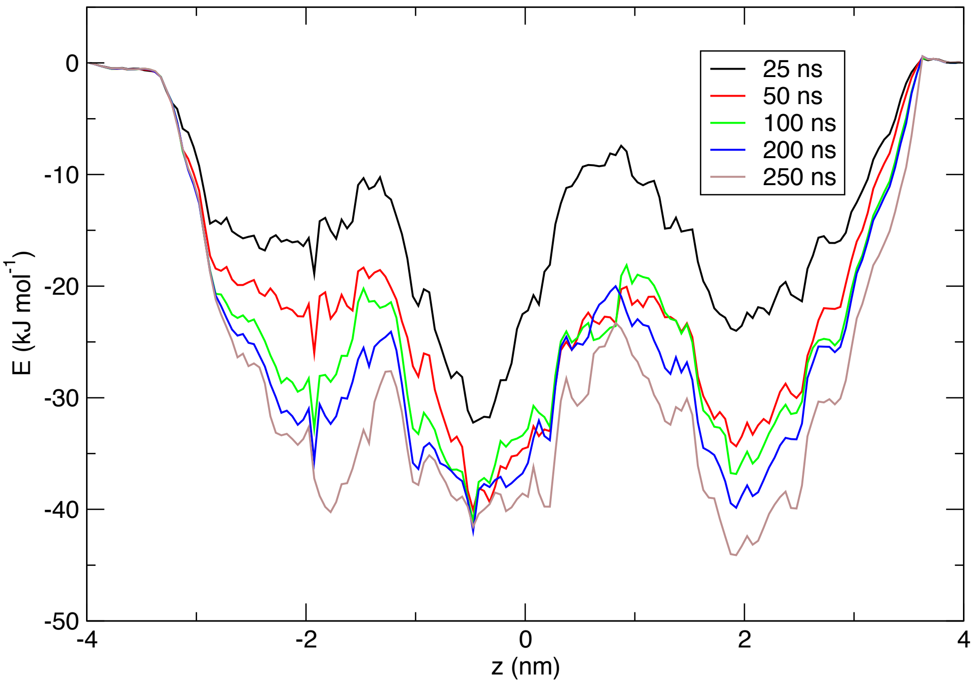


Figure 58: Analysis of the first 25, 50, 100, 200 and 250 ns blocks demonstrates poor convergence of the guanine mononucleotide PMF as each new block presents different energies and features along the PMF.

4.4 Discussion and Conclusions

For all systems, the histograms show a high level of sampling throughout the reaction co-ordinate, the autocorrelation times are small compared to the simulation lengths and have been covered multiple times whilst bootstrap analysis presents small error bars throughout the profiles. (See appendix a for histograms, autocorrelation times and bootstrap analysis.) Therefore, we consider all systems, excluding the guanine mononucleotide, converged.

The PMFs sampled fit into three groups, the phosphate, the nucleobases and the mononucleotides; as such they are discussed separately.

The energetics of translocation observed for the phosphate through αHL are very different to those reported for the phosphate specific channel OprP [171]; the phosphate fragment is at positive energy across the entirety of αHL compared to favourable energetics seen for OprP. Although the study through OprP used PO_4^{3-} compared to our use of $H_2PO_4^-$, which would create large changes near the charged amino acids. Hydrophilic interactions do not appear to dominate, as the energy does not follow the width of the pore, although hydrophobic regions of the pore do result in more unfavourable energetics.

For the nucleobases, there are similar regions in the PMFs. The pore clearly has multiple interaction sites resulting in unique energy profiles for each nucleobase. The main interaction sites discussed previously are conserved across all profiles, excluding the pocket at z=-0.9 nm due to the five membered ring of the purines.

With the mononucleotides, as the guanine mononucleotide profile is not converged it is not possible to compare the energetics. However, for the converged cytosine PMF, the profile for the nucleobase seems to dominate; the profile calculated is within thermal error of the nucleobase counterpart due to the base hydrogen bonding with the protein, forcing the phosphate to remain solvated. However in regions of negative charge, such as the pore exit, the electrostatics of the phosphate are still observed.

4.5 Future work

The guanine mononucleotide PMF should be run using an enhanced sampling technique. This would allow the mononucleotide to sample more space in the pore, without getting trapped in conformations. One such technique is replica-exchange umbrella sampling, which allows movement of the harmonic restraint and therefore encourages the substrate to sample different areas of the pore [185].

Ideally, the remaining two mononucleotides will be simulated. Further work should also consider the cap region, as for MspA it has been shown that mutation in this region can have a considerable effect on the rate of translocation, slowing translocation to one-hundredth its previous amount. This suggests that additional important binding sites can be observed outside the beta-barrel [84].

As discussed in section 1.3.4, many experiments including αHL include cyclodextrin rings, either bound covalently to the protein or free [101, 102]. It would therefore be of interest to establish the energetic contribution cyclodextrin has on translocation through αHL.

Chapter 5 Free energy of translocation through MspA

Abstract

Mutant of the MspA pore are also promising protein nanopores for sequencing, again, like αHL, previously the energetics of translocation were unknown. We undertook free energy calculations in order to further understand the translocation of mononucleotides and provide a logical starting point for further mutation of the protein. The wild-type MspA pore does not allow DNA translocation and as such, PMFs were performed on the crystal structure and a sequencing mutant. However, as the MspA protein is so large, model pores were created in order to simulate systems for reasonable timescales.

We explain energetically why the A96R mutant does not allow DNA translocation, whilst the D90N, D91N, D93N, D118R, D134R, E139K mutant does [94]. We also theorise that the additional stabilisation observed previously for guanine is due to acidic residues, as none are present in the D90N, D91N, D93N, D118R, D134R, E139K pore, guanine remains at the same energy as the other mononucleotides.

5.1 Introduction

MspA is a mycobacterial outer membrane protein, which fulfils the role of a hydrophilic channel into the cell [93]. Like aHL it is resistant to changes in pH and temperature, indeed it has been referred to as “one of the most stable channel proteins known to date” [186].

The octomeric porin MspA’s transmembrane region consists of 2 beta barrels disrupted by a proline residue [92]. There is a region of highly dense negative charge, as twenty-four aspartic acid (D90, D91, D93) residues form the narrowest point in the pore providing a large electrostatic barrier to the DNA phosphate group. The original MspA sequencing paper mutated the protein to remove this barrier, whilst also replacing further negatively charged residues with positive charges, producing the D90N, D91N, D93N, D118R, D134R, E139K mutant, referred to as the M2 mutant [84].

It is believed due to the shorter and narrower constriction compared to aHL, shown in Figure 59, that MspA has the potential to sequence DNA with more accuracy [94]. One reason for this, is that the beta barrel of aHL has up to 10 bases in the pore at a time, whilst MspA with its shorter constriction holds up to three [63].

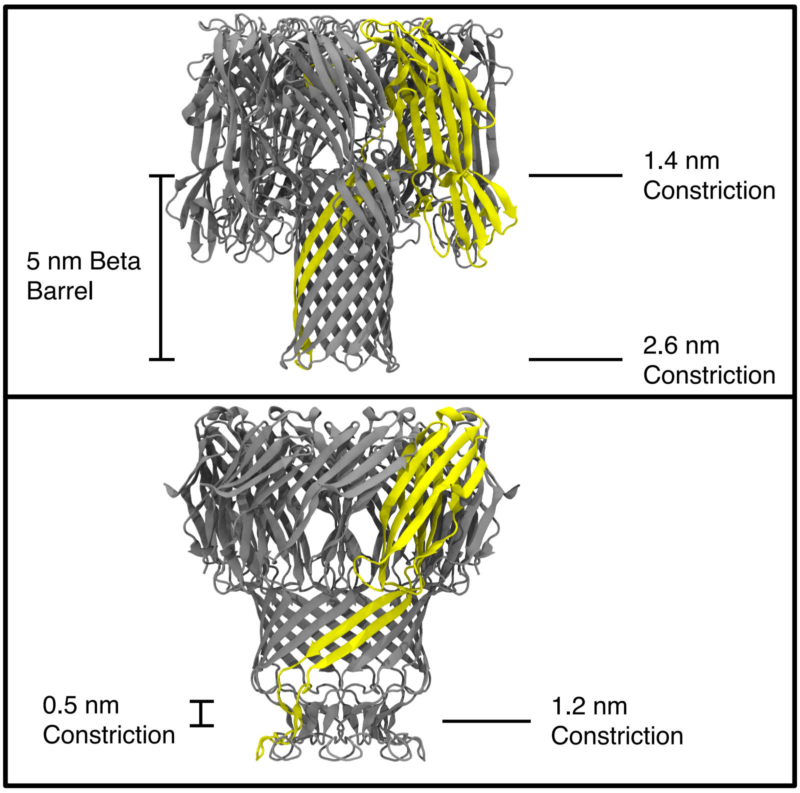


Figure 59: A comparison of the aHL (Top) and MspA (Bottom) proteins, note MspA has a shorter constriction region than aHL.

Full mononucleotides were used in these simulations, due to previous results with the cytosine base and the mononucleotide demonstrating similar autocorrelation times. The remaining mononucleotides were generated, with the 5’ hydroxyl parameters based on those in RNA, as the Gromos53a6 forcefield does not contain mononucleotide parameters [109].

5.2 Method

5.2.1 Model pore creation

The aim of the model pore was to reduce the time required to simulate the proteins; as such the main purpose was to reduce the atom count. Firstly using VMD [172], the beta barrels were removed from the remainder of the protein, reducing the atom count from twelve thousand for the full protein, compared to the four thousand of the model pore. The protein was then embedded in a membrane mimetic united-atom methane slab, created using a script provided by the Bond group. The system was then minimised, with the protein frozen in place, in order to allow the slab to distribute itself equally around the pore. The mutant pore was generated using modeller [173] and is shown in Figure 60.

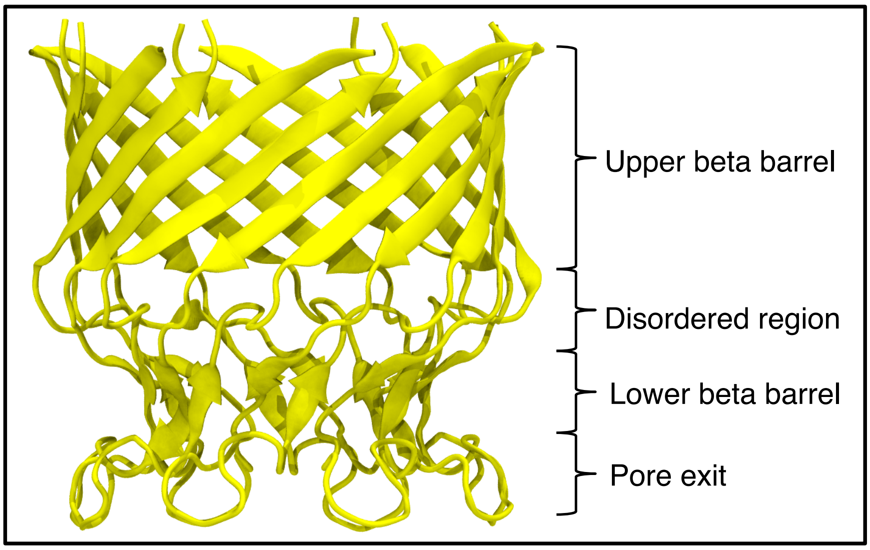


Figure 60: The MspA transmembrane domain.

5.2.2 MspA transmembrane domain

MspA consists of several regions within its transmembrane domain, namely two beta barrels separated due to proline residues, with an inflexible loop at the pore exit. The inwards facing residues are detailed in Figure 61, along with the Z co-ordinate they correspond to in the PMF. The C_α of G112 is defined as zero for the PMF calculation.

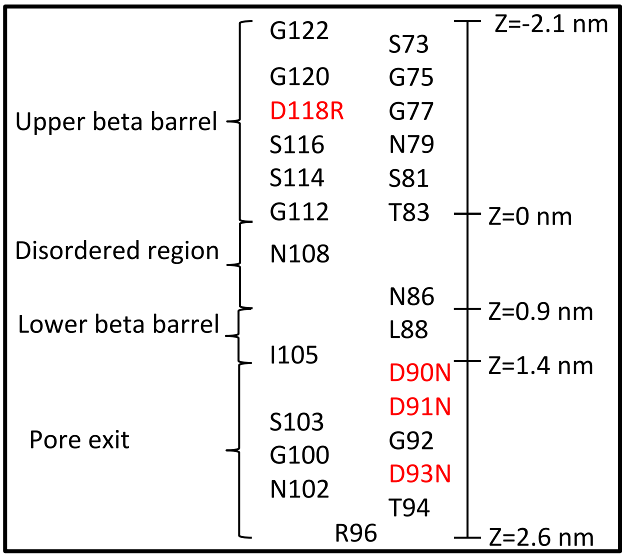


Figure 61: The locations of internal residues in MspA, with the mutated residues labelled in red.

5.2.3 Setup of the umbrella sampling calculations

The WT protein does not have a crystal structure; the crystal structure for MspA is currently the A96R mutant [92], this mutation is introduced to cause crystallisation. The mutation introduced is far from regions commonly modified in order to make a sequencing pore. In order to avoid unnecessary homology modelling, the crystal structure was used to represent the WT.

Each umbrella sampling wind for the MspA PMFs was spaced 0.25 Å apart. The harmonic potential was applied with a force constant of 2000 kJ.mol^{-1}.nm^{-2}. This is different from the previous PMFs previously described as the windows are closer together and held with a tighter restraining force, resulting in more extensive sampling, a comparison of histograms is shown in Figure 62.

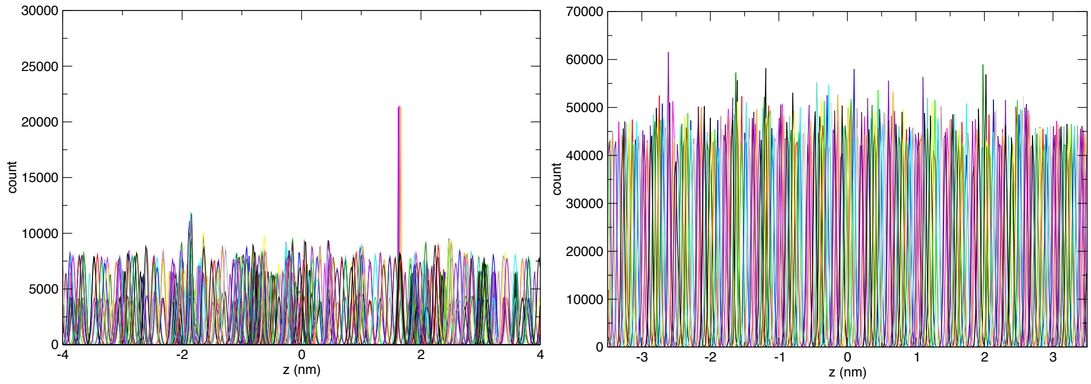


Figure 62: A comparison of αHL and MspA PMF methods, cytosine PMF through αHL on the left, CMP through MspA WT on the right, note the MspA histograms are more consistent, with fewer regions of low sampling and higher sampling in general.

In the WHAM calculation, the profile was calculated from the pull force, using bins equal to the number of windows and with the first 1 ns discarded for temperature and pressure equilibration.

5.3 Results and discussion

5.3.1 Convergence testing

As previously discussed in section 4.3.1, before inspecting the energetics it is necessary to confirm the simulations are converged.

Initially in order to work out the ideal simulation length via block analysis, a PMF was performed with 250 ns per window. This was intended to be far in excess of the minimum required. This result is shown in Figure 63, for the PMFs produced from the first 25, 50, 75, 100, 150 and 250 ns.

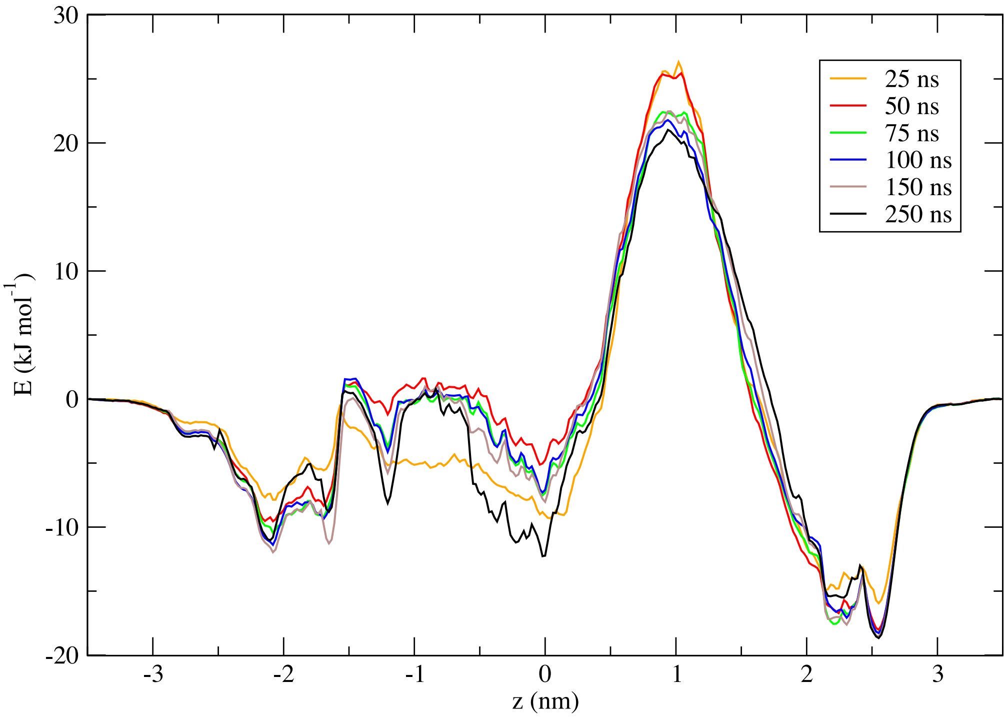


Figure 63: Block analysis of cytosine through A96R crystal structure. Note the 25 and 50 ns blocks do not capture all the peaks and troughs accurately, the 250 ns appears to get trapped in conformations e.g. z=-0.6, whilst the 75, 100 and 150 ns are all within thermal error, which is approximately 2.58 kJ.mol^{-1} at 310 K.

From inspection of the block analysis, we can see that the 25 and 50 ns blocks are not yet converged, as the energies are varying drastically from longer simulations, especially around Z=-1.25 nm and Z=1 nm. In contrast the 250 ns block appears to sample peaks and troughs not seen by the other simulations e.g. Z=-2 nm and Z=-0.6. Therefore, based on the block analysis, the remaining simulations were run for 75 ns, as this is the shortest simulation length that samples fully, whilst not getting trapped in conformations. I.e. it is within thermal error of the remaining blocks and therefore converged.

As shown in Figure 62, the histograms demonstrate extensive coverage of the principle axis of the protein. The autocorrelation times, see Figure 64, are below 20 ns, with only 3 frames where it is above 10 ns.

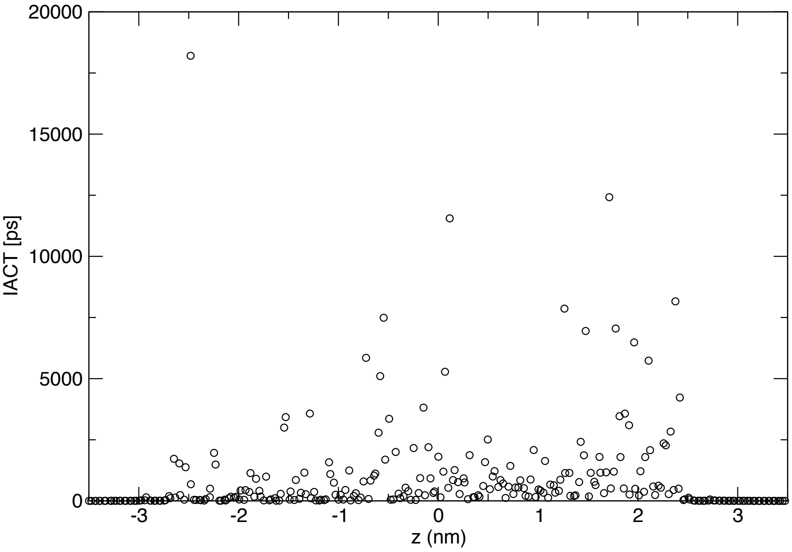


Figure 64: The autocorrelation times for the cytosine mononucleotide through the A96R mutant.

The average of the 500 bootstraps are shown in Figure 65. There are negligible errors in the bulk water and a maximum of 2.7 kJ.mol^{-1} in the protein, as thermal error is 2.58 kJ.mol^{-1} at 310 K this was deemed acceptable. For all other histograms, autocorrelation times and bootstrap analysis see appendix C.

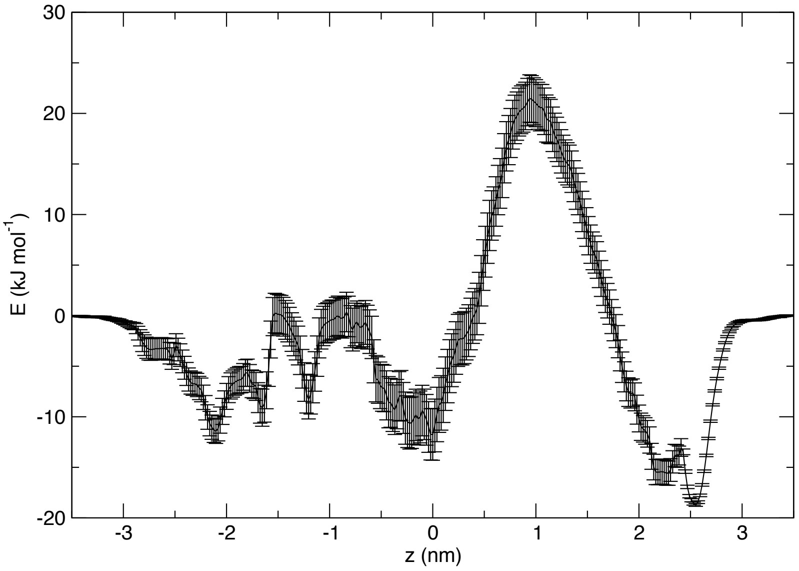


Figure 65: The bootstrap analysis for the cytosine mononucleotide through the A96R mutant.

5.3.2 A96R mutant

The crystal structure is a non-sequencing mutant, comparable to the wild type. The arginine residues at the trans exit promote the formation of ordered crystals but also, due to the mutation of alanine to arginine there will be a charge influence that wouldn't be observed with the WT. Despite this, the mutant should be comparable to the WT in all other regions.

5.3.2.1 Cytosine

The PMF of cytosine translocation through the A96R mutant is shown below in Figure 66, the immediately apparent features are the regions of high energy, i.e. those above 0 kJ.mol^{-1}, located at Z=-1.5 nm, Z=-1 to Z=-0.5 nm and Z=0.5 to Z=1.5 nm. It is unsurprising that such features exist, as the WT MspA protein is non-sequencing [94].

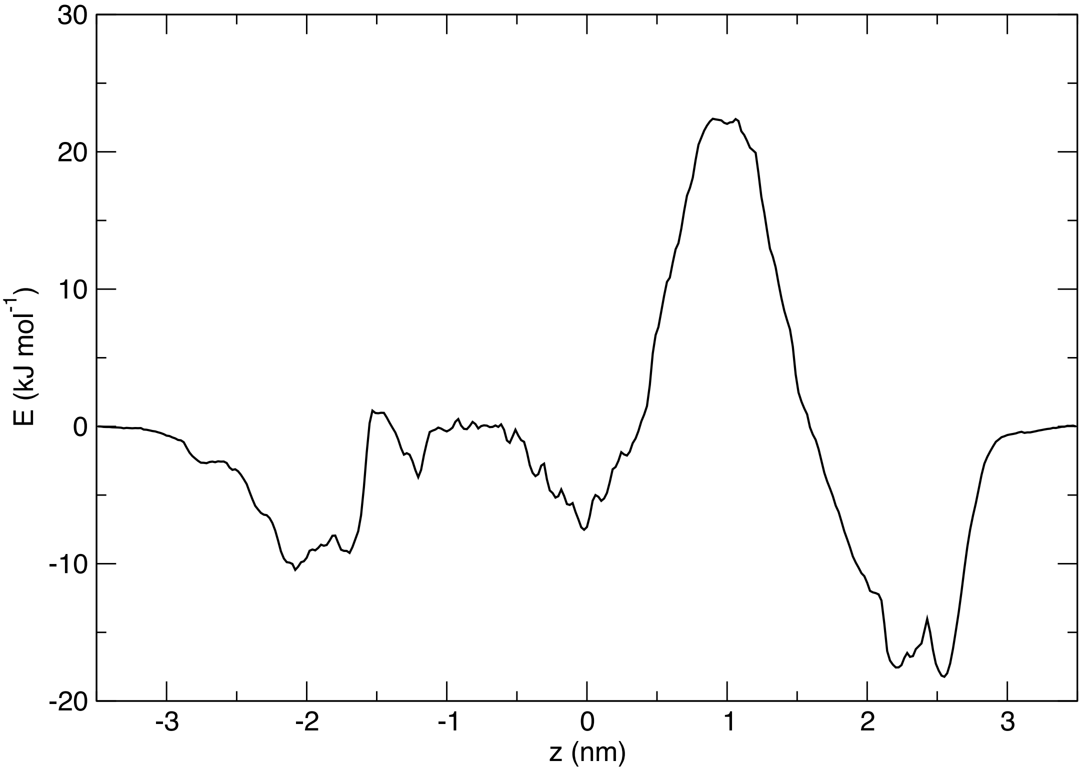


Figure 66: Cytosine PMF through A96R MspA mutant. Due to the barriers to permeation this protein is non-sequencing.

The cis entrance of the pore is at Z=-2.2 nm, whilst the trans exit is at Z=2.6 nm. As well as the unfavourable regions previously mentioned, there are regions of favourable energy, notably Z=-2.2 nm, Z=-1.6 nm, Z=-1.2 nm, Z=0 n, Z=2.2 and Z=2.6 nm. The various regions of interest are now discussed in turn.

At Z=-2.2 nm, the mononucleotide interacts favourably with the cis entrance to the pore, preferentially binding to the same protein monomer. Hydrogen bonding is observed over the duration of the simulation between the phosphate group and either S73 or N121. Whilst the phosphate is interacting with N121, the hydroxyl on the sugar hydrogen bonds to S73. The nucleoside remains orthogonal to the principal axis of the protein, extending over the membrane mimetic slab for the duration of the simulation as shown in Figure 67.

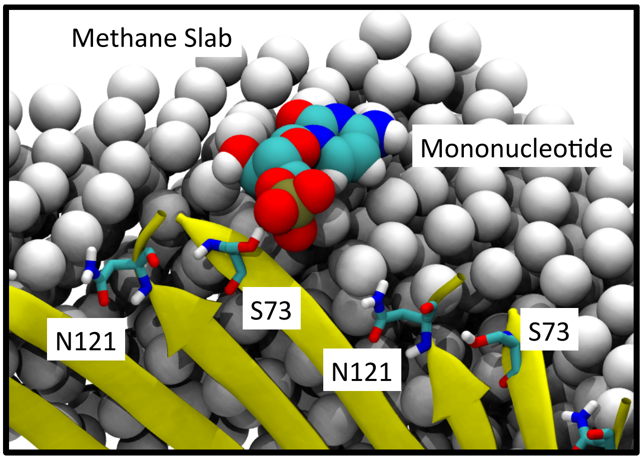


Figure 67: The dominant interaction between the cytosine mononucleotide at Z=-2.2 nm, N121 and S73. The mononucleotide is predominantly stabilised by hydrophobic exclusion by positioning the nucleoside face on to the methane slab, shown as white spheres.

At Z=-1.6 nm the mononucleotide is free to sample the entire octomer, both with regards to translation and rotation. Predominately, the mononucleotide remains close to the protein, interacting in various manners with S73, N121, V76 and D118. It is worth noting that the phosphate preferentially distances itself from the aspartic acid residue due to electrostatic repulsion. The mononucleotide is largely orientated with the phosphate either towards the cis entrance, or orthogonal to the principal axis, rarely orientating towards the trans exit, closer to D118.

At Z=-1.5 nm, the mononucleotide samples freely. At this position the substrate regularly alternates binding mode with the protein, with each binding mode lasting no longer than 20 ns. D118 and SER 73 provide some hydrogen bonding, but the electrostatic repulsion from D118 is likely the reason for no stabilisation in this region.

At Z=-1.2 nm the mononucleotide strongly binds to N121, S73, D118, S116 and N79 as characterised in Figure 68. The nucleoside alternates hydrogen bonding between the backbone of N121 and the sidechain of S73. The sugar's hydroxyl group hydrogen bonds to D118 and the phosphate hydrogen bonds with either S116 and N79, or S116 and D118. This results in a moderate, local stabilisation.

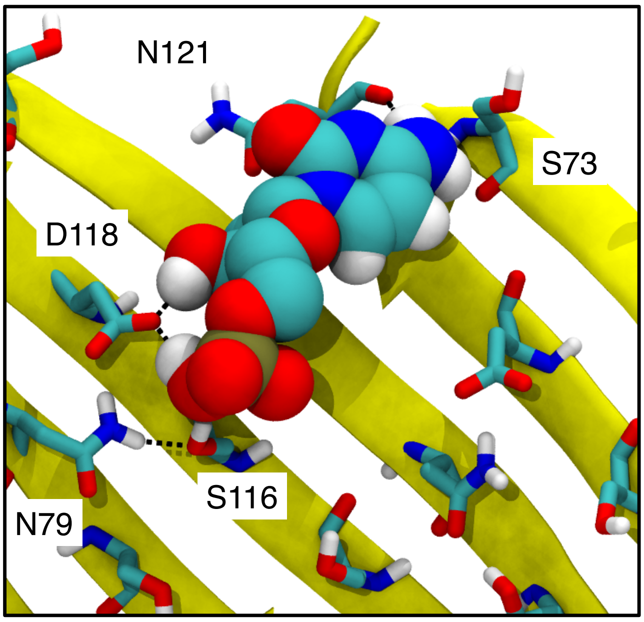


Figure 68: The hydroxyl group on the sugar hydrogen bonds to both S116 and D118, whilst the phosphate hydrogen bonds to S116 and N79, the base hydrogen bonds to the backbone of N121.

At Z=0 nm two T83 residues hydrogen bond to the nucleoside and the sugar hydroxyl group. Infrequently short lived hydrogen bonds between the sugars hydroxyl group and N108 are observed, twice over the simulation for approximately 5 ns each. The mononucleotide remains flat in this region with the phosphate pointing inwards, remaining solvated, as shown in Figure 69.

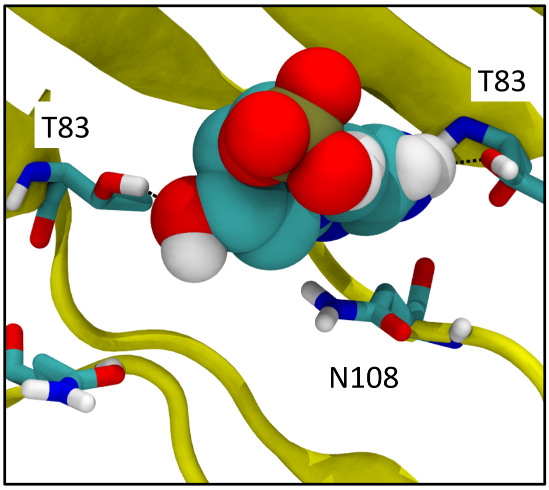


Figure 69: At the bottom of the upper beta-barrel, Z=0 nm, the mononucleotide is stabilised by hydrogen bonds to T83.

At Z=0.5 nm to Z=1.5 nm we encounter a large energetic barrier, due to the D90, D91 and D93 residues electrostatically repelling the phosphate group. We observe the base hydrogen bonding to each aspartic acid residues, whilst the phosphate remains central in the pore. The base lies orthogonal to the principle axis for the duration of the simulations. At Z=2.2 nm several short lived, approximately 10 ns, hydrogen bonding modes are observed between the base and D91, D93, A96, S103, N102 or the phosphate and A96, S103 or N102.

Whilst the nucleotide is at Z=2.6 nm we note hydrogen bonds with D93, S103 and N102 present for greater than 50 ns, The phosphate group on the mononucleotide hydrogen bonds to either the side chain of S103, the side chain of N102 or the backbone oxygen of D93, as shown in Figure 70. The hydroxyl on the sugar hydrogen bonds to N102 or D93 and the base sits in a pocket formed by R96, which surprisingly results in the exclusion of water.

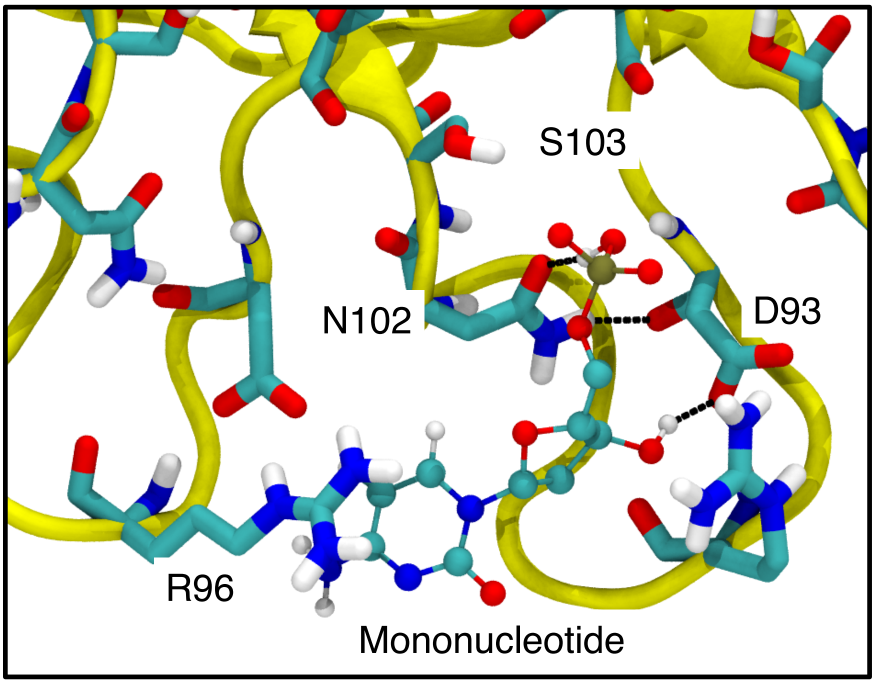


Figure 70: As the nucleotide exits the pore, a favourable binding site is observed with N102, R96 and D93 residues.

5.3.2.2 Comparison of A96R PMFs

The PMFs of the mononucleotides through the MspA A96R mutant are shown in Figure 71. As observed for αHL, aside from guanine, the nucleotides are mostly within thermal error. As with the guanine mononucleotide profile through αHL, guanine finds additional stabilisation in comparison to the cytosine profile by approximately 10 kJ.mol^{-1}. We postulate due to the additional hydrogen bond donor group, it is more stabilised than the other mononucleotides. In the wide upper beta barrel of MspA, at Z=-1.6 nm, this stabilisation is due to hydrogen bonding to D118 similarly, in the constriction, Z=1 nm, additional hydrogen bonds to the aspartic acid residues D90 D91 D93 causing reduced destabilisation.

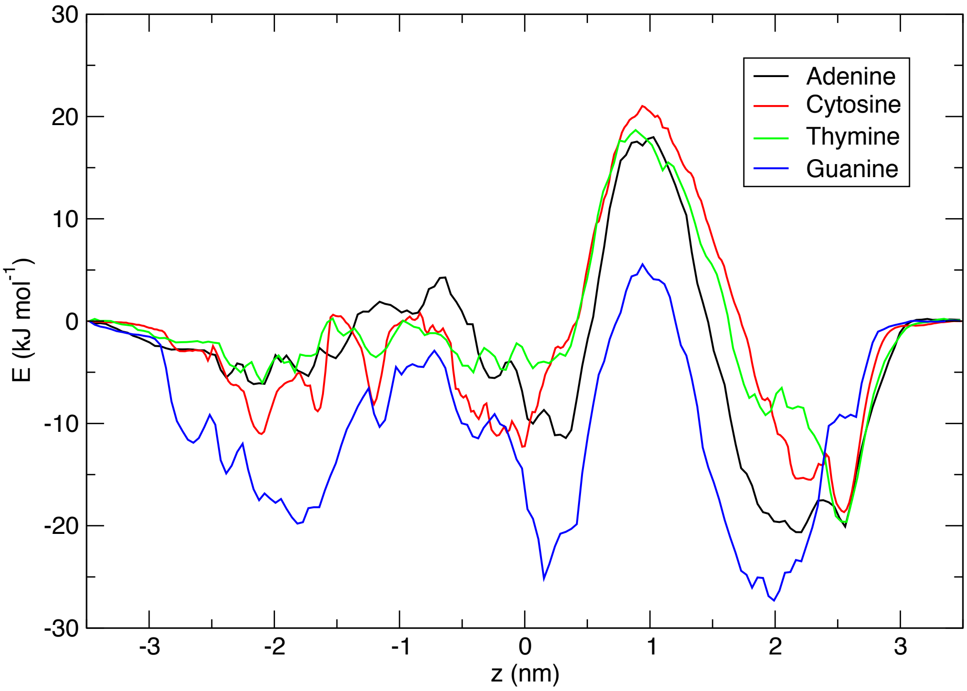


Figure 71: The mononucleotide PMFs through the MspA A96R mutant.

The adenine mononucleotide PMF curve is very similar to that of cytosine. The profiles remain within thermal error, 2.6 kJ.mol^{-1} apart from in the following regions; Z=0.3 nm and Z=2 nm. This appears to be due to the increased size of adenine; the phosphate group is able to remain further away from D93, D91 and D90, reducing the electrostatic repulsion.

As expected of the other pyrimidine, thymine produces a very similar PMF to cytosine. The main differences is at Z=-0.2 nm. This is due to the nucleoside and the sugar forming hydrogen bonds whilst the phosphate remains close to the edge of the pore, for thymine. In comparison, for cytosine, the phosphate is more solvated and therefore has more hydrogen bonding with waters, which is the cause of this stabilisation.

5.3.3 D90N, D91N, D93N, D118R, D134R E139K mutant

5.3.3.1 Cytosine

Cytosine PMF through the D90N, D91N, D93N, A96R, D118R, D134R, E139K mutant (M2) is compared to the A96R profile shown below in Figure 72. The M2 profile has no regions of positive energy, which explains why this mutant is sequencing; the mononucleotide has favourable interactions along the entire pore.

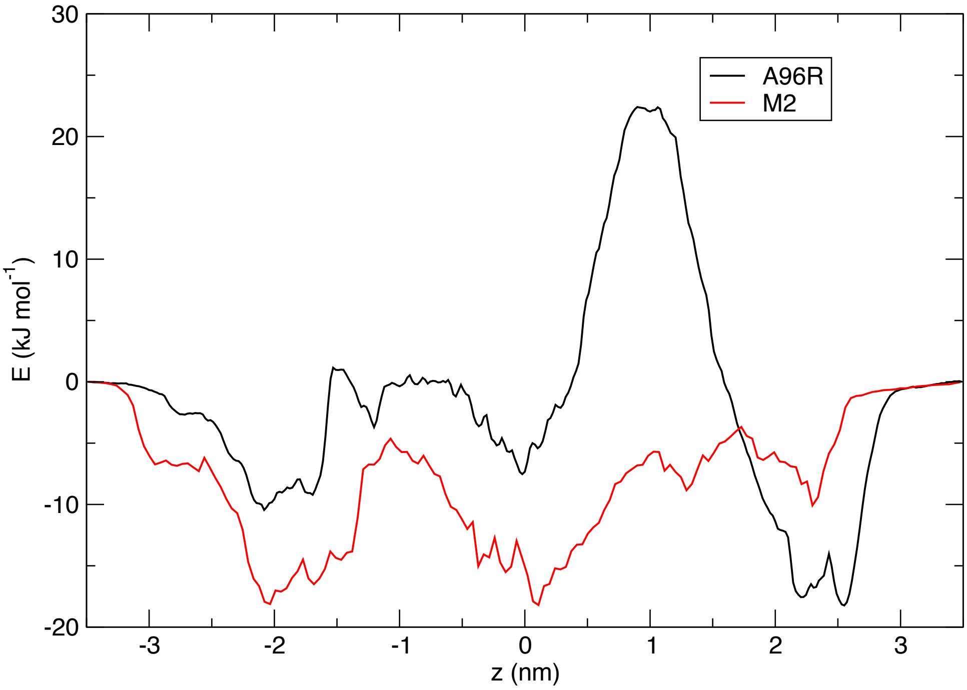


Figure 72: PMF of a cytosine mononucleotide translocating the A96R MspA mutant, in black, in comparison to the M2 mutant, in red.

The regions that are different between the cytosine PMFs through the M2 and A96R MspA mutants are Z=-1.5 nm, Z=-1.1 nm, Z=-0.5 nm, the region Z=0 nm to Z=2 nm and Z=2.5 nm.

At Z=-2.0 nm, the cis entrance of the protein, the binding mode is very similar to that observed with the wild type as shown in Figure 73. The nucleotide is stabilised by hydrogen bonding from either the phosphate group or sugar hydroxyl to the sidechain of S73 or the backbone of N121. Therefore we suggest that the stabilisation in this region is due to the electrostatic contribution of the D118R mutation.

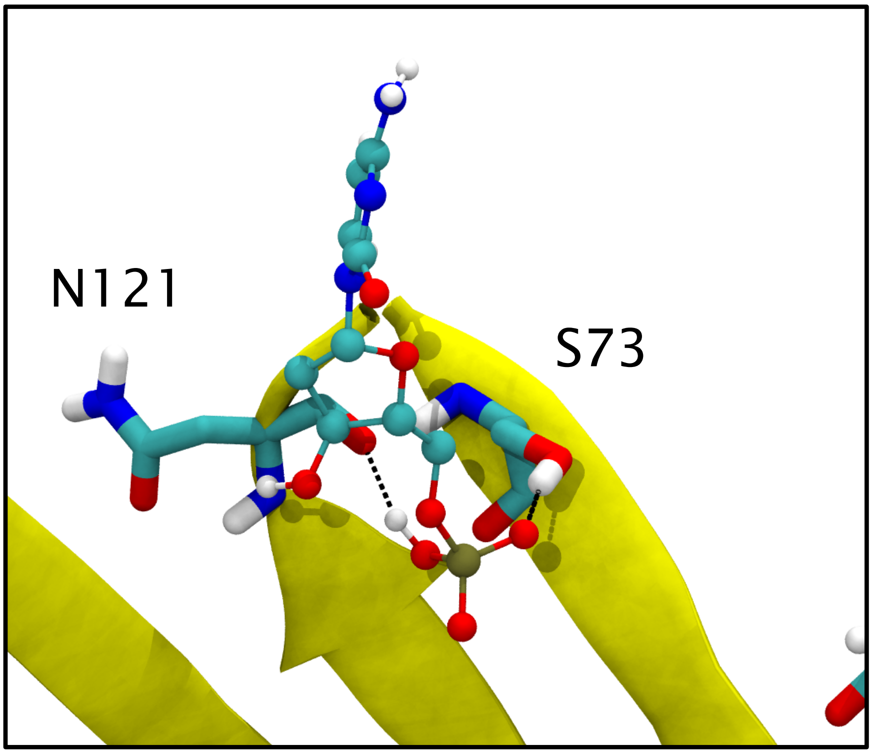


Figure 73: A similar binding mode between the A96R and M2 pores suggest the electrostatic contribution from the local D118R mutation.

At Z=-1.5 nm the D118R mutation results in a stabilisation due to favourable electrostatics, this effects the entire upper beta barrel region. When the mononucleotide samples Z=-1.2 nm due to the mutation D118R the favourable binding location for the nucleoside is no longer present, hence there is no reduction in energy in this region.

Along Z=0 nm to Z=2 nm, the D90N D91N D93N mutations and associated loss of electrostatic repulsion results in massively more favourable energetics by approximately 25 kJ.mol^{-1}. There is still an increase in the energy for the M2

PMF moving from Z=0 nm to Z=2 nm due to the decrease in solvation of the phosphate group. However at no point does the energy become positive and hence unfavourable.

Near Z=2.6 nm, the trans exit of the protein, the D91N and D93N mutations along with the loss of the A96R mutation results in less favourable energetics than the A96R protein. The hydrogen bonding of the phosphate and the hydroxyl group to N121 and N93 whilst the base hydrogen bonds to N91, as shown in Figure 74, do not make up for lack of electrostatic stabilisation from the missing arginine.

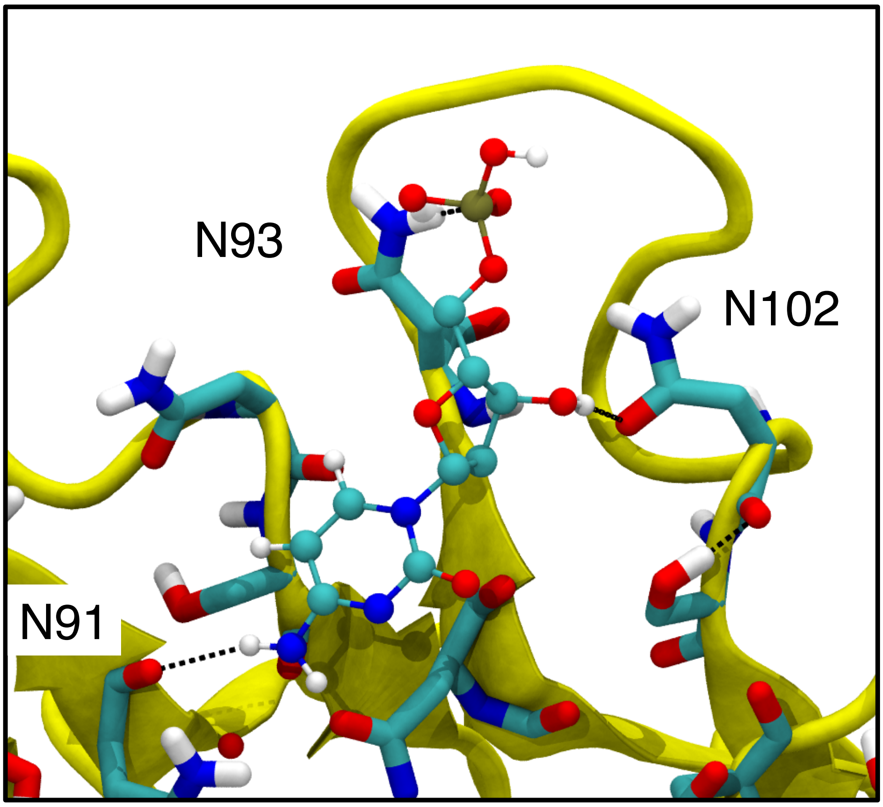


Figure 74: Stability is increased in the region Z=-2.0 nm partially due to hydrogen bonding to N93, mutated from D93.

5.3.3.2 Comparison of M2 PMFs

The mononucleotide PMFs through the M2 MspA mutant are shown below in Figure 75. We observe favourable energetics for all mononucleotides throughout the pore, suggesting this pore can sequence any DNA strands, as confirmed by experimental data [85]. The PMFs through the M2 pore are noisier than those through the A96R mutant as the binding locations are stronger and therefore the sampling is more limited in comparison.

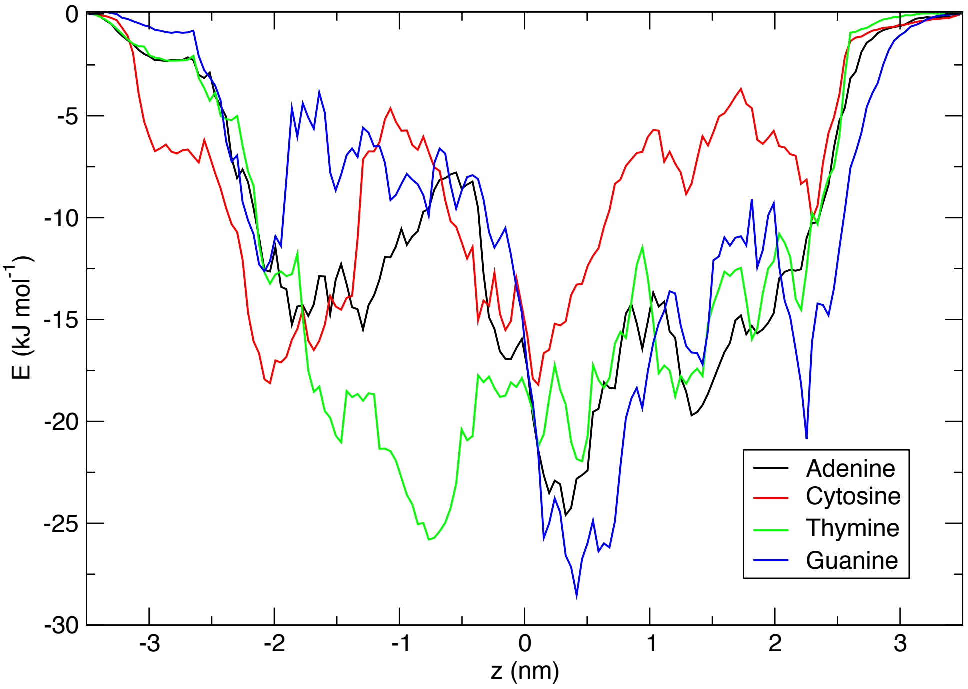


Figure 75: A comparison of mononucleotide PMFs through the MspA M2 mutant.

Adenine appears to display similar energetics to the other mononucleotides throughout the pore. In contrast thymine is strongly stabilised at Z=-0.7 nm, due to a pocket formed by N79, S114, S116 and R118 which the other nucleotides fail to sample, shown in Figure 76.

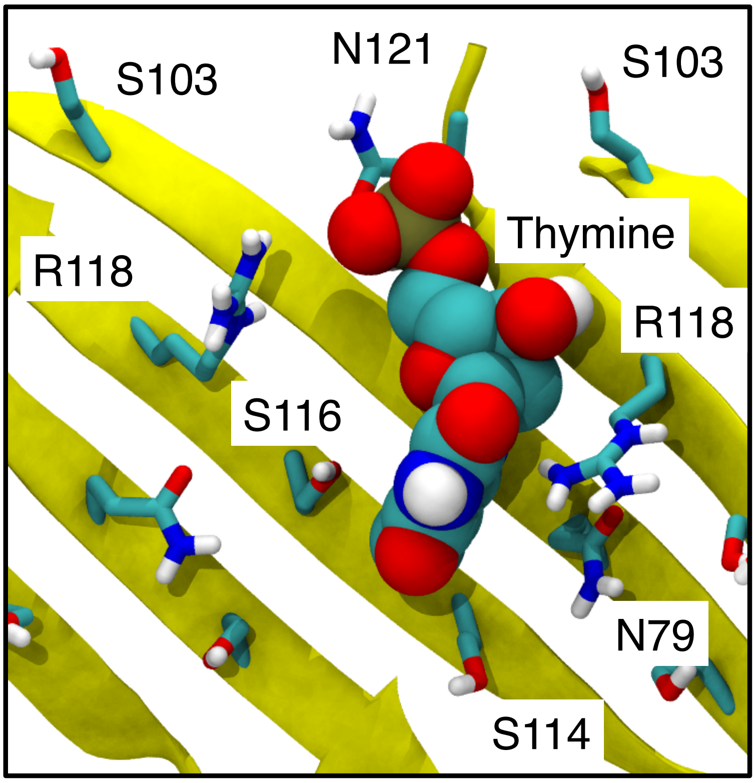


Figure 76: A pocket only sampled by the thymine mononucleotide, at Z=-0.8 nm.

As shown in Figure 75, cytosine finds less favourable energetics at Z=-1.1 nm and the region Z=0 nm to Z=2 nm by 7.5 kj.mol^{-1} in comparison to the other nucleotides. From Z=-1.3 nm to Z=-1.1 nm the phosphate moiety is stabilised by hydrogen bonds and electrostatics from R118. At Z=-1.1 nm there are no other hydrogen bonds to the mononucleotide, whilst at Z=-1.3 nm there are hydrogen bonds from the nucleotide to N121 and S103, as shown in Figure 77. In the region Z=0 nm to Z=2 nm cytosine has fewer hydrogen bonds with N90 N91 and N93, when compared to the other mononucleotides as shown in Figure 78.

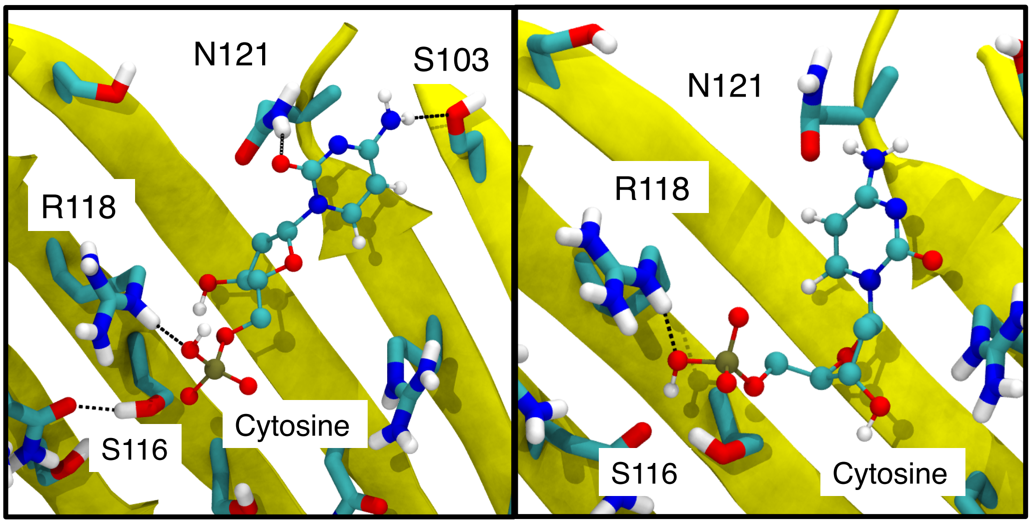


Figure 77: Binding modes for cytosine at (Left) Z=-1.3 nm and (Right) Z=-1.1 nm. There are additional hydrogen bonds at Z=-1.3 nm which contribute to the stabilisation in this region

In the region Z=0 nm to Z=2 nm cytosine has a single hydrogen bond with N90 N91 and N93, compared to 3 observed for the other mononucleotides as shown in Figure 78.

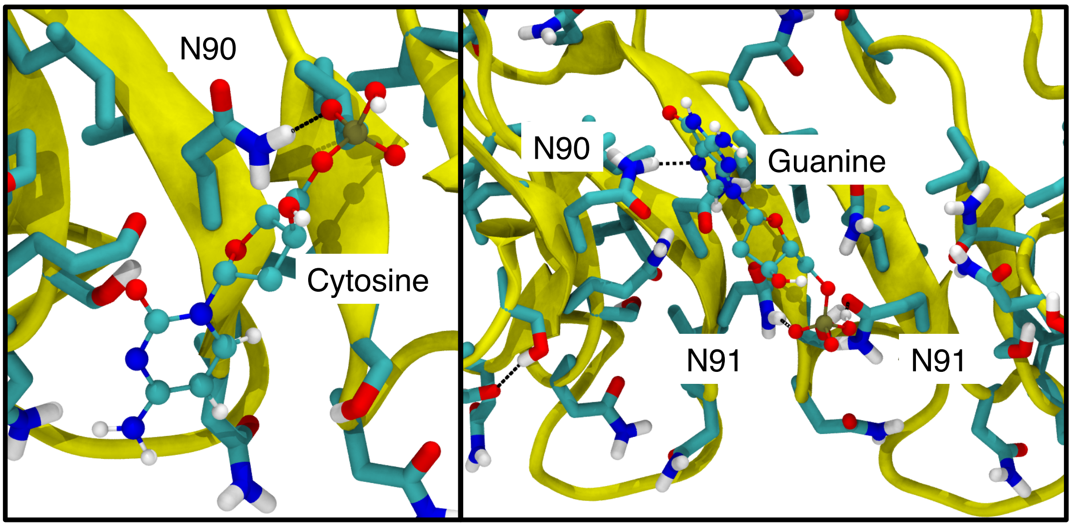


Figure 78: The hydrogen bonding pattern at Z=0.9 nm for (Left) cytosine and (Right) guanine show cytosine has fewer hydrogen bonds in this region.

For the first time, the guanine PMF displays similar energetics to the other nucleotides. We postulate that this is due to the lack of acidic residues in this pore, as they are strong hydrogen bond accepting residues, which would result in a strong stabilisation of the mononucleotide observed in αHL and the MspA A96R mutant.

5.4 Conclusions

We have provided rational explanations as to why the A96R is non-sequencing, whilst the D90N, D91N, D93N, A96R, D118R, D134R E139K mutant is. This is due to the removal of the barrier caused the residues D90, D91 and D93.

We also explain why guanine experiences different energetics for αHL and the MspA A96R mutant; acidic residues provide additional stabilisation due to hydrogen bond acceptance.

5.5 Future work

As the entire pore was not simulated, there are still some energetics that remain unknown. The differences between two mutant pores, M1 and M2 described by Butler et al [84], the inspiration for this work, are in the cap region outside the model pore. These two proteins bind DNA with very different strengths; M2 binds DNA approximately twenty times stronger then M1, therefore to explain why the binding strength is so different between the two mutants further simulations would need to include the cap region.

Alternatively, future simulations could predict the energies of alternative mutants with higher binding energies. A mutant we would suggest would include addition of positive charges near the narrowest point of the pore, e.g. D93K, although as the pore is so narrow the addition of longer sidechains may influence translocation adversely.

Chapter 6 Contribution of water to the energetics of DNA translocation through protein nanopores

Abstract

Water is the ubiquitous solvent in biology and because of this the contribution it presents to energetic barriers is often overlooked. Despite this, water frequently can be the dominant energetic barrier; studies of protein-ligand binding have investigated the displacement of waters extensively.

As a continuation of the pore dimensions analysis performed on αHL using HOLE, a more detailed method of investigating solvation was developed. This chapter presents a novel technique of looking at the displacement of water and the energetics involved. Using this new technique, the water displaced during translocation through the αHL and MspA model pores was analysed. We show that the displacement of waters does not provide a thermodynamically significant energetic barrier through either protein.

Chapter 6

Notes

This chapter was produced in collaboration with Dr Ross from the Essex group at the University of Southampton. I performed all simulations, analysis of data and wrote the preliminary program for obtaining free energies. Dr Ross derived the free energy equation Equation 25. Collectively we developed the method for obtaining free energies.

6.1 Introduction

Water is an important solvent for biological systems [129]. Studies of X-ray structures of protein-DNA complexes have shown that six percent of the crystallographic waters are involved in the protein-DNA binding process [128]. It has been shown that the ejection of water from the protein-DNA interface is energetically favourable, however those that remain can be important for shielding from electrostatic repulsion [187]. This emphasises that water plays a vital role mediating the protein-DNA boundary as it can screen unfavourable electrostatics, reduce the stability of the complex by disruption of protein-DNA hydrogen bonding or mediate hydrogen bonds [188].

When considering protein-ligand binding, water is equally important and complex energetically. Waters in ligand binding sites can be energetically favourable, requiring energy to displace, whilst in contrast other protein pockets remain dry [189]. Therefore when designing drugs, studies frequently consider if water molecules in the binding pocket should be conserved or displaced, as the energetics of ligand binding may not be favourable enough to remove the water [190, 191].

DNA translocating through protein nanopores is an intricate process, which is predicted to occur through a ‘binding and sliding’ mechanism for αHL [98]. In our previous studies of the PMFs through protein nanopores we observed several binding modes for a given point in the pore, often with water mediated hydrogen bonds. It was noted in section 4.3.2 that where the pore narrowed, generally favourable energetic regions were observed, suggesting that hydrophobic effects dominate for mononucleotides.

Therefore wishing to investigate how water contributes to the energetics of DNA translocating through nanopores, current methods to obtain the free energy of water binding were assessed.

Cluster analysis is a method that can provide the locations of preferred water binding [192], this allows for analysis of important regions, such as enzyme active sites [193]. Cluster analysis however does not provide any energetics; a free energy technique is required to calculate the free energy of binding.

Alchemical methods, such as exponential averaging [194], thermodynamic integration [195] and the Bennett acceptance ratio [196], BAR, allows for the study of changes in free energy between states by sampling across a reaction co-ordinate, lambda. Lambda windows frequently correspond to how “switched on” or “off” a molecule is and determines the strength of the interactions the molecule experiences. It has been used to study the free energy of solvation of simple molecules [197] as well as protein ligand binding [198]. Unfortunately due to the requirement to simulate many lambda windows for each water molecule of interest, of which there will be many in each region of the nanopore, therefore these techniques are too computationally expensive to address how water contributes to DNA translocation.

Radial distribution functions, RDFs, measure how the density around a solute in an isotropic homogeneous fluid varies as the radius is increased [112]. Therefore RDFs measure density changes along a single axis. Clearly a protein is not an isotropic system, neither is it homogeneous, therefore RDFs cannot be used to investigate our systems. However the proteins we investigated, αHL and MspA, are heptemeric and octomeric pores, therefore we have sevenfold and eightfold rotational symmetry in these systems along the principle axis of the protein, which reduces the degrees of freedom. We designed a method similar to RDFs, by measuring densities along the principle axis of the protein in contrast to a radius.

6.2 Methods

The simulations analysed were the cytosine mononucleotide systems discussed in section 4.3.6 and 5.3.2.1.

6.2.1 Terminology and examples

In this chapter several terms are used to describe the process of nucleotides translocating and how it affects the water density across the porin, these terms are defined in Table 1.

Term	Definition	Example
Z	A co-ordinate in the simulation box,	The nucleotide is at Z = 6 nm
p(x)	The probability density of (x)	p(Z)
K_x	The harmonic restraint position in the box, each corresponding to a PMF simulation window	Under K_5 the nucleotide is restrained around Z=5 nm
|	Given	p(Z|K_5), the probability density of finding the nucleotide at Z given the harmonic restraint at Z=5 nm
ζ	Greek letter "zeta", refers to the water Z-coordinate across the simulation box	ζ varies as the nucleotide is moved through the pore

Table 1: The definitions of terms used through this chapter

The simulation setup is the updated PMF method described in 5.2.3.

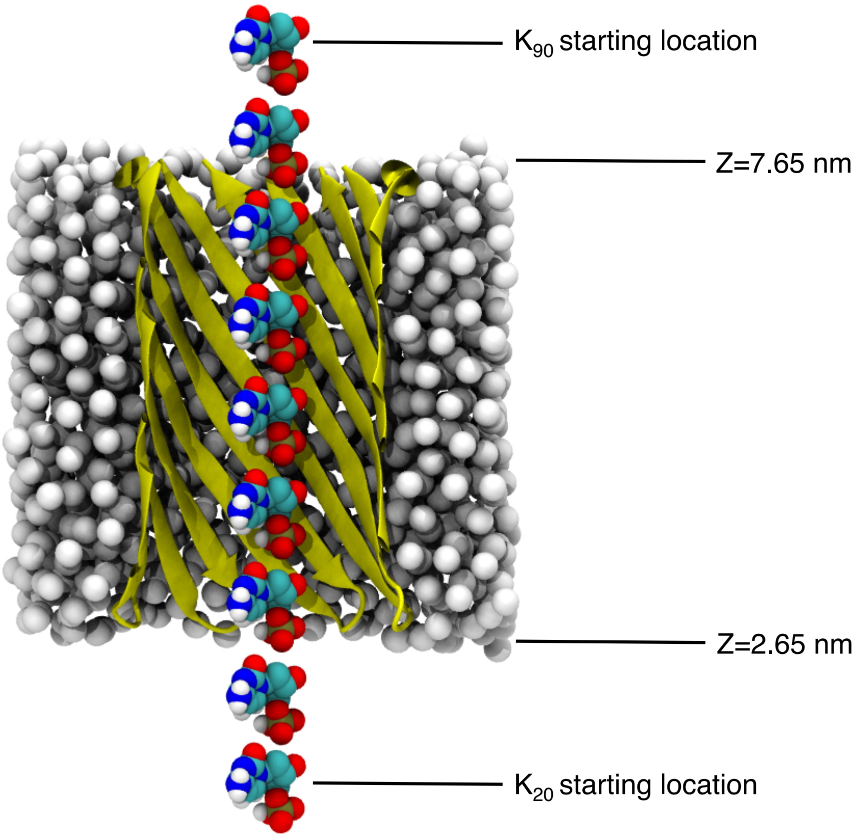


Figure 79: An overview of the system setup for αHL, each mononucleotide is in a different simulation, with water and ions omitted for clarity.

The location of residues in the simulation box for αHL are shown in Figure 80. Note E111 and K147 form the narrowest region in the pore.

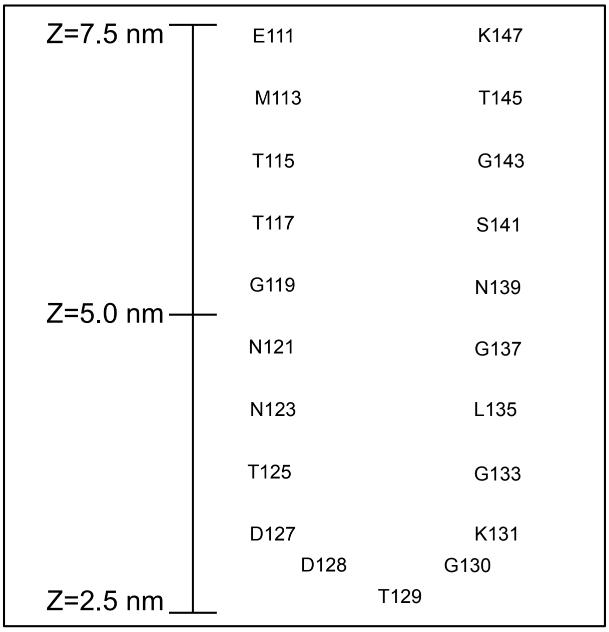


Figure 80: The locations of inwards facing residues in the αHL model pore.

Density plots were calculated using the native GROMACS tool, g_density. For water density plots the box was split into 100 “slices” along the Z axis, whereas for the nucleotide density the box split was into 500 slices. These plots reveal how the density varies across the entire simulation box, an example water density plot is shown below in Figure 81. Note that this corresponds to a single simulation, which only contains 500 output frames, therefore the noise observed in bulk water, although undesirable, is not wholly unexpected. The pore entrance is at Z=7.65 nm and the exit is at Z=2.65 nm.

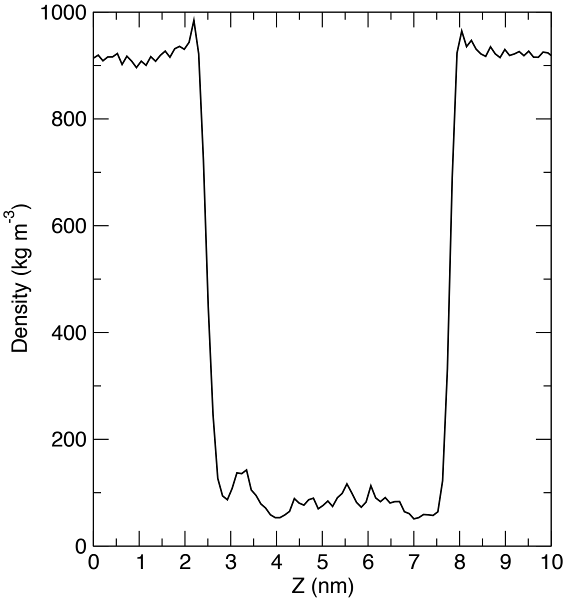


Figure 81: p(ζ | $K_{1.13}$), the reduction in water density is due to the water entering the protein from bulk water.

Figure 81 does not show the expected density of bulk water, 1000 kg.m^{-3}. This is because the simulations were run in 1M NaCl, therefore the density of both water and ions was calculated to demonstrate the accuracy of the simulations, shown in Figure 82. Here we observe bulk densities of approximately 1060 kg.m^{-3}, which is very similar to the calculated value of 1058.5 kg.m^{-3}, see appendix C part 1 for the calculation of the expected density of a 1 M NaCl solution.

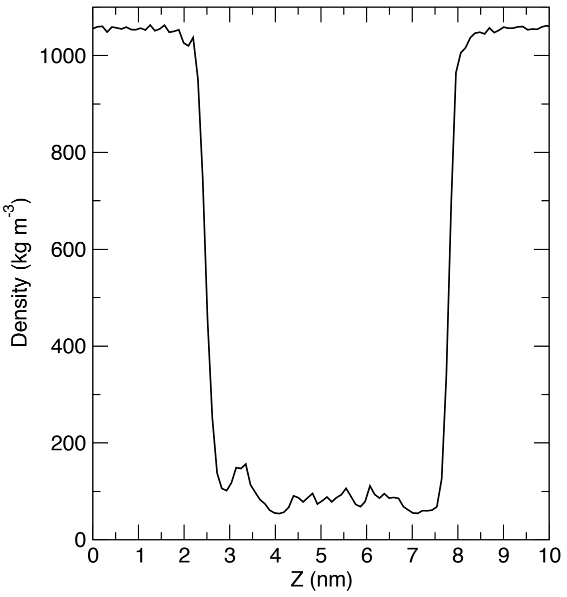


Figure 82: The density profile of the 1 M NaCl solution.

An example nucleotide density plot, given K_1, is shown in Figure 83. The mononucleotide is approximately 1.1 nm across at its maximum width, see the appendix C, part 2, and is free to rotate. As the harmonic restraint is applied to the centre of mass, even with slight perturbations from the centre of the harmonic well, the observed histogram width of 1.4 nm is only slightly larger than the width of the molecule. We therefore note that for a given value of K the nucleotide occupies a large number of Z-coordinates.

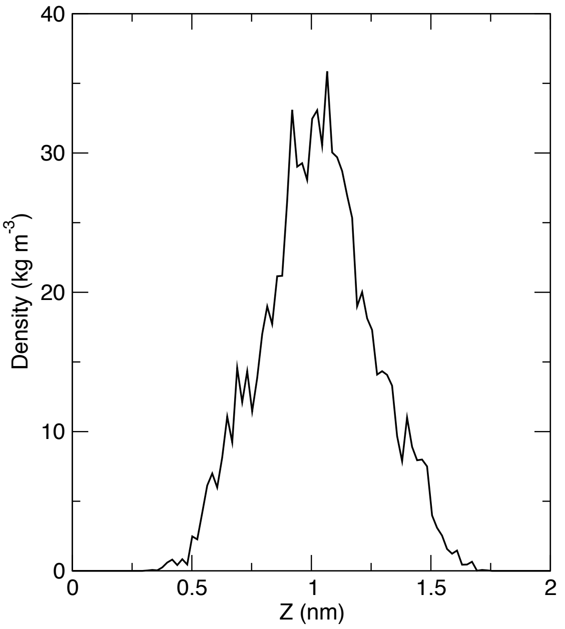


Figure 83: The p(Z| $K_{1.13}$) plot, zoomed in for clarity.

A comparison of p(ζ|K_∞), i.e. an empty pore, against p(ζ|K_5),

Figure 84 shows that water was excluded where the nucleotide was present. Note the simulation of the empty pore had a much higher output frequency and therefore contains more simulation frames; hence the bulk water densities are smoother.

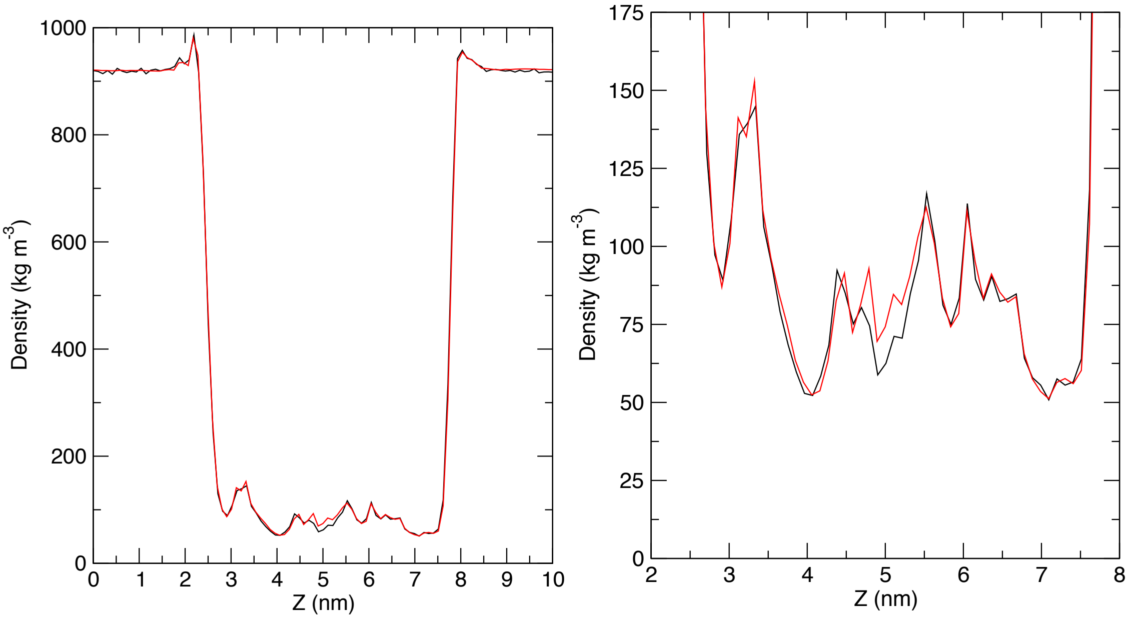


Figure 84: How the cytosine mononucleotide affects the local water density, zoomed in on the pore region on the right for clarity. The black line is p(ζ|K_5) and the red line is p(ζ|K_∞).

6.2.2 From p(ζ|K) to p(ζ|Z)

As water is being displaced due to the nucleotide translocation, we sought to investigate the energetics associated with displacing water. However there were two major issues to overcome; firstly the noise levels were significant and secondly, the nucleotide was not in a single position per simulation.

For the free energy of translocation, the data obtained, the probability density of water given a harmonic is not useful; the probability density of water given a nucleotide location is required. However, from p(Z|K), the nucleotide we can calculate p(ζ|Z), by taking averages as shown in Equation 21. This therefore shows how the water density changes as a nucleotide translocates through the protein.

$$p(\zeta|Z)=\frac{\Sigma_k p(\zeta|K)p(Z|K)}{\Sigma_k p(Z|K)}$$

Equation 21: The re-weighting scheme used to calculate p(ζ|Z).

As previously shown, each nucleotide density histogram occupies several z-coordinates and this collection of z-coordinates contributes to a (ζ|K) plot. These histograms overlap, as shown in Figure 85.

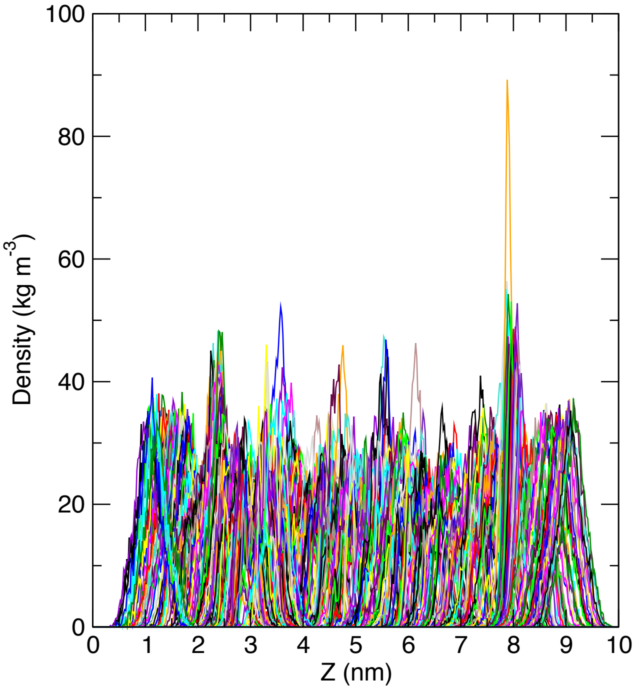


Figure 85: The nucleotide density plots show significant overlap between simulations.

For each Z co-ordinate between Z=1 nm to Z=9 nm there are approximately 58 simulations in which a nucleotide occupies that position in the box and for each of those simulations, there is a corresponding water profile. A scheme was devised by which each water profile was weighted by the nucleotide density for each Z co-ordinate occupied over the simulation shown in Figure 86.

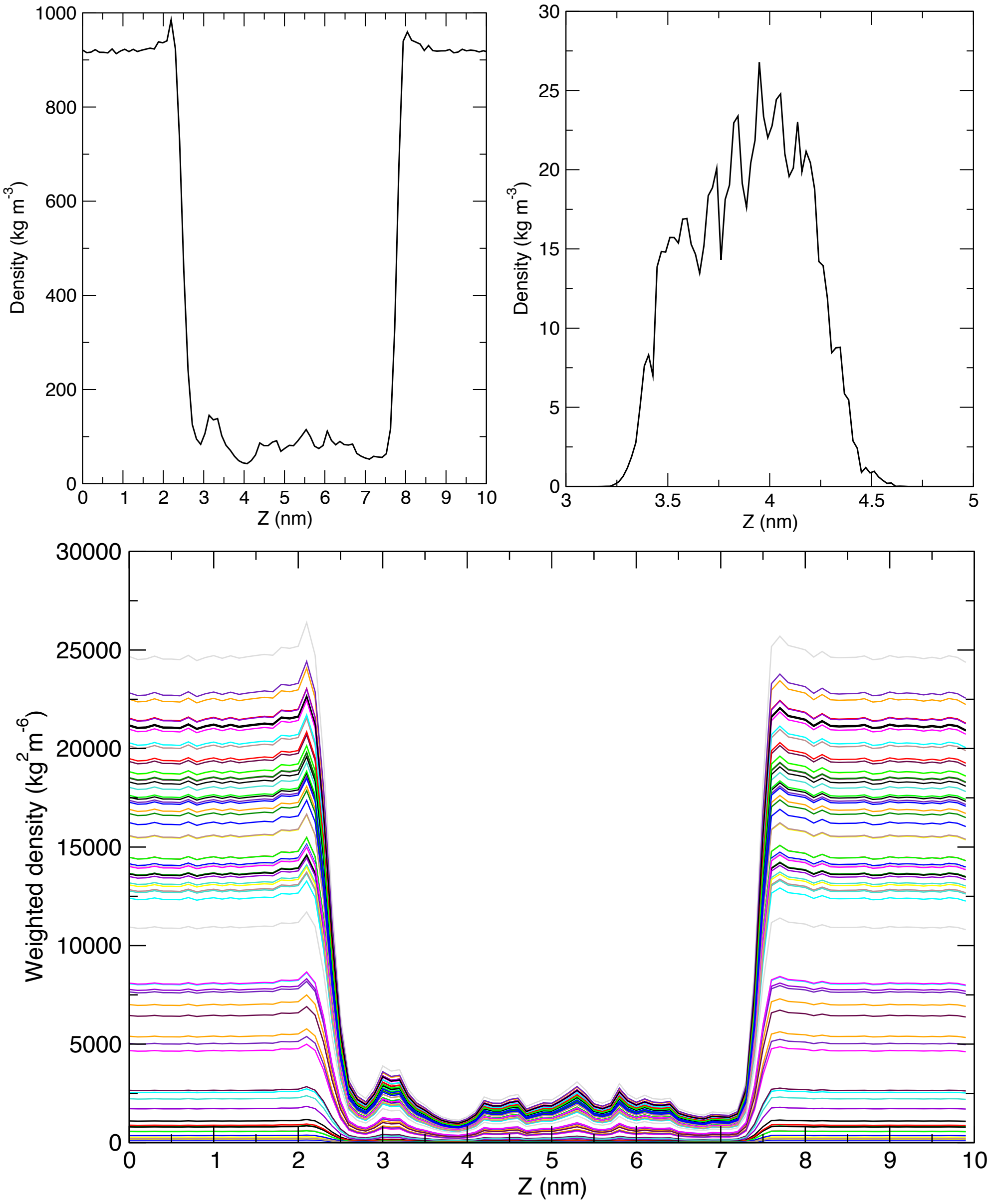


Figure 86: Each simulation provides (Top left) a water profile and (Top right) a nucleotide density plot, the water profiles were then weighted by the nucleotide density plot to provide weighted profiles (Bottom) for a given z co-ordinate.

The weighted water profiles for each Z co-ordinate were summated and normalised, by dividing by the sum of the nucleotide densities at that Z co-ordinate. This process is shown in Figure 87.

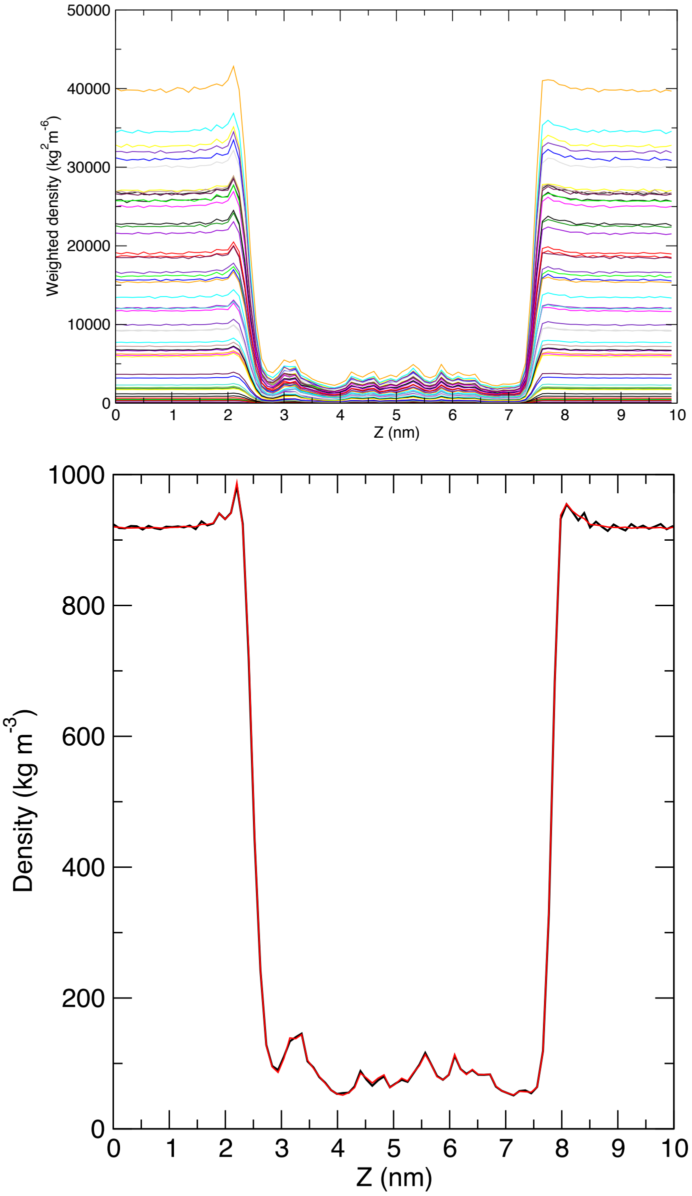


Figure 87: (Top) the weighted profiles for Z=5.5 nm, are summated and divided by the total nucleotide density in this region to provide a plot comparable to original p(ζ|K) plots. (Below) The p(ζ|$Z_{5.5}$) plot, in red, in comparison to the p(ζ|$K_{5.5}$) profile, in black, which provided the major contribution. Note the bulk water is much smoother for the normalised profile and there are minor differences in the pore region.

In summary, for every Z co-ordinate, a calculation is performed on every simulation that samples that Z value. Each p(ζ|K) is weighted by the height of the p(Z|K) histogram at that Z, these weighted ζ values are then summated and divided by the sum of the nucleotide density to provide an average $p(\zeta|Z)$. This process is shown in Equation 21.

6.2.3 Comparison between water density profiles

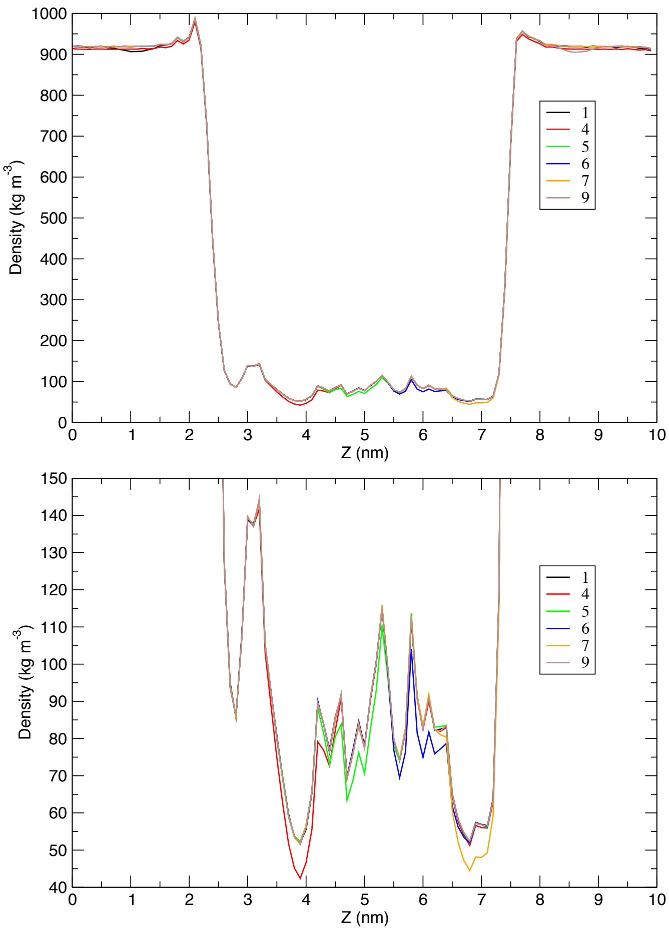


Figure 88: Plots of $p(\zeta|Z_1)$, $p(\zeta|Z_4)$, $p(\zeta|Z_5)$, $p(\zeta|Z_6)$, $p(\zeta|Z_7)$, $p(\zeta|Z_9)$. (Top) The plots over the entire box, note the minor drops in the bulk density caused by the nucleotide at Z_1 and Z_9. (Bottom) Zoomed in on the pore region, the nucleotides only cause a reduction in density in the 0.5 nm either side of their location.

This enables comparison of water profiles for nucleotide locations instead of for harmonic locations. This assisted observation of the degree of solvation in a given region; if the nucleotide did not reduce the water density it was likely to be in a dry pocket. Several p(ζ|Z) plots are compared in Figure 88. The p(ζ|Z) plots demonstrate that in αHL there is not an extended water network, as the reduction in water density is localised. Also, each p(ζ|Z) has a decrease in water density associated with that nucleotide location, suggesting that across the pore cavitation of the water occurs; there are no pockets large enough to accommodate a mononucleotide.

6.2.4 Free energy derivation

Equation 22 can be used to relate probability to free energy.

$$P(r)=\frac{1}{Y}e^{-\left(\frac{G(r)}{k_bT}\right)}$$

Equation 22: The relationship between the probability density of a state, P(r) relates to the partition function, Y, the free energy of the state, G(r), the Boltzmann constant, K_b and temperature, T.

Rearrangement of Equation 22 to make free energy the subject yields Equation 23.

$$G(r)=-k_bT\ln\big(P(r)\big)-k_bT\ln(Y)$$

Equation 23: The free energy of r, given the probability of r.

Free energies are more useful to express as the change between two states instead of just a single energy. So we subtract the free energy of a reference state, however laws of logarithms allow us to simplify this to a ratio; shown in

Equation 24.

$$\Delta G=-k_bT\,\frac{\ln(P(r))}{\ln(P(r_{ref}))}$$

Equation 24: The free energy change between state r and a reference state.

The data was normalised so over all measured space it sums to one and hence are probabilities. Therefore, for our system, we can now calculate the relationship between the energy required to displace water and p(ζ|Z), using Equation 25. Note the integral between the co-ordinates Z_{max} and Z_{min} is used to ensure we are measuring the change in energy across the pore region.

$$\Delta G_{\text{displacing water}}=-\Delta G_{\text{ water binding}}=-k_bT.\ln\left(\frac{\int_{Z_{min}}^{Z_{max}} p(\zeta|Z_1)}{\int_{Z_{min}}^{Z_{max}} p(\zeta|Z_\infty)}\right)$$

Equation 25: Relationship between probability and free energy of water binding.

Equation 25 shows the change in energy required to remove a water molecule from between the co-ordinates Z_{min} and Z_{max} with the nucleotide in position Z_1 in comparison to when the nucleotide is outside the pore. This therefore gives the energetic barrier the nucleotide must overcome to displace water during translocation through the pore.

6.3 Results

6.3.1 Cytosine mononucleotide translocating through αHL

The results of measuring the free energy of cytosine displacing water across the pore, with Z_{min} and Z_{max} set to Z=2.5 nm and Z=7.5 nm respectively, are shown in Figure 89. The profile produced is very smooth, suggesting the data is converged.

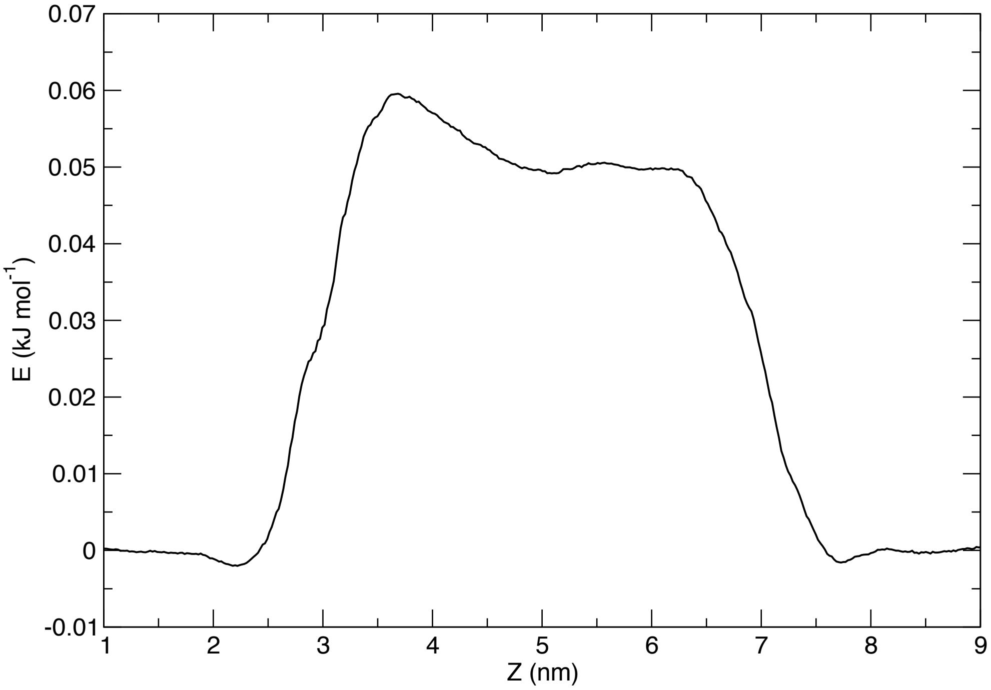


Figure 89: Change in the free energy of water binding associated with displacing waters as the cytosine mononucleotide translocates through αHL.

The energies observed due to displacing waters are very small, the largest is 0.06 kJ.mol^{-1} and therefore significantly less than k_bT, which is approximately 2.6 kJ.mol^{-1} at 310 K. The displacement of water is consequently not a meaningful obstruction to translocation, as these barriers will be overcome through thermal noise. This is not surprising, due to the large volume of the pore. Integration of the number density of water across the empty pore an average of 330 waters in the pore, whilst the maximum number of waters displaced is 8.5 at Z=5 nm. Therefore fewer then 2.6 % of waters are relocated from the pore and into the bulk.

As an error estimate, the free energy was calculated with a limited data set as an estimate of the error bars in a similar manner to bootstrapping. The data was split into alternating simulations and reveal the errors in the produced profile are negligible, these results is shown in Figure 90.

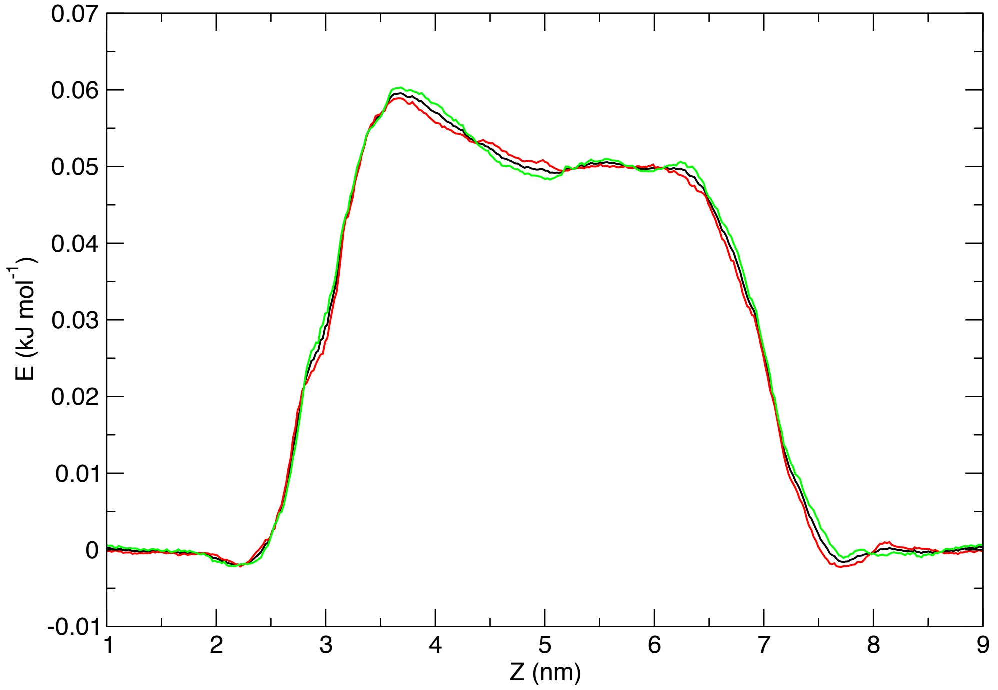


Figure 90: Analysis of partial data sets, red and green lines, reveal insignificant differences when compared to the total data set, shown in black.

Seeking to investigate if the water forms a structured network, the investigated region, Z_{min} and Z_{max}, was reduced. This allowed observation as to whether the nucleotide caused changes to the energetics whilst not in that region.

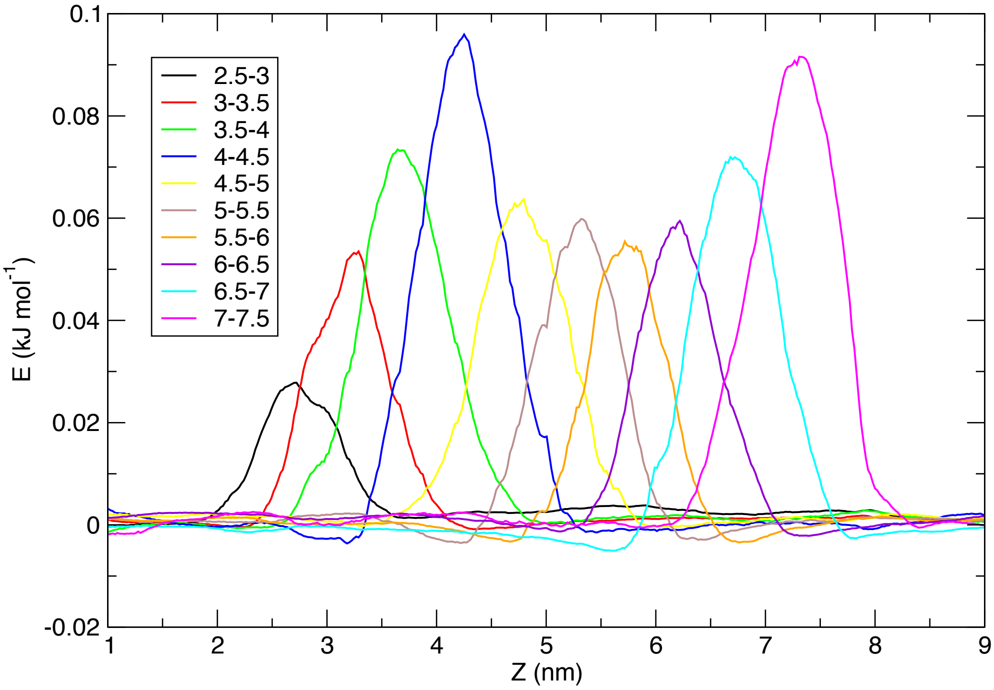


Figure 91: The free energy associated with displacing waters in specific regions as a cytosine mononucleotide translocates through αHL verifies the lack of a water network in αHL.

The pore was split into 10 regions of 0.5 nm each and the free energy of displacing water measured, shown in Figure 91. This shows that there is no extended water network as the disruption to each region is localised. There is no change in energy beyond the length of the nucleotide, 1.13 nm, away from the specified regions.

The energies observed for the 0.5 nm regions are larger than those observed across the entire pore. This is because when the disruption across the entire pore is measured, the change is comparatively smaller due to the lower proportion of waters affected by nucleotide.

6.3.2 Cytosine mononucleotide translocating through MspA

This technique was also applied to cytosine translocating through MspA, as shown in Figure 92. Again the energetic barrier that water provides is appreciably smaller than thermal error and therefore not a thermodynamically significant barrier to translocation. We suggest the more favourable energies observed at Z=8.5 nm are due to movement of the R96 residue at the exit of the pore. As the mononucleotide moves away from the pore, the arginine sidechain points towards the mononucleotide, instead of towards D93, enabling additional waters to enter the pore.

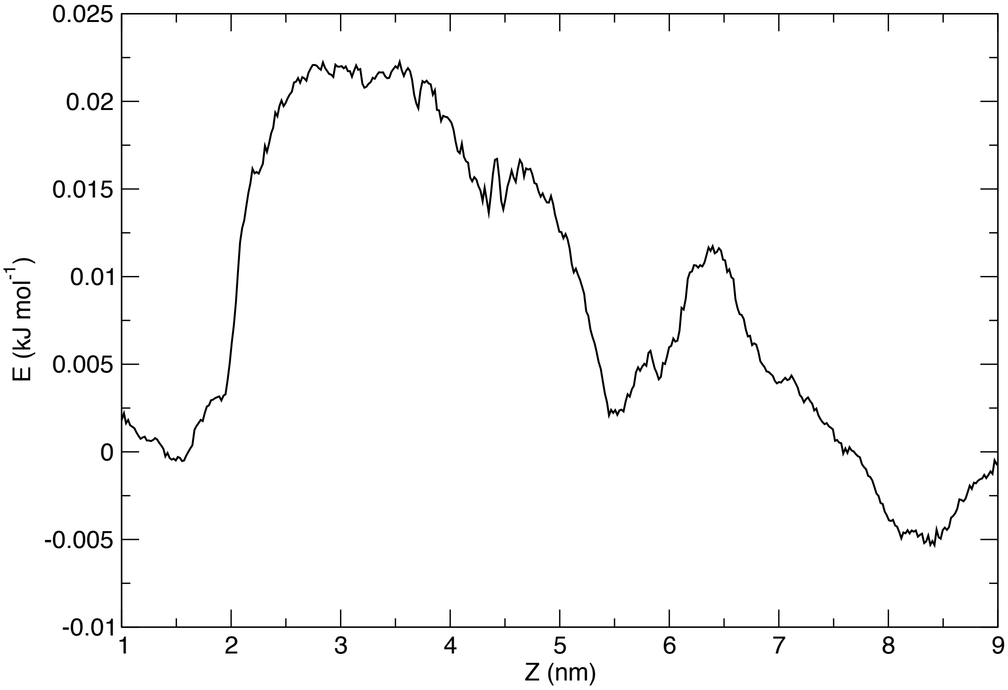


Figure 92: The energy of displacing water for cytosine translocating the MspA A96R model pore.

The energetics of this profile are much noisier than observed for αHL, this is due to the decreased number of snapshots output by the simulations. For αHL the co-ordinates were output every 200 ps, whilst for MspA they were output every 1000 ps. This means the water density profiles contain a fifth of the data and therefore do not average as smoothly for MspA as they do with αHL.

As a reference, the locations of inwards facing residues and the regions of MspA are shown in Figure 93.

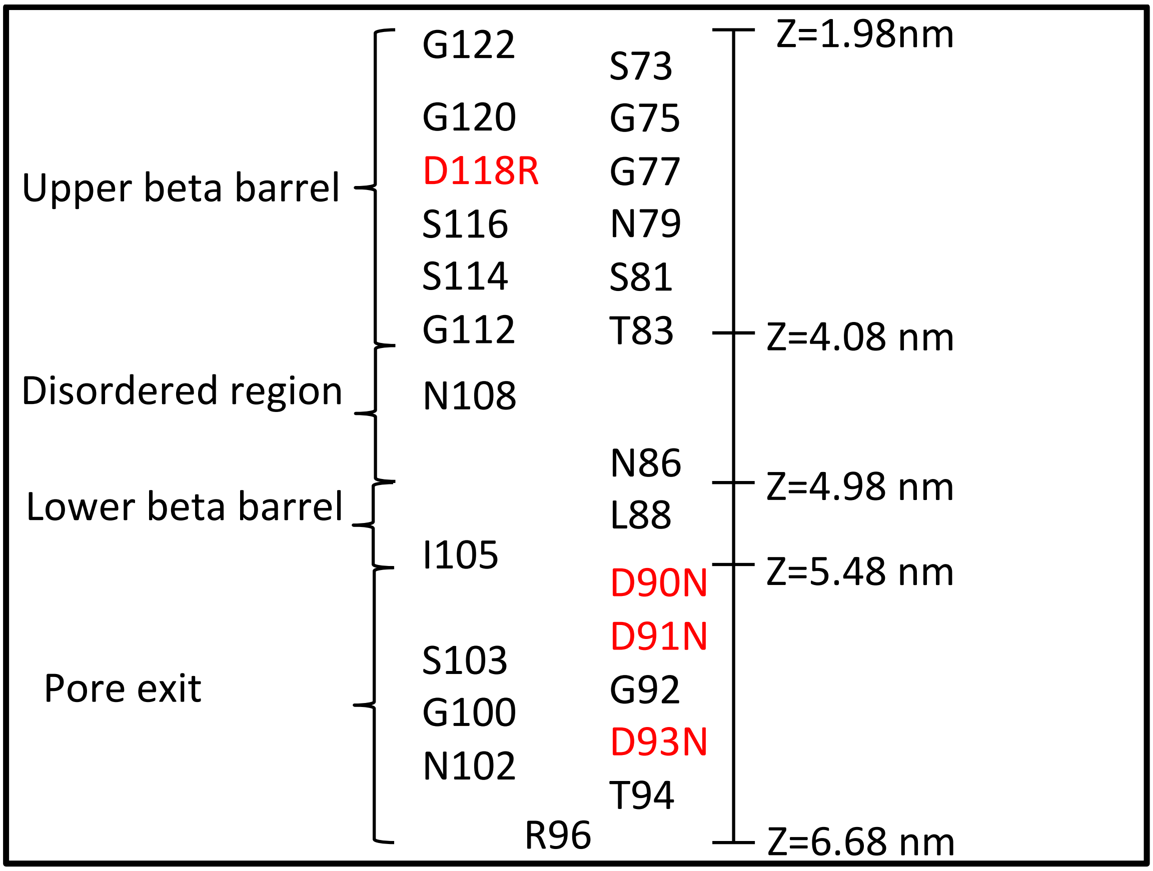


Figure 93: The Z co-ordinates of residues for MspA, with mutated residues shown in red.

6.4 Conclusion

A new analysis method is presented, which enables users to study changes in density across a reaction co-ordinate and the associated free energy for that density change.

Our investigation shows that water provides a small energetic barrier, that is easily overcome by thermal noise, to the translocation of a cytosine mononucleotide through either αHL or MspA.

We suggest only the energetics of the DNA and protein needs to be considered when discussing translocation through protein nanopores. Also, with mutation of αHL or MspA we would suggest significant changes are required to make displacement of water a barrier to translocation.

6.5 Future work

Further testing on systems should be performed and compared to literature values, we suggest an ion channel such as the potassium channel from *Streptomyces lividans* [199], as loss of solvation is know to be important energetically in this example. This channel is much smaller than those of studied in this chapter, only 0.3 nm compared to 1.2 for αHL and 1.4 for MspA, as such a nucleotide would be unsuitable and potassium should be the studied substrate.

More robust error analysis should be implemented, such as bootstrapping and block analysis. This would enable accurate error bars to be calculated and convergence testing to be performed.

Chapter 7 Nanopore Design

Abstract

Despite nanopores being the subject of much scientific scrutiny, there has not been a thorough computational investigation into how the internal charge distribution affects the ability of the pore to sequence DNA. Experimental studies [86, 102, 200], can assist in designing nanopores, however they are not able to provide insight into why the DNA responds on the molecular level, leading to potential problems for sequencing such as secondary structure formation [126]. Using a toy system, synthetic nanopores were designed to break down the complex relationship between the number and location of charges with speed of translocation and degree of secondary structure. The process of building our system towards a more protein-based representation was then started, to show these models are still valid with amino acids as well as point charges.

From this study, it was revealed that the location is as important as the charge of the ion introduced, with negative charges able to slow DNA translocation as much as a positive charge for certain locations in the pore. This is in contrast to current opinion, which dictates that positive charges should be used to slow DNA and negative charges being commonly mutated out [86, 94, 169, 200, 201].

To facilitate our studies into longer strands of DNA, a method to create periodic DNA with no terminal residues is discussed. This allows for simulations to be designed to study bulk DNA, not near terminal ends.

7.1 Introduction

As previously discussed, in 1.3.4, mutations are introduced to αHL [101] and MspA [85] in order to reduce the rate of DNA translocation, allowing for accurate discrimination between bases [86]. There has not yet been a systematic computational study into the complex relationship between the number of charges in the nanopore, their location and how DNA translocates through the nanopore.

Intuitively, it is expected that positive charges on the nanopore would promote binding, through electrostatic attraction to the DNA phosphate groups, which would result in a decrease in translocation rate, whilst the opposite is expected for negative charges [98, 126, 134, 201]. It is not immediately apparent as to how the location of a charge would influence translocation; both Maglia et al [200] and Rincon-Restrepo [86] et al demonstrate that multiple positive charges result in reduced rate of translocation, down to 1/18th the rate observed with the wild-type protein. Due to the presence of other binding locations introduced in these studies it is not clear as to the effect of individual charges.

From our PMF studies we predict that there is approximately a 10 kJ.mol^{-1} reduction in free energy from entering the αHL pore and a similar increase from leaving, therefore altering the energetics in these regions will likely have more influence than elsewhere. A project to establish how charges influence the rate of DNA translocation was undertaken, whilst also investigating how such charges influence the degree of secondary structure present.

In the context of DNA translocation, simulations have been limited to studying small strands, up to 60 bases long [134], with even the longest strands being orders of magnitude smaller than those in experimental work, which have used over 6000 bases [126, 134, 201]. Due to the limited length of the DNA used, there is a limited duration simulations can be run for; once translocation is complete it is more efficient to run a repeat than to wait for capture to occur through the periodic image. A solution would be to increase the length of the DNA, but this is only practical to a certain point as for longer DNA strands the system size must be increased, thereby slowing the simulation.

In order to avoid the limited duration of translocation experiments, and the requirement of a large quantity of data to calculate statistically useful results, the length of the DNA strands had to be increased without influencing the size of the simulation box. This was achieved through the use of DNA bound across periodic images, which results in an infinite length polymer. This idea came from simulations performed on carbon nanotubes [202], these studies recommended that the system must be run in the NVT ensemble, as pressure coupling along the principle axis of the polymer result in stress along the molecule.

7.2 Method

7.2.1 Generation of DNA strands

B-DNA helix co-ordinate files were downloaded from an online resource [203]. The structures were then edited to be recognisable by the GROMOS 53a6 forcefield prior to simulation setup, as described in appendix D. The DNA strand was then extended, to the length of the box, 45 nm, which was performed by applying an electric field of 0.1 V.nm^{-1} to the helix for 10 ns in 1 M NaCl solution, with the 5’ terminus restrained, as shown in Figure 94.

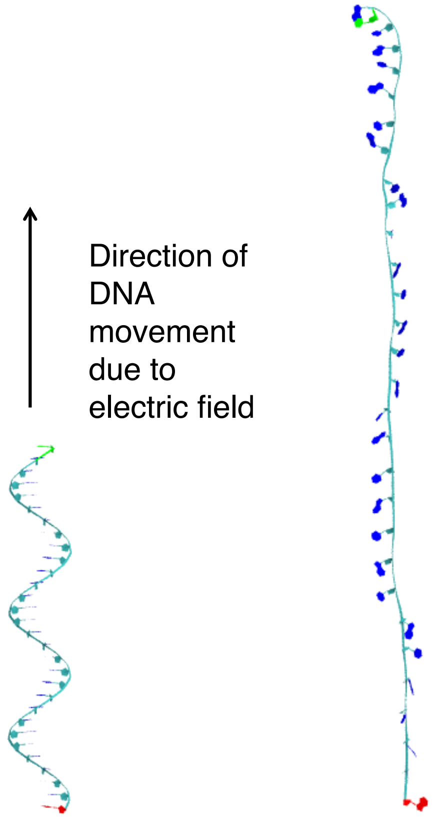


Figure 94: A downloaded DNA strand, prior to, and post straightening, with the 3’ terminal residue in green and the 5’ terminal residue in red.

Periodic DNA was created by removing the terminal residues from the structure, the topology of the strand was edited to create bond definitions between the two new terminal ends, as summarised in Figure 95. Therefore any definitions required between the two residues, such as dihedrals, are identical to those through the remainder of the strand.

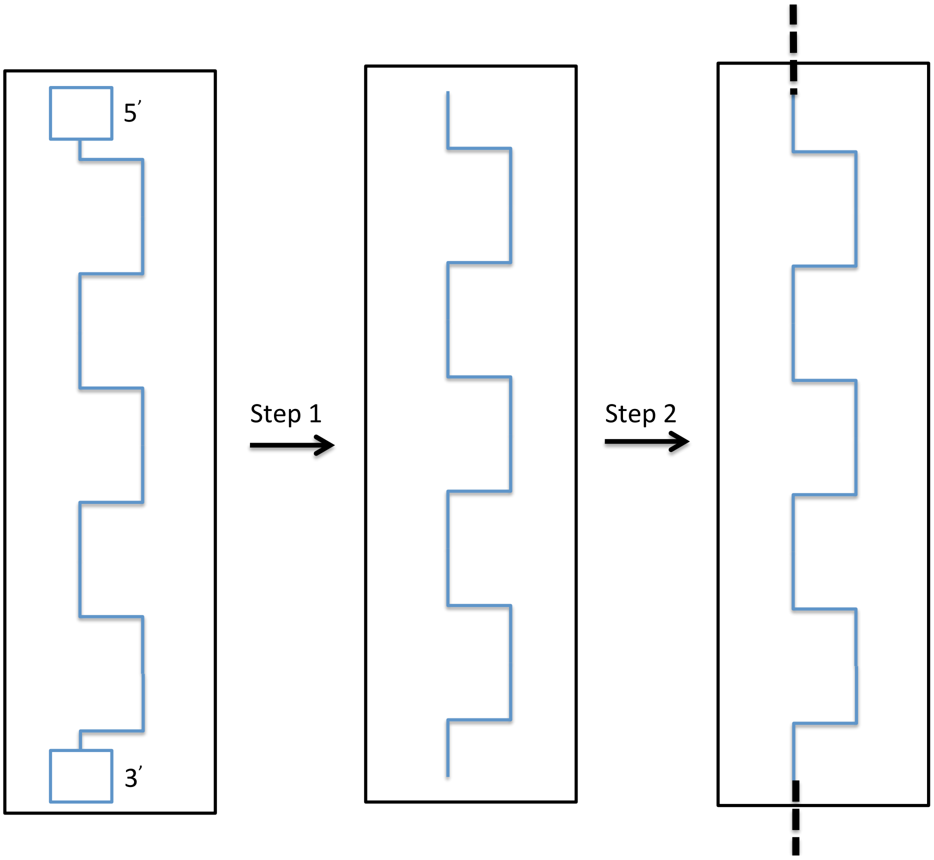


Figure 95: Schematic overview of infinite DNA setup. The black boxes represent the periodic boundary whilst the blue lines represent a DNA helix. In the first step, the DNA helix terminal ends are removed. In the second step bond definitions between the first and last residues are introduced across the periodic boundary.

Despite the infinite length of the DNA, there are a set number of bases in each periodic image, hence for poly(A)$_{60}$ the strand contains 60 bases in a single periodic image, but with no terminal ends. This is not the only method for creating infinite length DNA, another group [119] used the CHARMM [204], and NAMD [205] codes by utilising “pseudo-covalent bonds” from IMPAtch. It appears that the linker they use to attach their strand to itself across a periodic image is in fact just a harmonic potential, but using our method the full DNA bond definitions are used, meaning the residues are governed by the force-field in the same manner as those between every other nucleotide pair.

7.2.2 Pore design

Pores were created via a script provided by the Bond group, which creates a square of atoms around concentric rings. This script allows for adjustment of the width of the pore, the number of rings, space between atoms and thickness of the pore. This is the same script from which the methane slabs for the model pores were created [98]. The resulting pores are made of united atom methane; this was done for simplicity as these systems are designed to merely mimic hydrophobic interactions. The pores have a 3.2 nm diameter throughout and are 3 nm thick. The membrane mimetic slab is 11.4 by 11.4 nm. Systems were set up with various charge distributions, this was done by adding charges to a ring of 5 atoms in the pore as shown in Figure 96. Rings of charges were added at various heights in the pore, at one of the following locations the cis entrance, between the cis entrance and the centre of the pore, the centre of the pore, between the centre of the pore and the trans exit and finally, at the trans exit. Additional systems were created with a ring of charges at the cis entrance and another at the trans exit. In these simulations, DNA translocates from the cis entrance to the trans exit.

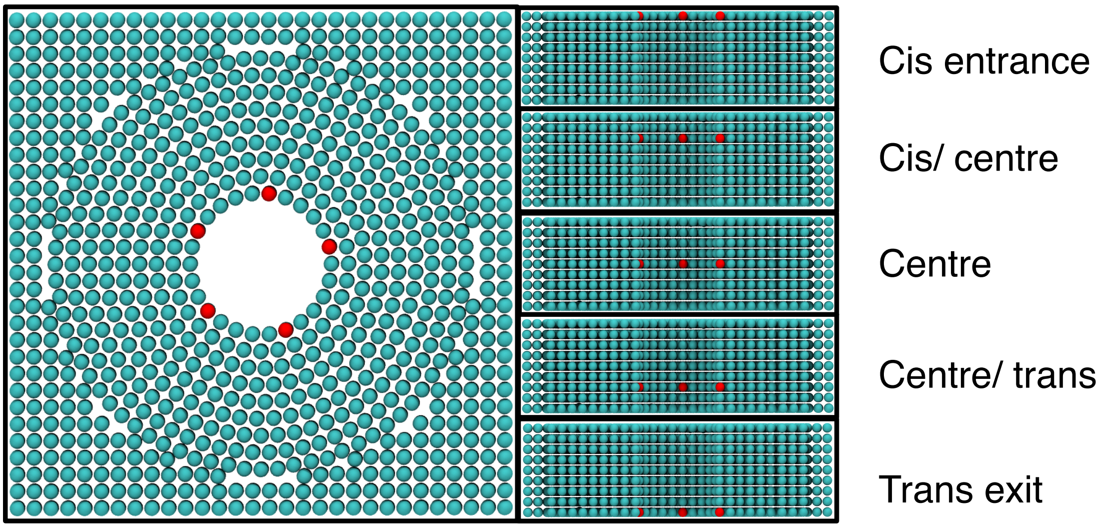


Figure 96: (Left)The locations of the charges looking down through the nanopore, for the charge distribution simulations. (Right) from top to bottom, the heights denoted by the terms “cis”, “cis/centre”, “centre”, “centre/trans” and “trans” of the nanopore. The red spheres represent the charged atoms.

7.2.3 DNA translocation

Periodic DNA was used in order to replicate large scale DNA passing through the pores, replicating that the majority of translocating DNA is not near terminal ends. A box length of 45 nm was used and 2 strands of poly(A) were created, with periodic images of 60 and 120 bases each respectively, shown in Figure 97. These lengths were chosen as for this box length they allow observation of DNA with varying degrees of secondary structure; poly(A)$_{60}$ has no secondary structure whilst poly(A)$_{120}$ starts approximately as a B-DNA helix. Note for this box length, 45 nm, a B-DNA helix would be most stable with 13.6 turns and therefore 136 bases, poly(A)$_{120}$, in comparison would be most stable with 11.4 turns and therefore has a minor negative twist induced, of approximately -2.2 turns from a standard B-DNA helix. This twist will promote writhe of the ssDNA [17] and the formation of secondary structures as discussed in section 1.1.3.

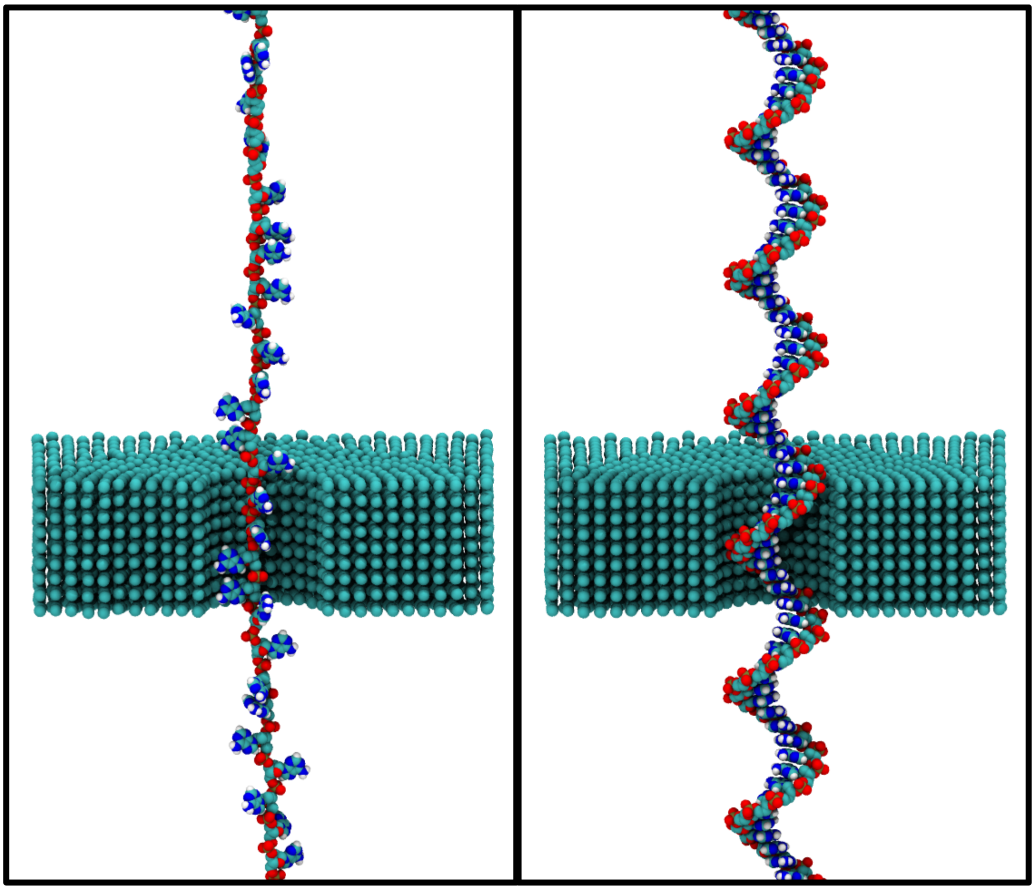


Figure 97: Initial setup for periodic DNA translocation, (Left) for poly(A)$_{60}$ and (Right) poly(A)$_{120}$.

These simulations were equilibrated in the NVT ensemble using the v-rescale thermostat [161] at 300 K for 1 ns. 3 repeat simulations were run for 105 ns and an electric field of 0.1 V.nm^{-1}, equivalent to 300 mV across the pore. The first 5 ns of simulation data were discarded to reduce the influence of the DNAs starting co-ordinates and velocities.

7.3 Results

7.3.1 Point charges

7.3.1.1 Rate of DNA translocation

The rate of DNA translocation was measured as the distance, per nanosecond, the phosphorus in the backbone travelled in the direction of the pores principle axis, averaged over the simulation length. The 3 repeats of 100 ns were split into 10 ns blocks, from which averages and standard error of the mean was calculated. These results are shown in Table 2.

Pore charge and location	Rate of translocation for poly(A)$_{60}$ (nm.ns^{-1})	Rate of translocation for poly(A)$_{120}$ (nm.ns^{-1})
Neutral	5.83±0.08	3.67±0.09
Positive (cis)	4.39±0.11	3.43±0.09
Positive (cis/centre)	5.33 ± 0.17	3.65±0.11
Positive (centre)	5.49±0.17	4.04±0.14
Positive (centre/trans)	6.21 ±0.15	3.84±0.08
Positive (trans)	5.49±0.09	3.59±0.12
Negative (cis)	6.17±0.10	4.09±0.06
Negative (cis/centre)	5.89±0.11	3.92±0.06
Negative (centre)	5.41±0.15	3.42±0.10
Negative (centre/trans)	4.52±0.13	3.60±0.06
Negative (trans)	5.13±0.10	4.21±0.12
Positive (cis) and negative (trans)	4.24±0.17	3.48±0.09
Negative (cis) and positive (trans)	6.60±0.14	4.27±0.09

Table 2: The average rate of DNA translocation over three 100 ns repeats.

Initially, the rate of DNA translocation through the pores were inspected against time, shown in Figure 98, in order to validate that the DNA was not subject to a constant acceleration. This proves that the systems did not accelerate over time; rather the rate of translocation varied around the mean as expected.

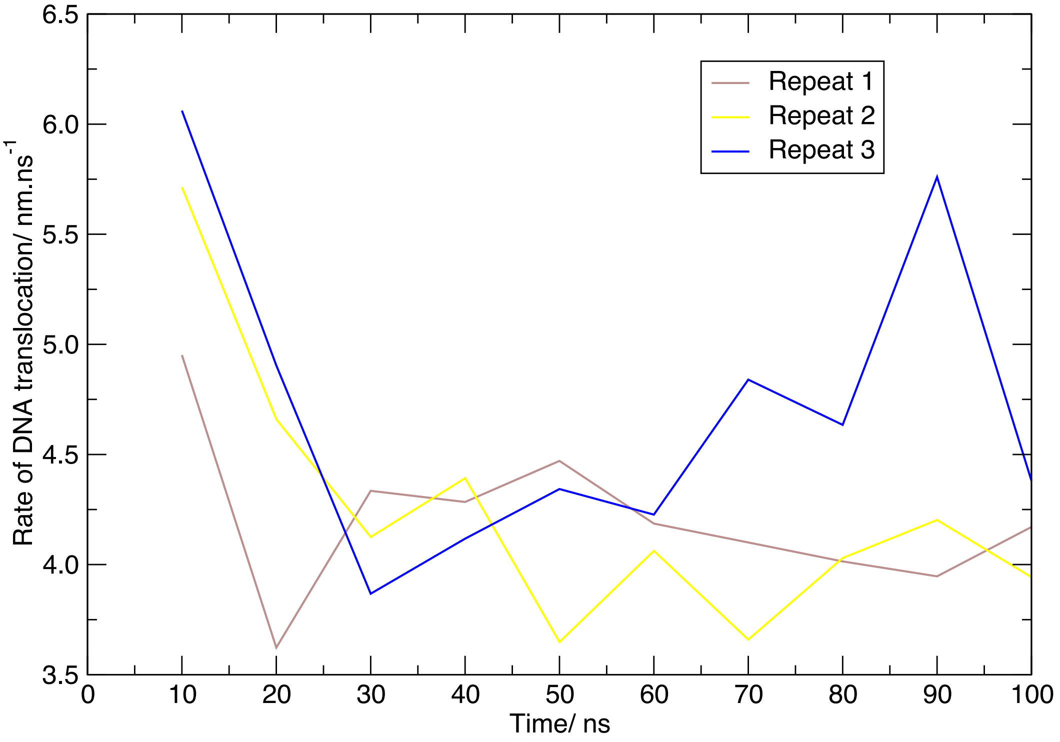


Figure 98: The rate of translocation of DNA through the pore with positive charges in the cis location against time.

In all systems, the poly(A)$_{120}$ translocates through the nanopores slower than the poly(A)$_{60}$, in general poly(A)$_{120}$ translocates at 72±10% the rate of poly(A)$_{60}$. This is due to the formation of secondary structures with poly(A)$_{120}$, which, as expected, cannot form for poly(A)$_{60}$, which create “drag” through the solvent, resulting in slower translocation.

For poly(A)$_{60}$ positive charges in the nanopore have a marked difference on the rate of translocation, although this is very location dependent. A positive charge at the cis entrance of the pore slows the DNA but this effect reduces as the charge is moved towards the trans exit of the pore. Notably positive charges introduced at the "mid/trans" location have the opposite effect on speed to the other charge locations, increasing the rate of translocation, although this is following the trend of charges lower in the pore resulting in faster translocation. Also, the middle of the pore and the trans exit appear to have the same effect, suggesting the exit to the pore has unique significance on the energetics of translocation as it breaks the observed pattern. When negative charges are introduced to the pore, unsurprisingly an opposite trend is observed. Negative charges towards the cis entrance result in increased rate of translocation, whereas charges further down the pore result in a reduced rate. Interestingly, again the trans exit appears to break the observed trend, emphasising that the energetics associated with exiting the pore are of great importance to the rate of translocation. These rates are compared in Figure 99.

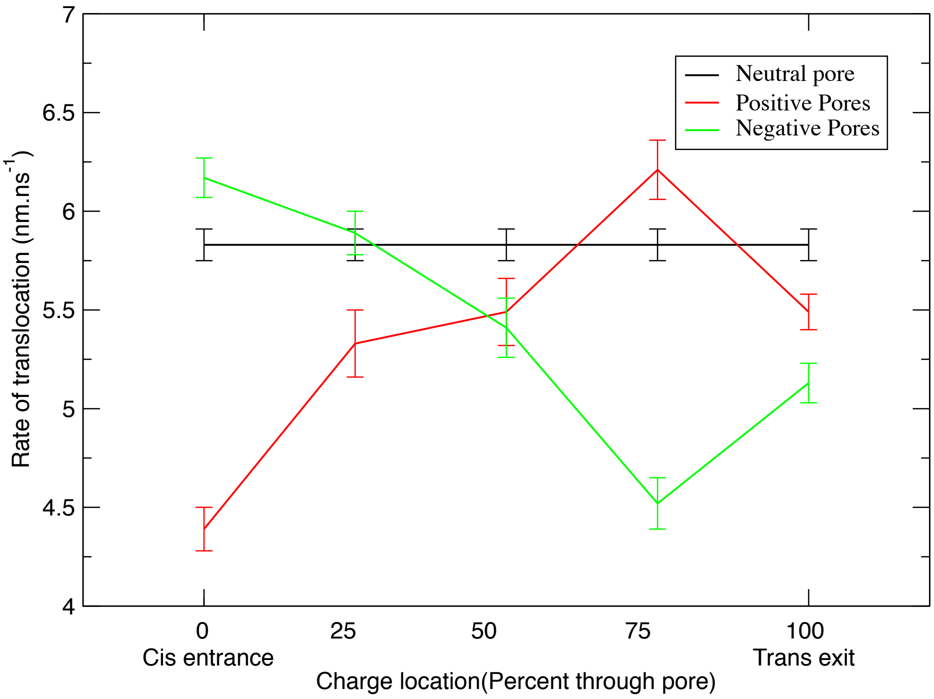


Figure 99: How the location of charges influences the rate of translocation for poly(A)$_{60}$. Note that positive and negative charges can reduce or increase the rate of translocation.

We suggest that these changes are due to energetics; at the entrance to the pore, positive charges encourage binding although negative charges reduce the favourability of entry to the pore. At the centre of the pore, positive charges slow DNA by binding, whilst negative charges introduce a barrier to permeation that must be overcome. At the trans exit, positive charges make the translocation process more difficult as the DNA has to overcome the energetics of exit, in addition to unbinding from the charge, breaking the observed trend. In contrast, although negative charges increase the energy, it may be that this makes the process of pore exit two steps; with a smaller barrier for each individual step.

Secondary structures were then introduced to the pore by using the poly(A)$_{120}$ DNA strand and re-measuring the rate of DNA translocation. The positive cis entrance results in slower translocation in comparison to the neutral pore, the rate of translocation then increases with the “cis/middle” charged location having negligible effect and the centre of the pore resulting in the highest rate of translocation. Beyond the centre of the pore to the trans exit the rate then reduces, although slightly slower than from the centre to the cis entrance. The negative charges have the opposite effect on poly(A)$_{120}$, the curve is now inverted, with the cis entrance and trans exit resulting in an increased rate of translocation compared to the centre. In this case the “mid/trans” location has negligible effect in comparison to the neutral pore. These rates are shown in Figure 100.

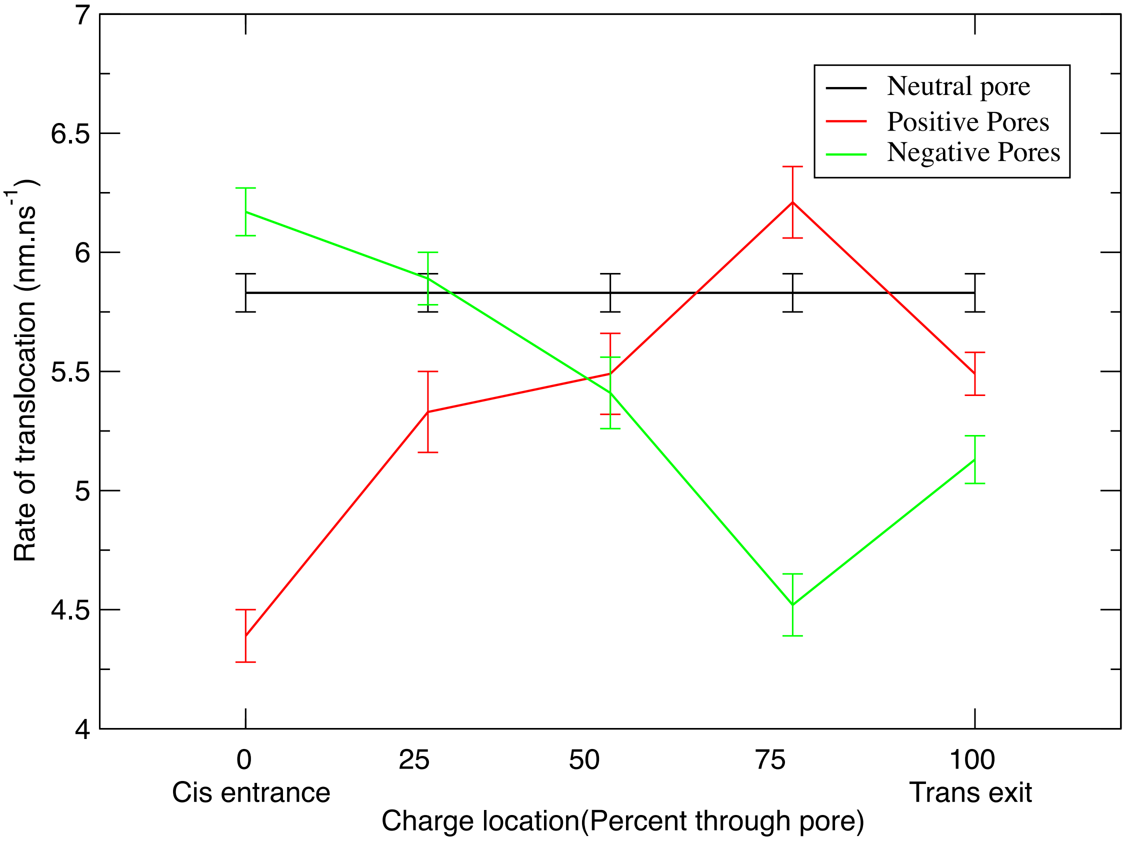


Figure 100: How the location of charges influences the rate of translocation for poly(A)$_{120}$. The negatively charged pores produce a v-shaped curve, with the middle of the pore as the minimum, whilst the positively charged pores produce the mirror, "n-shaped" curve.

The differences between the plots of rate of DNA translocation of poly(A)$_{60}$, Figure 99, and poly(A)$_{120}$, Figure 100, are two fold; the magnitude of the differences and the shapes of the curves. Firstly, the differences in rate of DNA translocation caused by charge with poly(A)$_{60}$ are approximately double those with poly(A)$_{120}$, this is likely due to the poly(A)$_{60}$ translocating more rapidly and therefore energetic barriers creating a larger effect. Regarding the shapes of the curves, as secondary structure is increased the curve is shifted left compared to the poly(A)$_{60}$. We speculate this is due to the increased secondary structure with poly(A)$_{120}$; the centre of the pore influences exerts a larger influence when the number of bases in the pore increases. This is different from poly(A)$_{60}$, where there are fewer bases in the pore and exit from the pore is a more dominant energetic barrier.

Testing for cumulative effects was performed by producing pores with two rings of opposing charge at the cis entrance and trans exit. For poly(A)$_{60}$ the pore with positive charges at the cis entrance and negative charges at the trans exit results in slower DNA than for their singly charged counterparts, both of which are slower than the neutral pore. In contrast the pore with negative charges at the cis entrance and positive charges at the trans exit increase the speed of translocation more than each individual pore. In this case the negatively charged pore increases the rate relative to neutral and the positive pore slows translocation down but to a lesser extent than the other locations with positive charges. With poly(A)$_{120}$ where the composite pores have contradictory effects, we observe that the resulting rates of translocation predominantly resemble that of the charge at the cis entrance. This suggests that when secondary structures are present, alterations to the energetics of pore entry outweigh those to pore exit.

7.3.1.2 Secondary structure

Localised bends in the DNA cause the DNA to exit in a different order to that which it entered in, resulting in difficulties for sequencing [126], secondary structures can also completely stop DNA from translocating [94]. Therefore the degree of secondary structure was analysed.

The degree of secondary structure was calculated as the average number of phosphorus atoms in the pore, over the simulation length. For the purposes of this calculation “in the pore” is defined by the co-ordinates of the atoms comprising the pore entrance and exit. Phosphorus was chosen so that comparison between the various nucleic acid monomers for future work would be trivial.

Pore charge and location	Average number of bases in pore for poly(A)$_{60}$	Average number of bases in pore for poly(A)$_{120}$
Neutral	28.25±0.04	42.80±1.36
Positive (cis)	28.33±0.04	42.37±0.74
Positive (cis/centre)	28.29±0.04	39.60±0.77
Positive (centre)	28.29±0.04	45.42±0.71
Positive (centre/trans)	28.25±0.04	37.40±0.73
Positive (trans)	28.29±0.04	43.50±0.95
Negative (cis)	28.22±0.04	36.93±0.58
Negative (cis/centre)	28.23±0.04	37.40±0.78
Negative (centre)	28.27±0.04	44.±1.27
Negative (centre/trans)	28.31±0.04	39.39±0.71
Negative (trans)	28.28±0.04	38.45±0.97
Positive (cis) and negative (trans)	28.33±0.04	39.23±0.77
Negative (cis) and positive (trans)	28.21±0.04	38.69±0.69

Table 3: The average number of nucleotides in the pore over three 100 ns repeats.

As observed in Table 3, there is negligible variation in the number of bases in the pore for poly(A)$_{60}$. The error bars between all simulations overlap and therefore we assume that for DNA under tension the internal charge of the nanopore has no significant effect on the formation of secondary structures.

As expected due to the increased density, with poly(A)$_{120}$ there are a more bases in the pore on average, corresponding to secondary structures forming and translocating through the pore. These secondary structures also cause a larger variation in values and therefore the large increase in error compared to poly(A)$_{60}$.

The secondary structures are not similar to a triple helix and the adenine does not form base pairs. The structures mainly seem to be stabilised through base stacking with very limited intramolecular hydrogen bonding. An example of a secondary structure is shown in Figure 101.

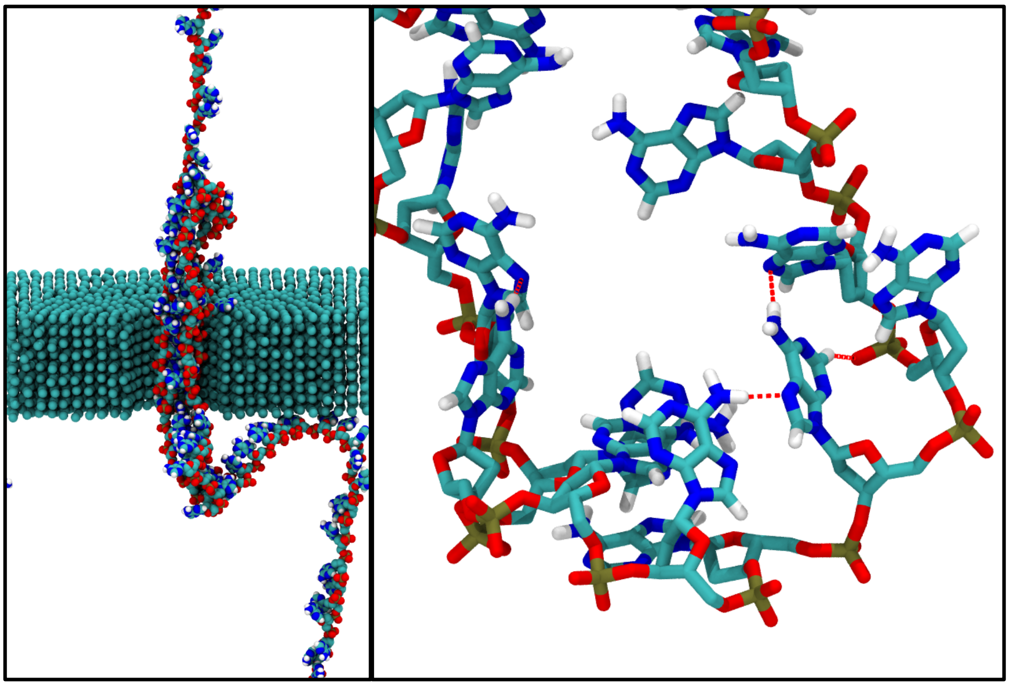


Figure 101: Examples of the secondary structures formed, (Left) translocating through the pore, (Right) the 4 of the 6 inter-strand hydrogen bonds, red dashed lines, are formed at the hairpin at the bottom of the structure in the Hoogsteen positions between bases, or between the nucleoside and the phosphate group.

With positively charged pores, the degree of secondary structure does not appear to be correlated to location within the pore. Charges at the cis entrance and trans exit are within error of the uncharged pore and therefore have no measurable effect. In comparison a positive charge at the centre of the pore increases the degree of secondary structure by 5 %, while those at the cis/centre and centre/trans locations reduce it by 7.5 % and 12.5% respectively. For negatively charged pores, in general the degree of secondary structure is reduced. This effect is weakened as the charge moves towards the centre of the pore, from either the cis entrance or trans exit, with the centre resulting in no change to the degree of secondary structure, within error of the neutral pore.

With pores containing both positive and negative charges, a decrease in secondary structure is observed. The magnitude of the decrease is smaller than when just negative charges were introduced, suggesting the outcome is an average of the two composite pores.

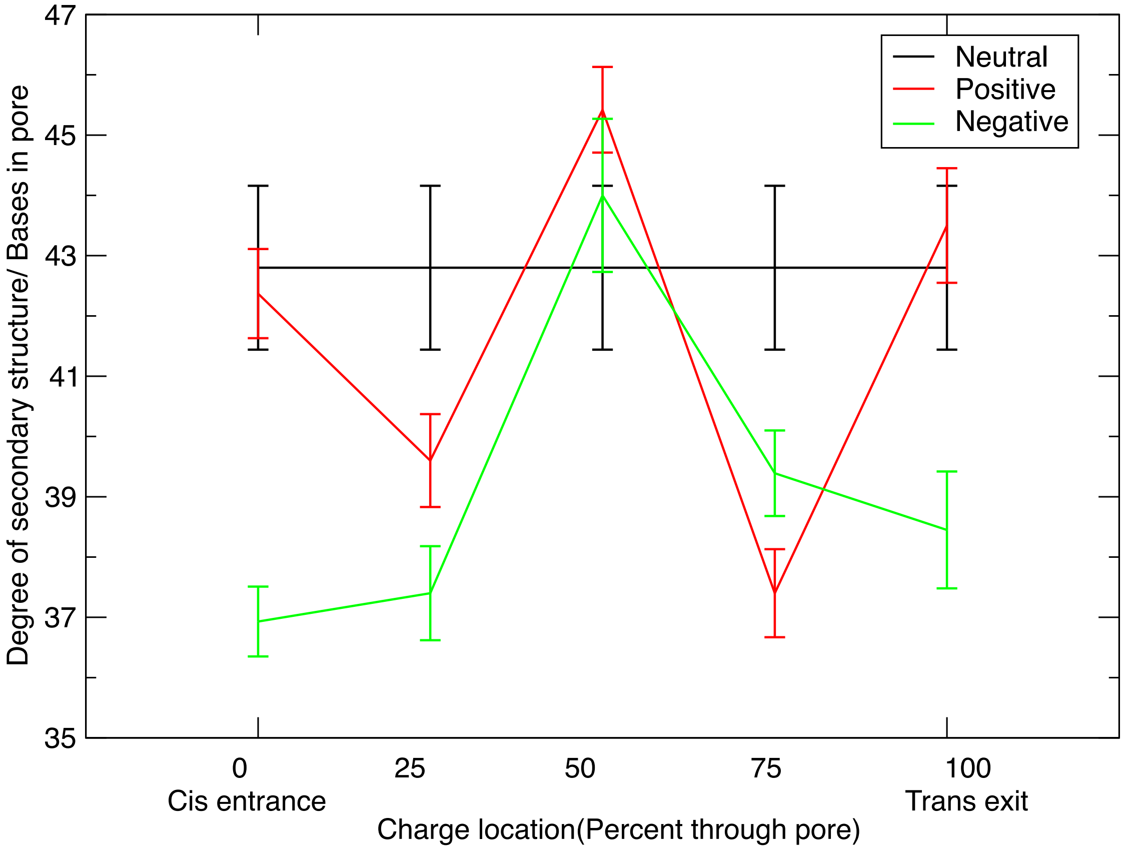


Figure 102: The effect of positive and negative charges in the nanopore on the degree of secondary structure based on their location.

The number of bases in the pore cannot easily be correlated to either the rate of translocation or the expected energetic effect of the added charges. With negatively charged pores, those pores that increase the rate of translocation result in a decrease in the degree of secondary structure, however this relationship does not hold true for positive pores. Similarly, we would expect negative charges to result in a decrease in secondary structure, as there is no preferable energetic binding location, which is observed, but in comparison positively charged pores do not increase the degree of secondary structure, going against the logic that DNA would attempt to bind to these regions.

7.3.2 Amino acid constructs

The work so far has demonstrated the ability of point charges to influence translocation, however point charges do not represent protein pores well. Amino acids contain flexible sidechains capable of moving several angstroms, especially under the influence of an electric field. In the study with point charges, we emphasised the importance of location with charges separated by 0.6 nm. One set of simulations was repeated with amino acids in place of point charges. The positively charged lysine amino acid was chosen due to its similarity to a point charge. Other amino acids contain more de-localised charges, for example the carboxyl group in acidic amino acids have the negative charge split over both oxygen atoms.

We therefore decided to attempt to replicate our previous simulations with positively charged pores, by replacing the positive point charges with 5 lysine residues in separate chains i.e. amino acid monomers. The pores were set up, comparably to the previous pores, with 5 charged residues spread equidistantly around the pore and at either the cis entrance, trans exit or centre of the pore and are shown in Figure 103.

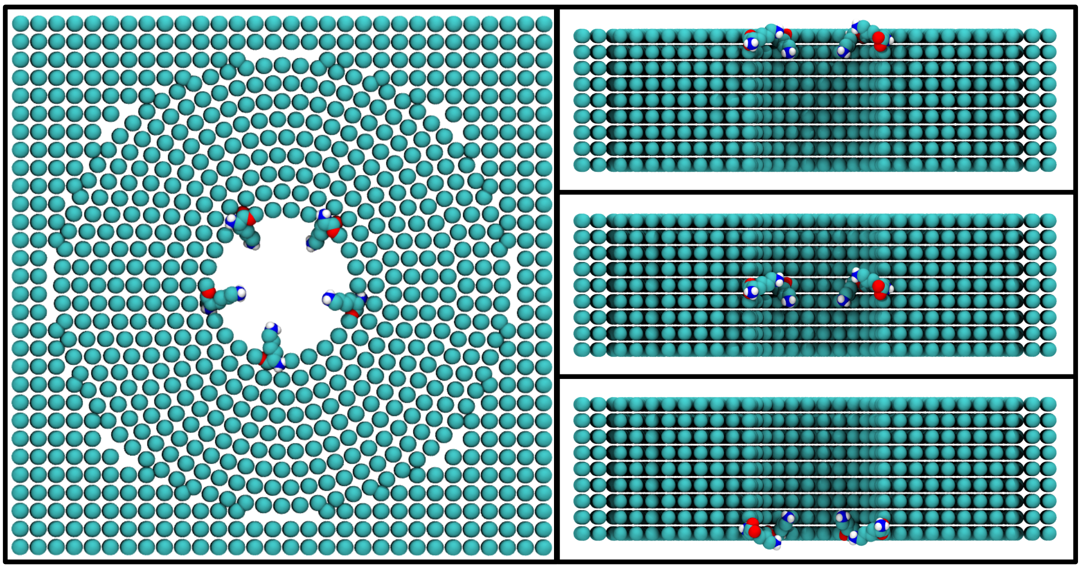


Figure 103: An overview of the charged pores containing lysine. (Left) through the pore, note the backbone of the amino acids are parallel to the membrane-mimetic slab, so there is no net effect from the charged carboxylic acid or amine groups of the backbone. (Right) The height denoted by the terms “cis”, “centre” and “trans”.

7.3.2.1 Rate of DNA translocation

Pore charge and location	Rate of translocation for poly(A)$_{60}$ (nm.ns^{-1})
Positive (cis)	4.39±0.11
Positive (cis/middle)	5.33 ± 0.17
Positive (middle)	5.49±0.17
Positive (mid/trans)	6.21 ±0.15
Positive (trans)	5.49±0.09
Lysine cis	5.22 ± 0.06
Lysine middle	6.07 ± 0.06
Lysine trans	6.83 ± 0.04

Table 4: The rate of DNA translocation, compared between positive point charges and lysine based pores.

The trend for amino acid based pores is similar to those observed with point charges, with an increase in speed as the charge is moved down the pore. The rate of translocation is higher with amino acids in comparison to point charges. The error bars are smaller than those observed for point charges and in addition the exit to the pore doesn't have a reduced effect. These differences between the rate of translocation for lysine pores and point charges can be somewhat accounted for by the length and flexibility of the sidechains.

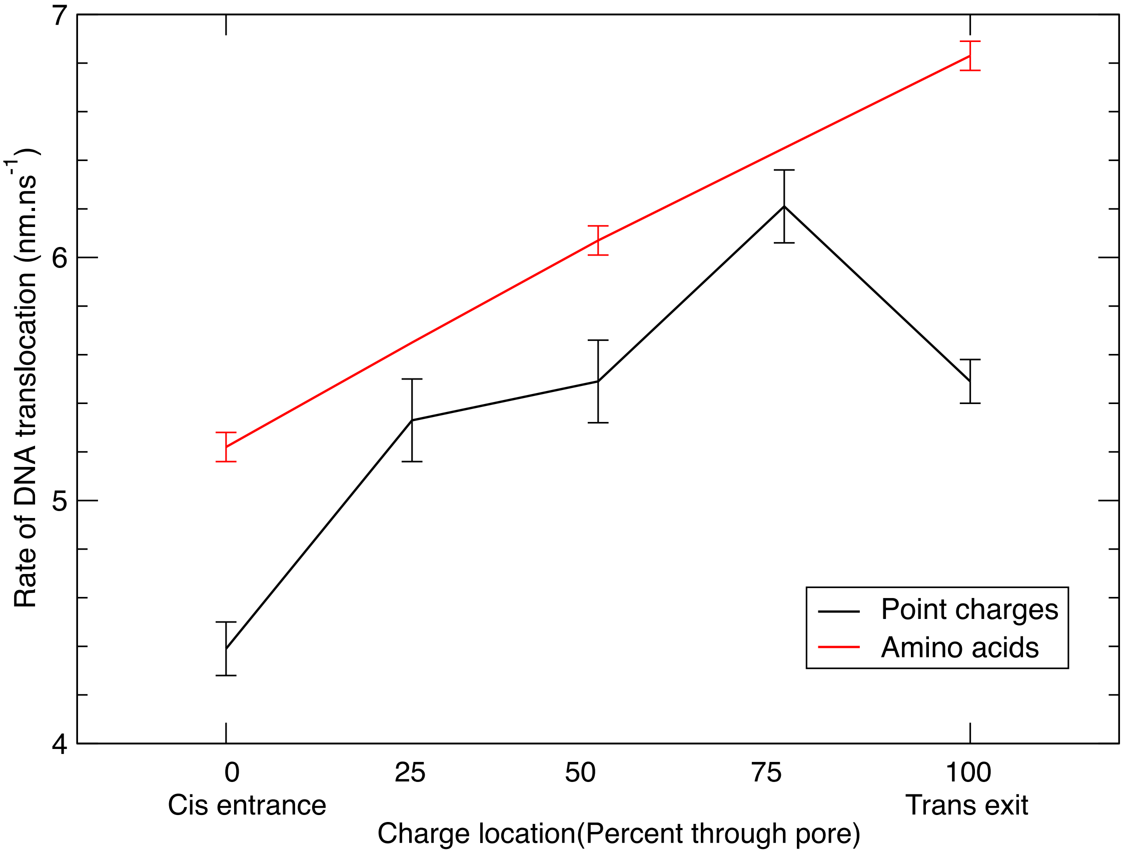


Figure 104: A comparison between the rate of DNA translocation for pores with positive point charges and those with lysine amino acids.

The reduction in error with lysine pores can be attributed to the reduced distance between the 5 charges and the phosphate backbone. The reduced effect of the trans exit is not present with the lysine pores, due to the electric field, the sidechains point towards the centre of the pore, again showing how minor movements of charge can result in values different from expected.

7.3.2.2 Secondary structure

Pore charge and location	Average number of bases in pore for poly(A)$_{60}$
Positive (cis)	28.33±0.04
Positive (cis/middle)	28.29±0.04
Positive (middle)	28.29±0.04
Positive (mid/trans)	28.25±0.04
Positive (trans)	28.29±0.04
Lysine cis	28.25±0.04
Lysine middle	28.24±0.04
Lysine trans	28.21±0.04

Table 5: The degree of secondary structure, compared between positive point charges and positive lysine based pores.

As shown in Table 5, similarly to the point charges, all simulations are approximately within error of each other. Therefore with DNA under tension, shown in Figure 105, it does not appear that the charge distribution has a noticeable impact on secondary structure.

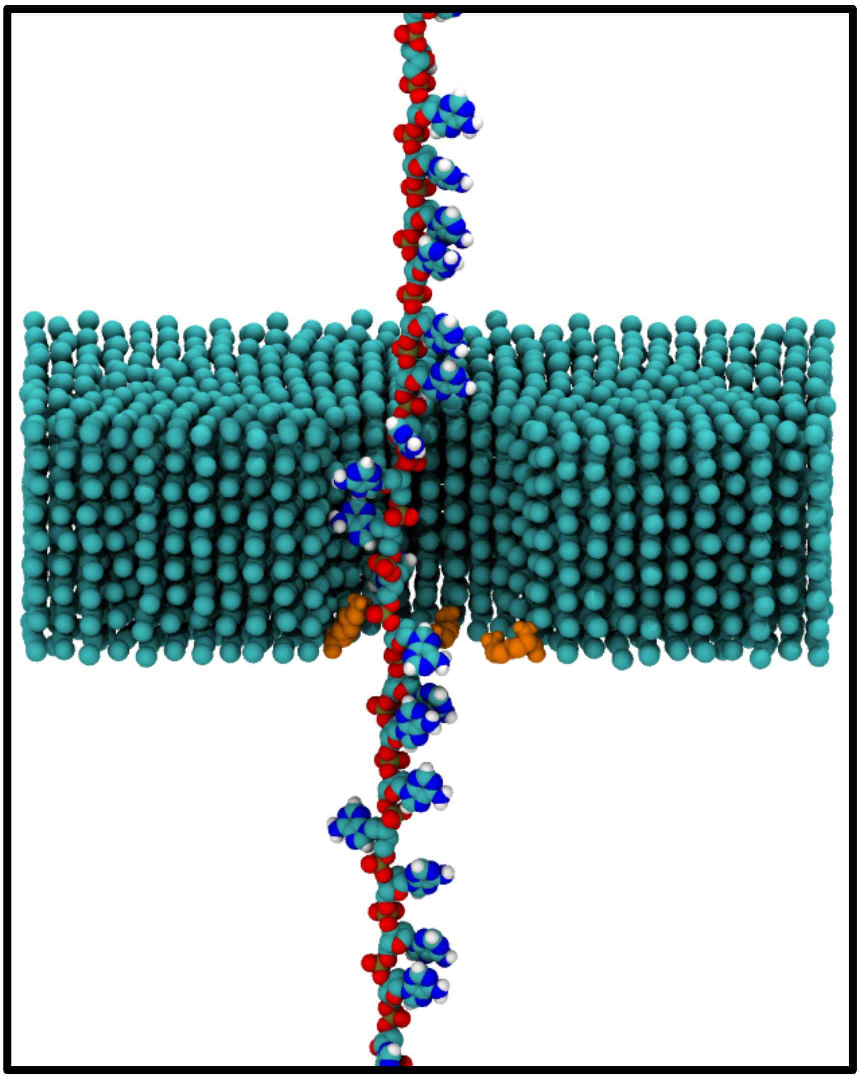


Figure 105: Poly(A)$_{60}$ translocating the pore with Lysine in the trans location, shown in orange, displays minimal secondary structure.

7.4 Discussion

Initial results suggest that the positions of charges can have just as large an impact on the rate of DNA translocation as the sign of the charge. The trans exit of the pore has a reduced impact compared to the remainder of the pore. It is plausible that changing the energetic barrier, prior to exit would alter the rate of translocation as the rate of a reaction is inversely proportional to the activation energy. In addition, the degree of secondary structure in the translocating DNA increases the complexity, as the effect charges have changes as secondary structure is increased. Of course this model is for simple pores, the αHL transmembrane domain has many charges i.e. the wild-type contains E111, K147 at the cis entrance, D127, D128 and K131 at the trans exit [80].

The effect of charge appears to be additive to some degree when secondary structures are not present; the pore with both a negative and positive ring at the locations that individually give the slowest DNA translocation, slows the DNA to a greater degree. This is also true for the pore with negative and positive rings at locations which cause the DNA to translocate fastest, the DNA translocates even faster. This is important as it shows that the effect of charge is cumulative; adding additional charges in different regions to already charged pores will still result in the same effect.

The observed trend for protein based nanopores is similar to those seen with point charges. Although these simulations are not representative of many experimental systems; it is similar to the αHL beta barrel as it is of constant width. Rincon-Restrepo et al [86] found mutation from neutral to a positive charge at the cis entrance of αHL, M113R, had negligible effect, possibly due to the presence of the positively charged K147 residue in the same region, on the opposite strand of the beta barrel. In the same paper they also found addition of multiple positive charges to the pore had a more substantive effect than single point mutations, validating our results that show the effect of charge are cumulative.

7.5 Conclusion

We have demonstrated that negative charges, like positive charges can be useful for reducing both the rate of translocation through nanopores and the degree of secondary structure. We have also shown that location is important when introducing charges into nanopores as both charges can speed up or slow down DNA translocation, with our model it depends entirely on where it is introduced. Therefore, using these results we expect to be able to slow DNA translocation through protein nanopores to a greater extent than previously recorded, allowing for more accurate reading of ionic currents and therefore improved DNA sequencing.

7.6 Future work

Many more repeats could be performed; by varying the DNA sequence used as well as multiple repeats. Performing simulations for the remaining 3 homopolymers and several other sequences, designed to either promote or reduce secondary structure formation will allow further study of the relationship between secondary structures and rate of translocation.

Non-periodic strands of DNA could be used in order to observe how the terminal regions of DNA react to the various charges. As these regions have increased flexibility, the effect conformation can be studied in greater detail. This would also investigate the energetics to pore entry. So far we are unaware of if our pores are sequencing; for example it may be the negatively charges prevent DNA translocation. However, PMFs may not be required as the previously discussed non-periodic DNA should demonstrate the ability, or lack thereof, of the pores to capture DNA strands.

So far we have exclusively used ‘toy’ systems, with either just point charges or individual amino acid residues. Further investigations should look at increasing the realism, potentially starting with a beta barrel model pore of glycines, for validation that this work can apply to a more realistic representation of a protein.

Chapter 8 Conclusions

Over the course of this thesis, several issues relative to nanopore sequencing have been studied. We have shown that exonuclease sequencing using aHL will require modification of the aHL protein to allow a greater percentage of nucleotides to enter the pore. We have revealed the energetics of DNA translocating through aHL and MspA, allowing rational targets for future mutations, whilst also showing that the displacement of water is not a significant energetic barrier. Finally, we demonstrated that when charged mutations are introduced to the nanopore, the location dictates the effect is has to the same degree as the sign of the charge.

We therefore believe this work has allowed for increased rational selection and choice for the design of future nanopore sequencers.

Appendices

Appendix A- Potential of mean force through αHL

Appendix B- Potential of mean force through MspA

Appendix C- Water analysis

Appendix D- Nanopore design

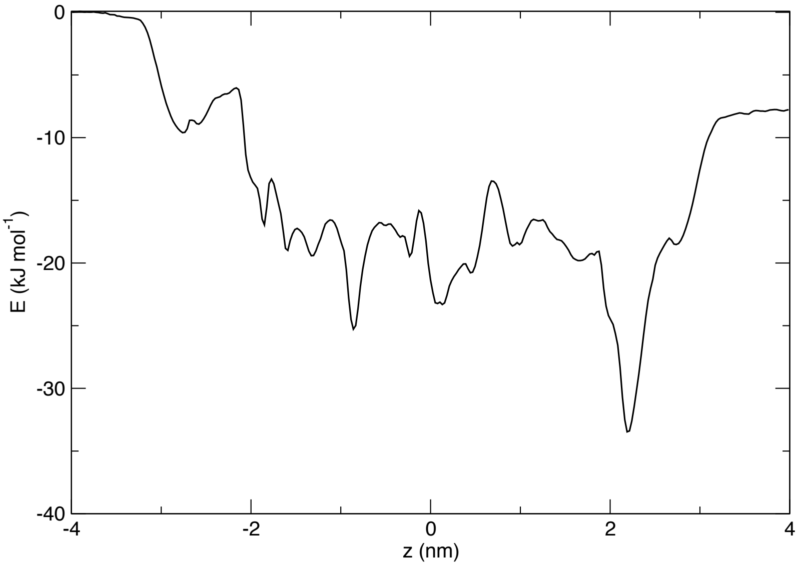
Appendix E- Published work

Appendix A PMFs through αHL additional data

αHL Adenine base

Non-cyclic profiles

The difference in bulk water energies is approximately 7.7 kJ.mol^{-1}.

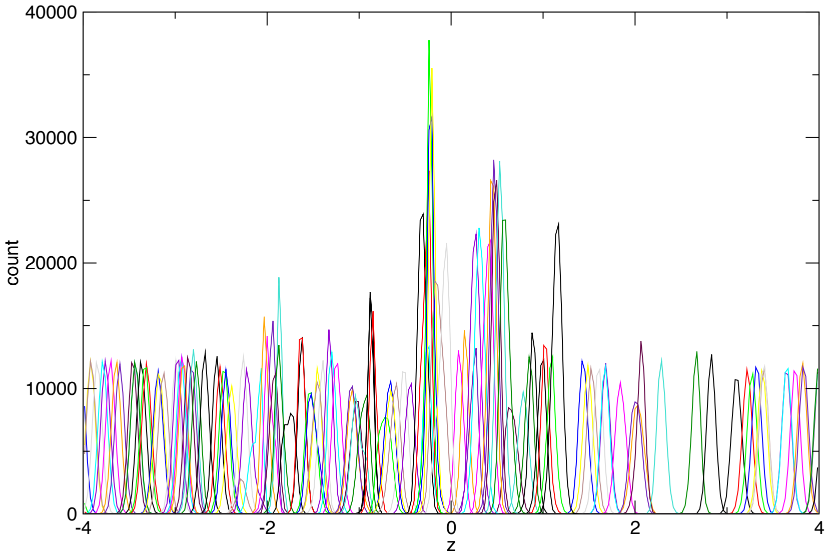


Appendix A

Histograms

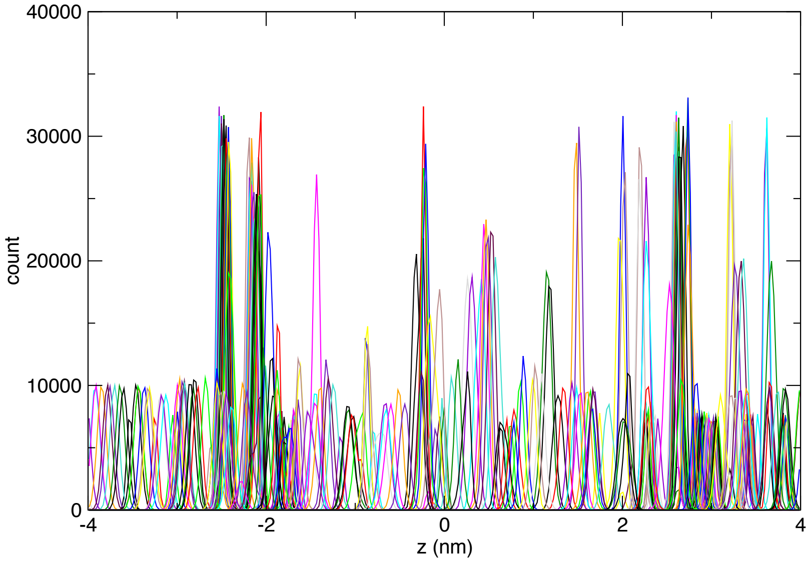
Prior to additional simulations

There are gaps of nearly zero sampling, predominantly from approximately z=1 to z=4.

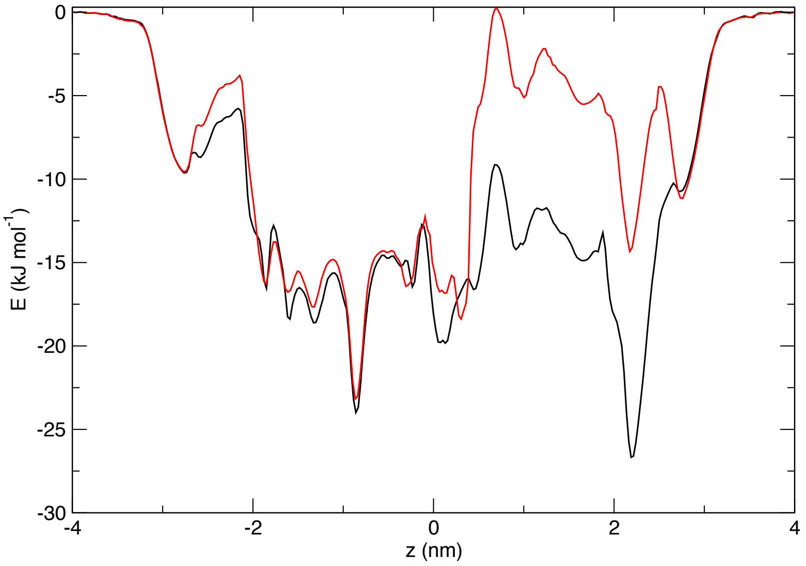


Post additional simulations

The gaps are now filled and as such the average sampling has increased to approximately 1200.



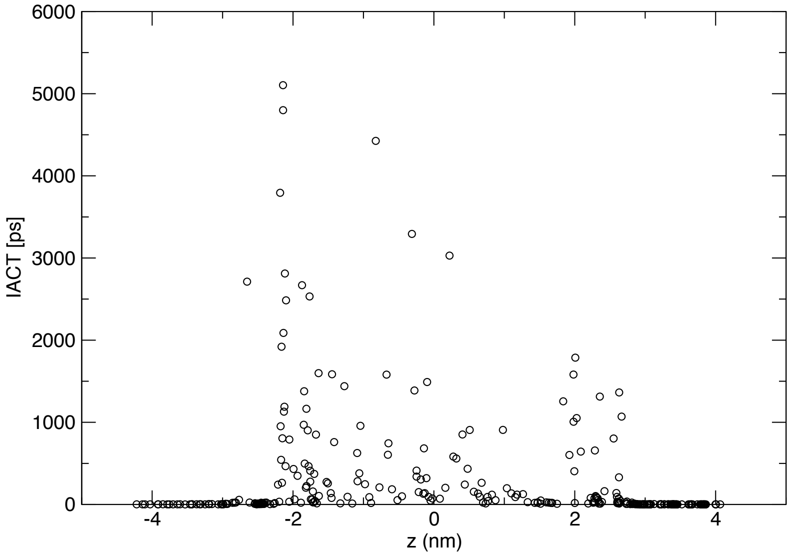
Comparison to Dr Guy's results



My final PMF in black, Dr Guy's in red. Notable changes to the profile are at z=-2.5, z=0, z=0.5 and z=2.5.

Autocorrelation times

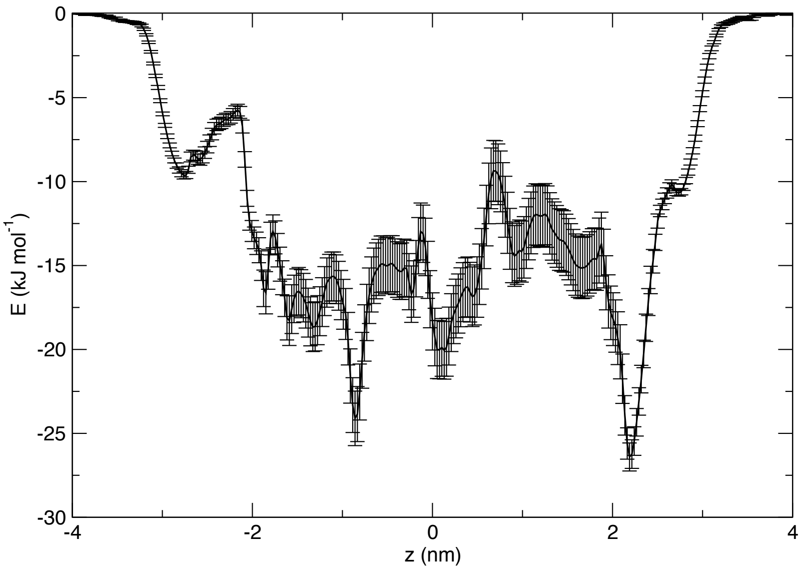
The highest autocorrelation time is below 5.5 ns, with only 1 frame where it is above 5 ns.



Appendix A

Bootstrap Analysis

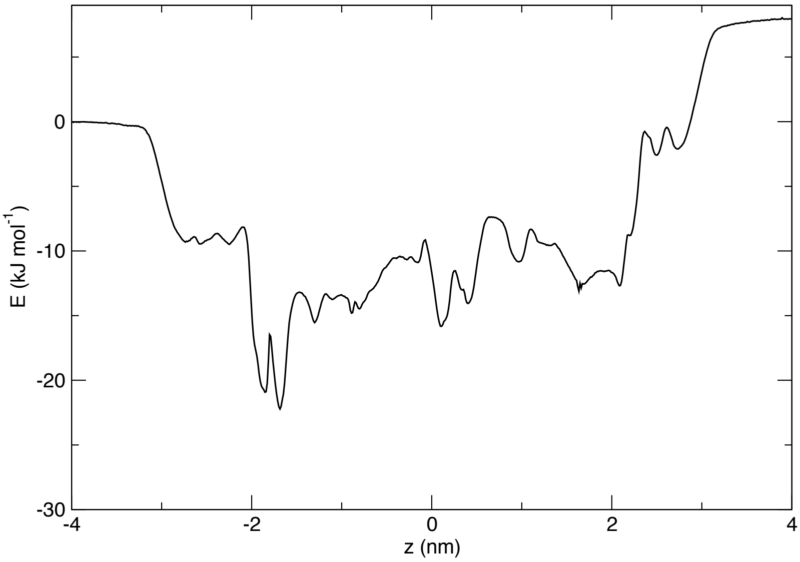
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 1.8 kJ.mol^{-1} in the protein



αHL Cytosine base

Non-cyclic profiles

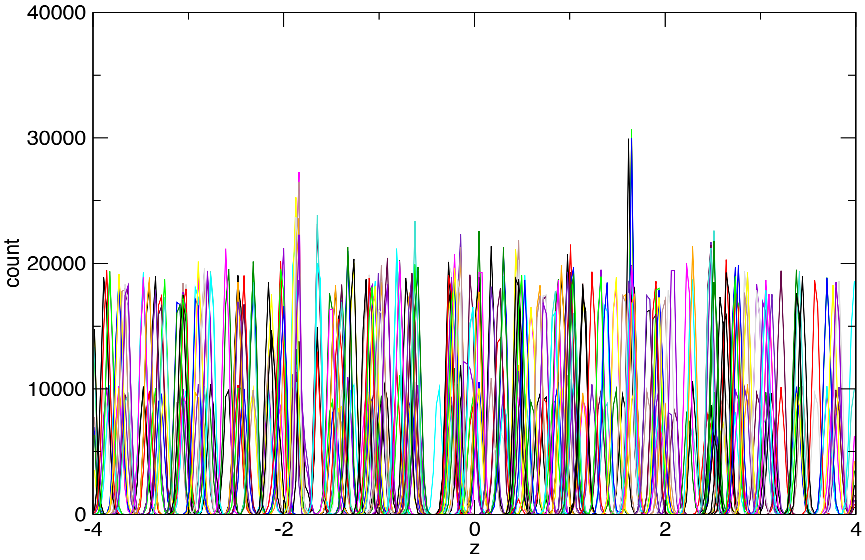
The difference in bulk water energies is approximately 8 kJ.mol^{-1}.



Histograms

Prior to additional simulations

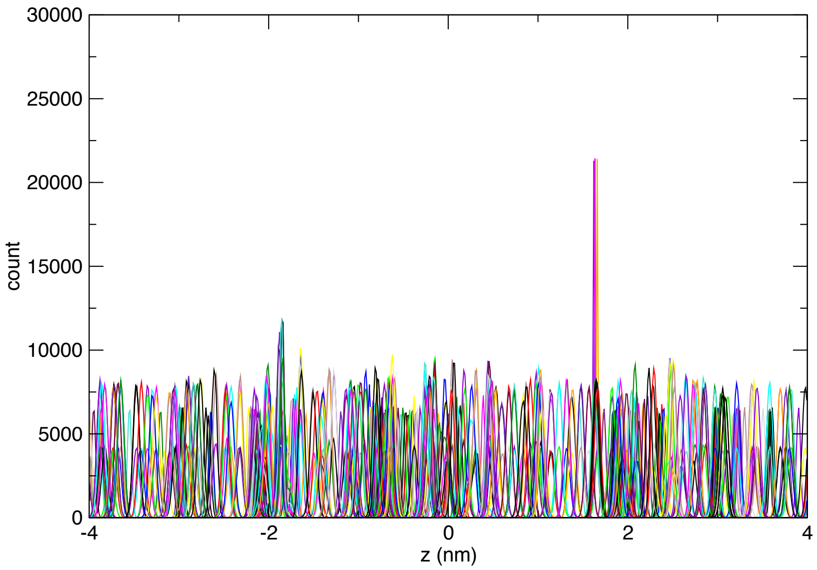
2 major gaps of limited sampling, at approximately z=-1.8 and z= -0.3, minor gaps of sampling less than 2500 in 15 other regions, otherwise average sampling of approximately 10000.



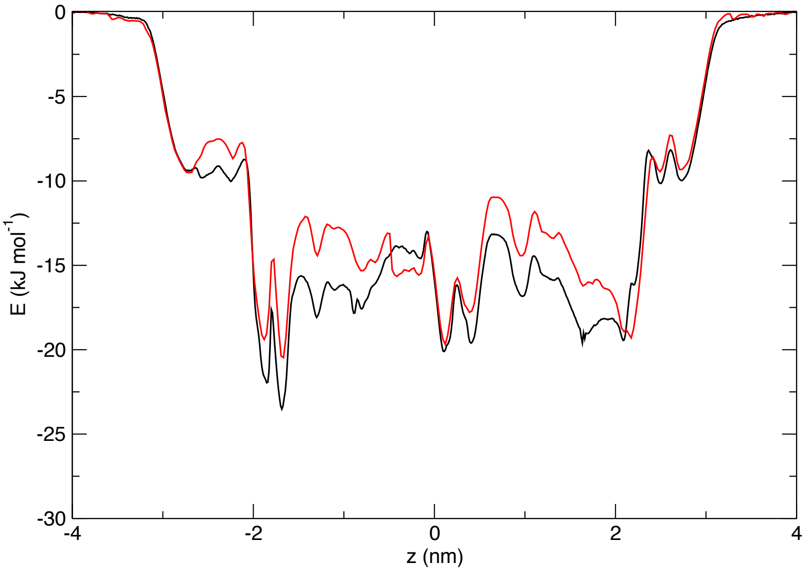
Appendix A

Post additional simulations

All the major and minor gaps are now filled.



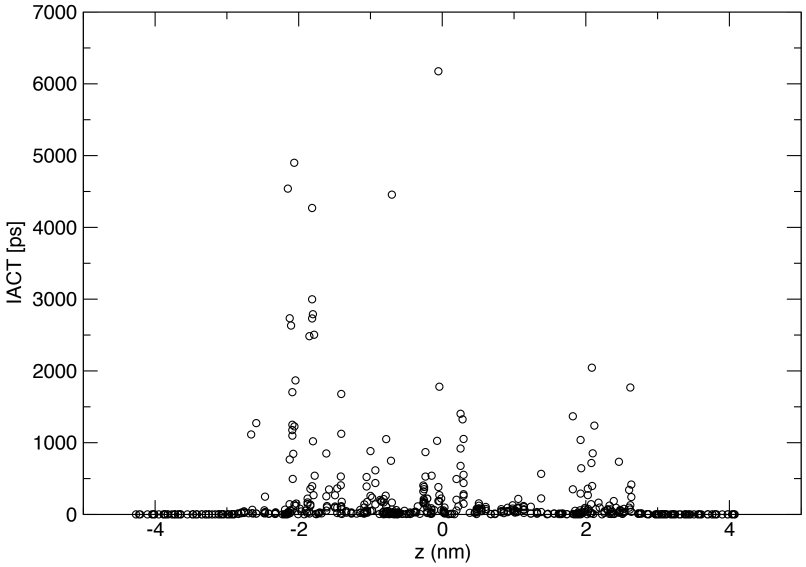
Comparison to Dr Guy's results



My final PMF in black, Dr Guy's in red. Notable changes to the profile are at z=-2.5, z=-1.5, z=0.2 and z=2.

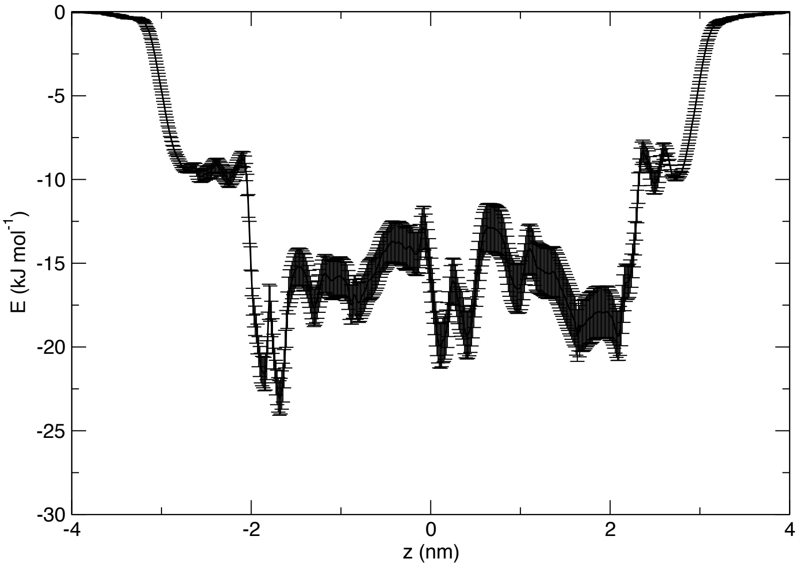
Autocorrelation times

The highest autocorrelation time is below 6.5 ns, with only 1 frame where it is above 5 ns.



Bootstrap Analysis

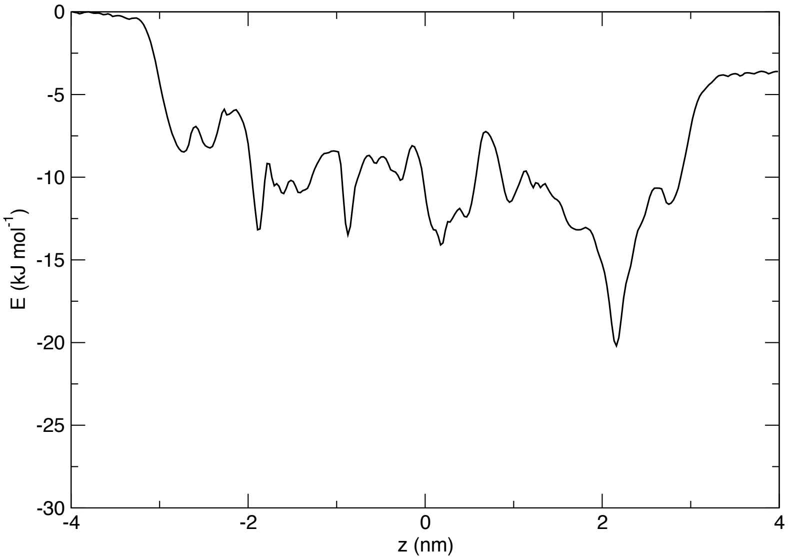
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 1.2 kJ.mol^{-1} in the protein.



αHL Guanine base

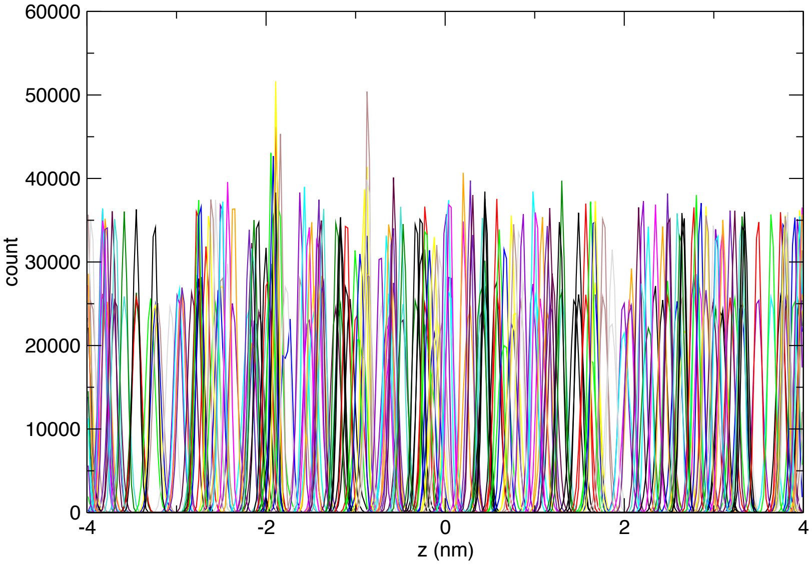
Non-cyclic profile

The difference in bulk water energies is approximately 3.6 kJ.mol^{-1}.



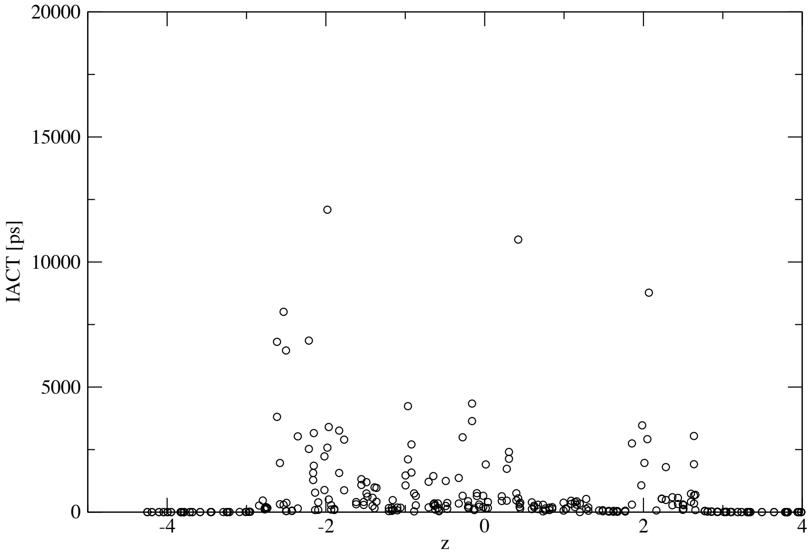
Histograms

Minimum sampling at any region is 10,000 counts, average is in excess of 20,000.



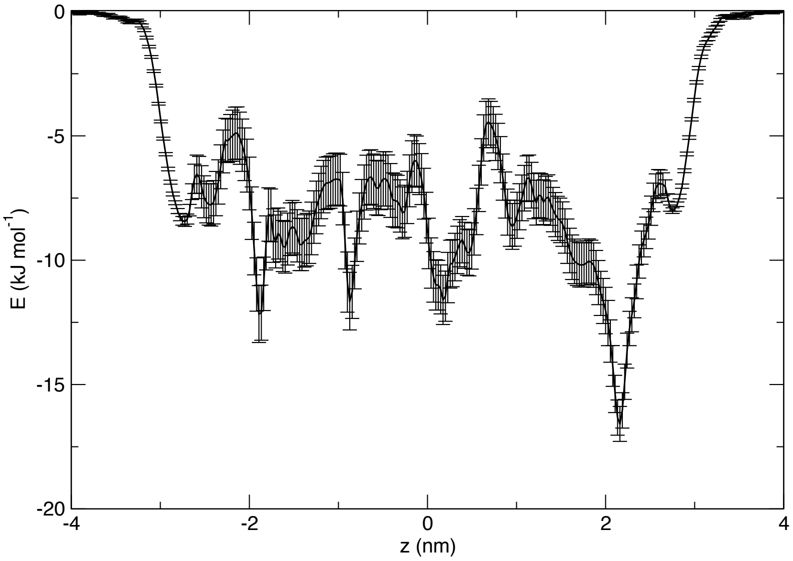
Autocorrelation times

The highest autocorrelation time is below 12.5 ns, with only 7 frames where it is above 5 ns, as the simulation length is 250 ns these have all been covered at least 20 times. Largest autocorrelation times are near the pore entrance and exits, where the fragment has an interaction with the protein but is free to sample the bulk water.



Bootstrap Analysis

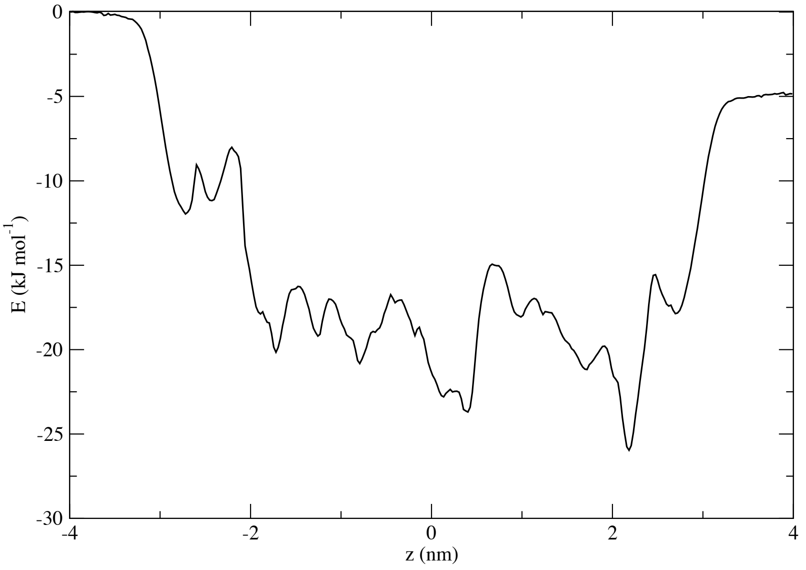
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 1.05 kJ.mol^{-1} in the protein



αHL Thymine base

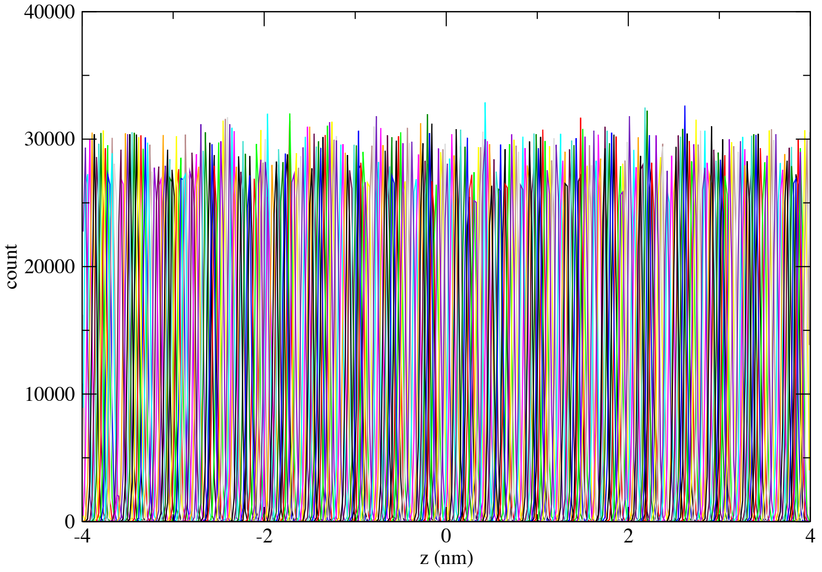
Non-cyclic profiles

The difference in bulk water energies is approximately 5 kJ.mol^{-1}.



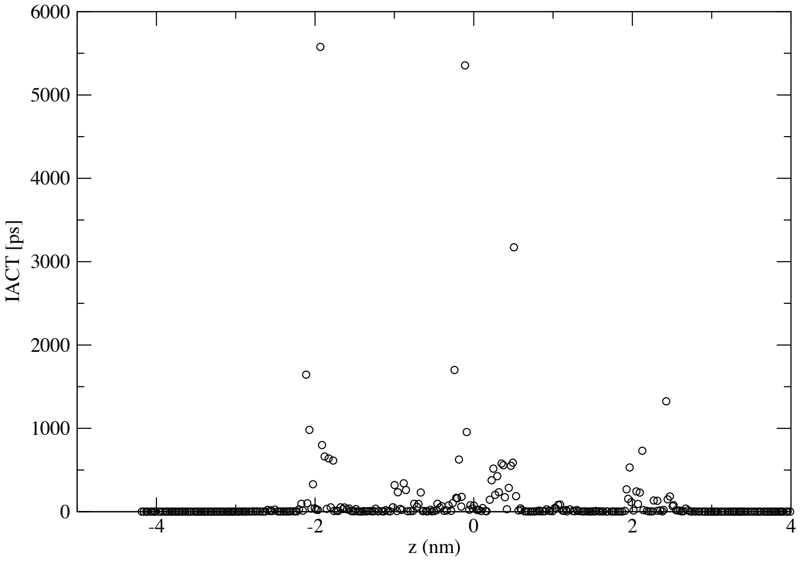
Histograms

The histograms demonstrate extensive coverage of the principle axis, associated with a change in the method for setup of the PMFs as this was the last setup; double density of windows, 0.25 angstroms between each with a 2000 kJ.mol^{-1}.nm^{-2} restraining force.



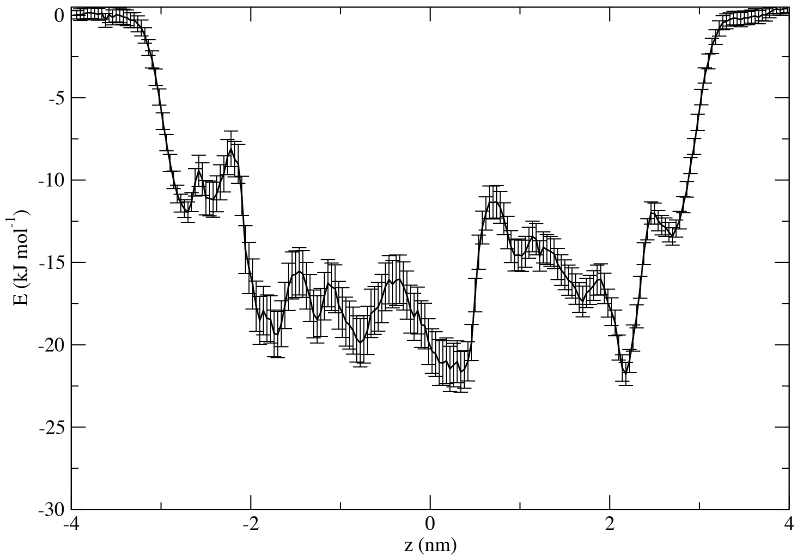
Autocorrelation times

The highest autocorrelation time is below 6 ns, with only 2 frames where it is above 5 ns.



Bootstrap Analysis

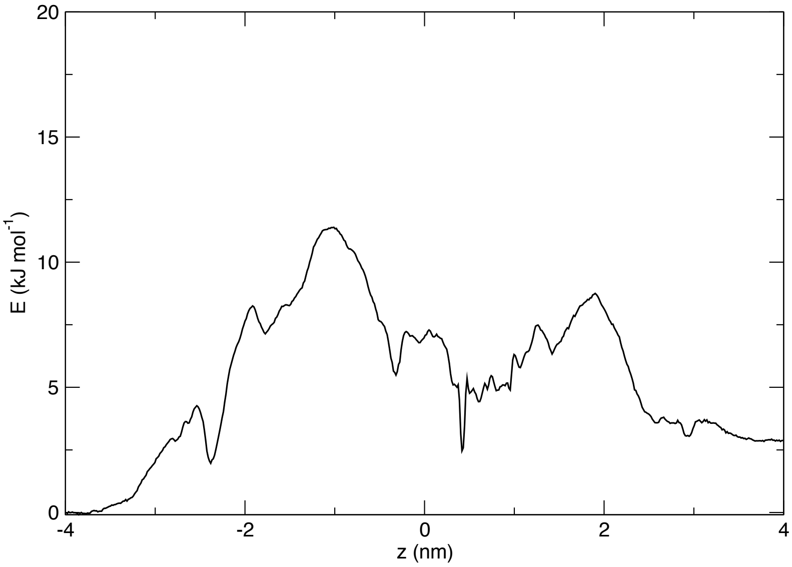
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 1.5 kJ.mol^{-1} in the protein.



αHL Phosphate

Non-cyclic profile

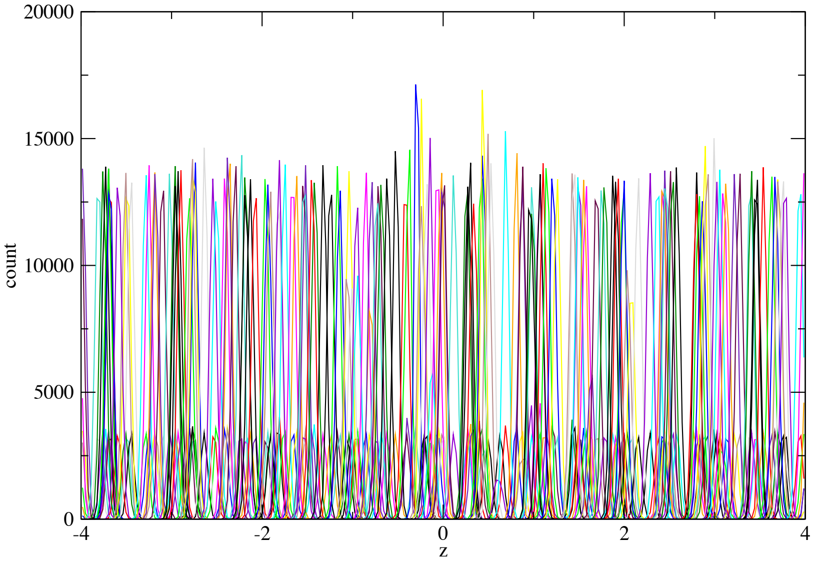
The difference in bulk water energies is approximately 2.6 kJ.mol^{-1}.



Histograms

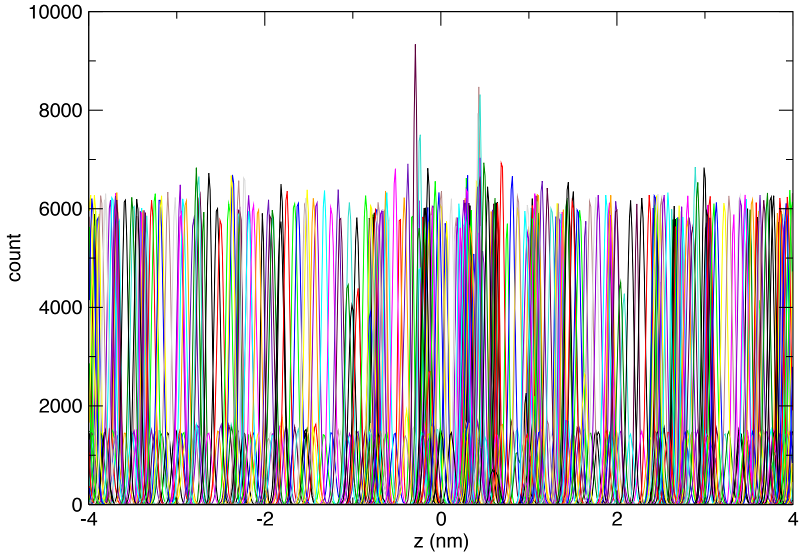
Prior to additional simulations

Nearly zero sampling at z=2.7, with 6 regions of sampling less than 2500 counts and an average sampling 5000 counts.

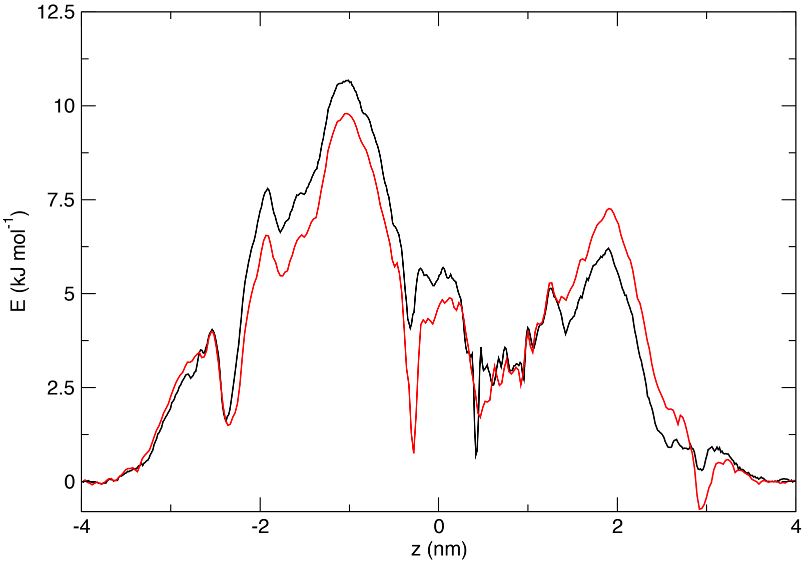


Post additional simulations

The gaps in the histograms are now filled.



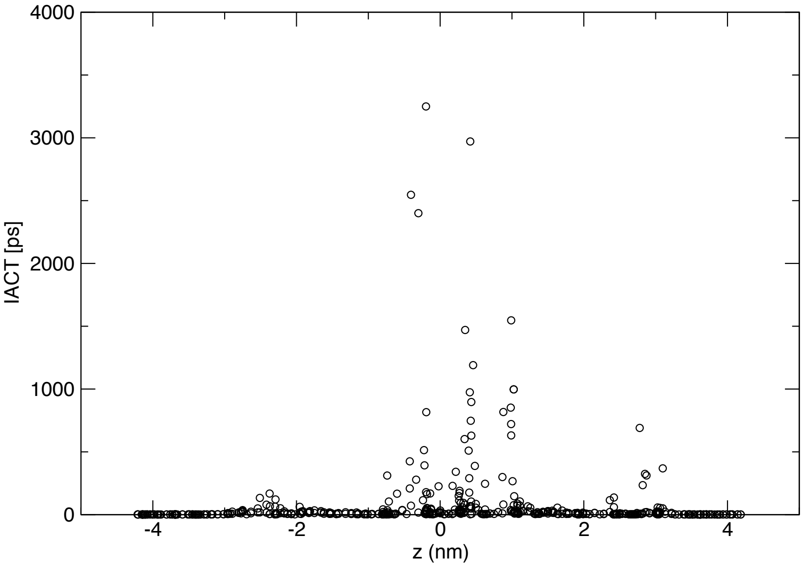
Comparison to Dr Guy's results



My final PMF in black, Dr Guy's in red. Notable changes to the profile are at z=-2, z=-0.25, z=0.5, z=1.2 and z=3.

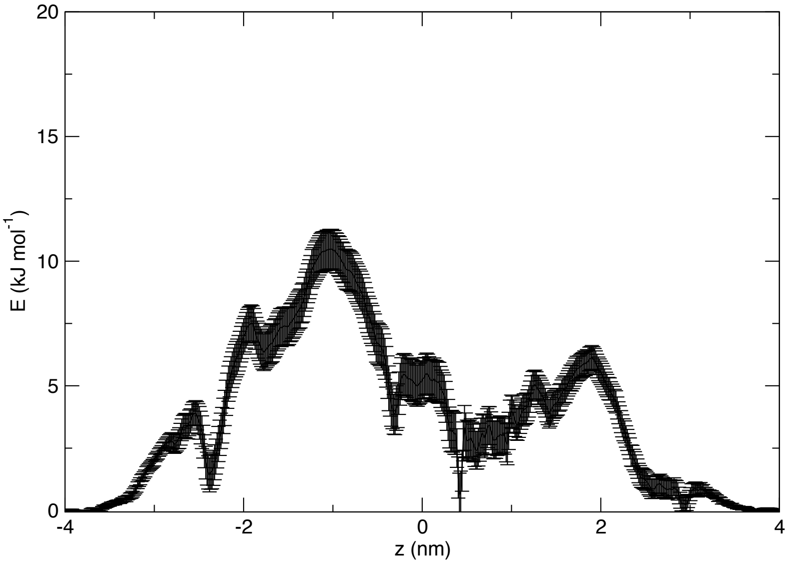
Autocorrelation times

The highest autocorrelation time is below 3.5 ns, with only 5 frames where it is above 1.5 ns. This is significantly lower than the other fragments, likely due to the reduced size and high degree of symmetry in comparison to the other fragments.



Bootstrap Analysis

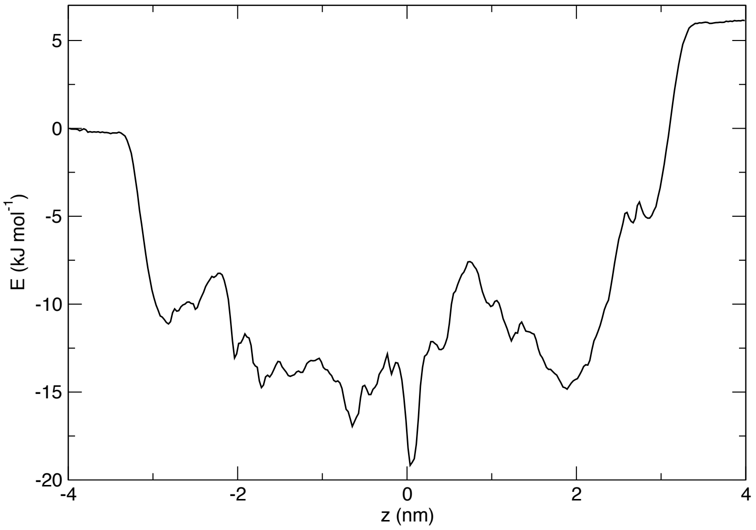
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 0.83 kJ.mol^{-1} in the protein



αHL Cytosine mononucleotide

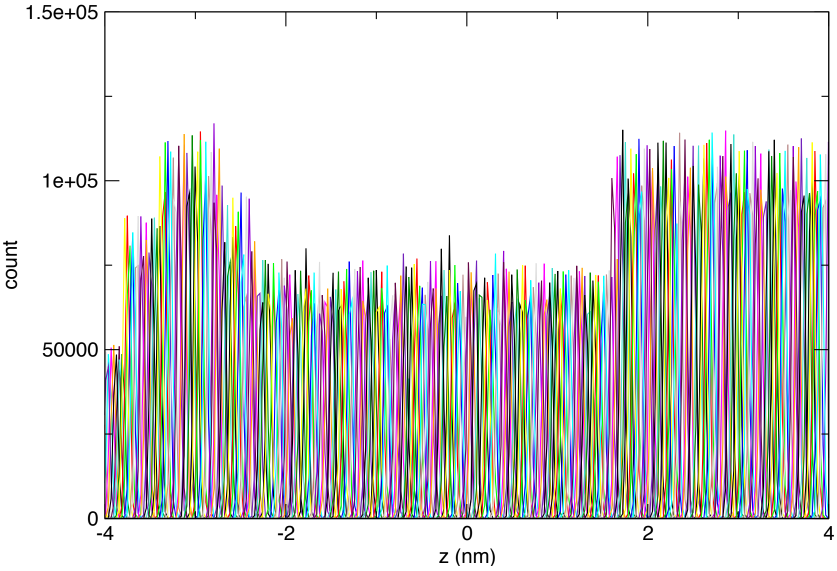
Non-cyclic profile

The difference in bulk water energies is approximately 6.5 kJ.mol^{-1}.



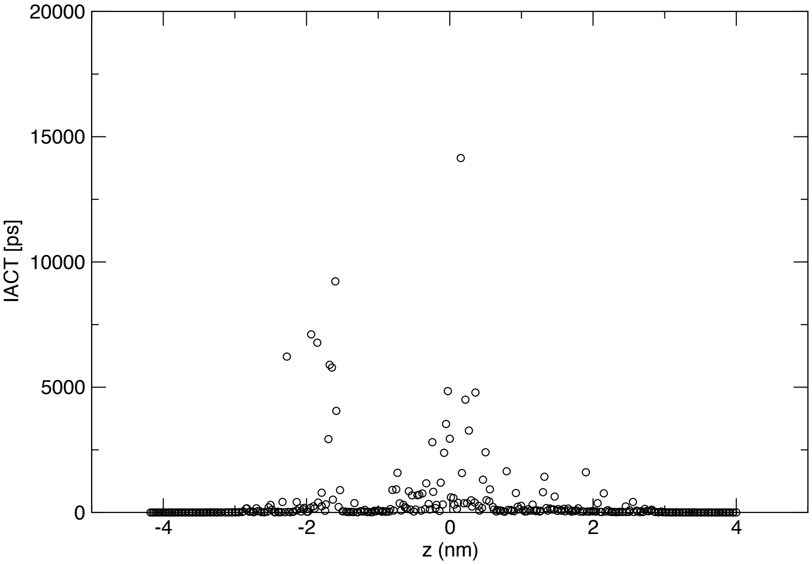
Histograms

The histograms demonstrate extensive coverage of the principle axis.



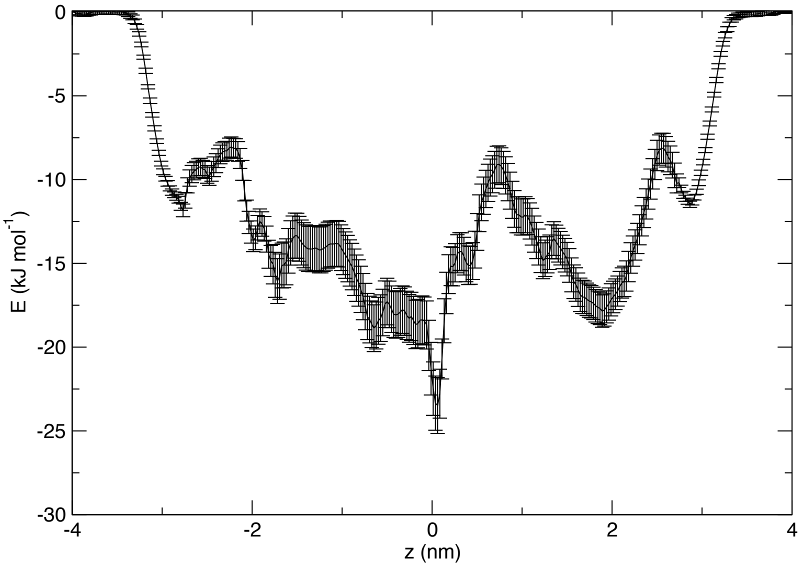
Autocorrelation times

The highest autocorrelation time is below 15 ns, with only 8 frames where it is above 5 ns and 1 above 10. These are larger than for the base as expected, but only by a factor of 2.5.



Bootstrap Analysis

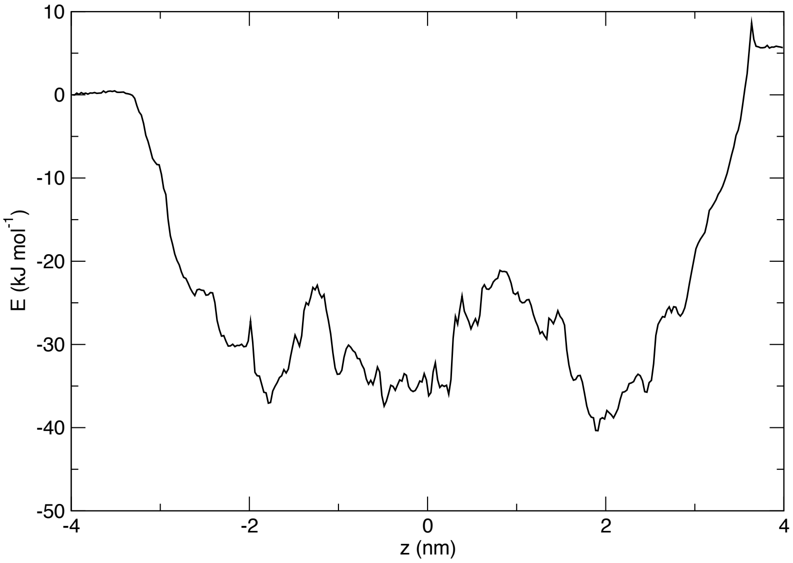
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 1.6 kJ.mol^{-1} in the protein



αHL Guanine mononucleotide

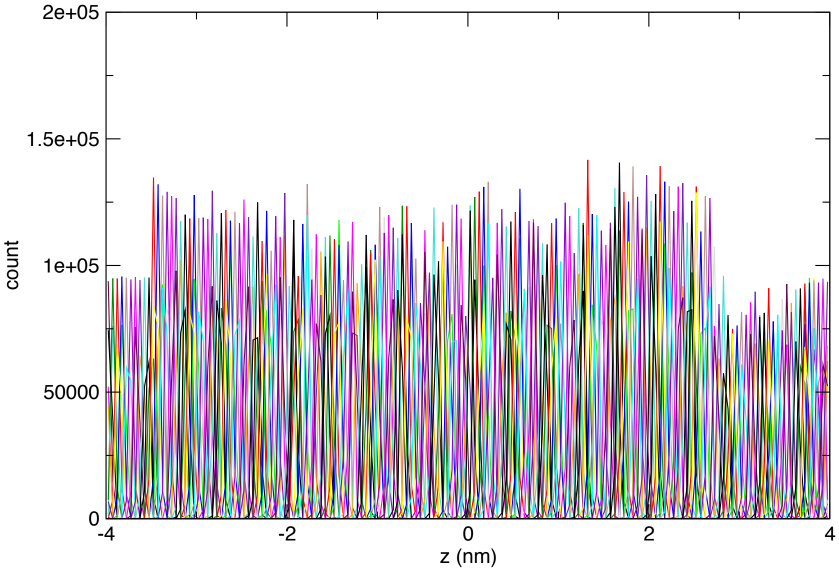
Non-cyclic profile

The difference in bulk water energies is approximately 5.5 kJ.mol^{-1}.



Histograms

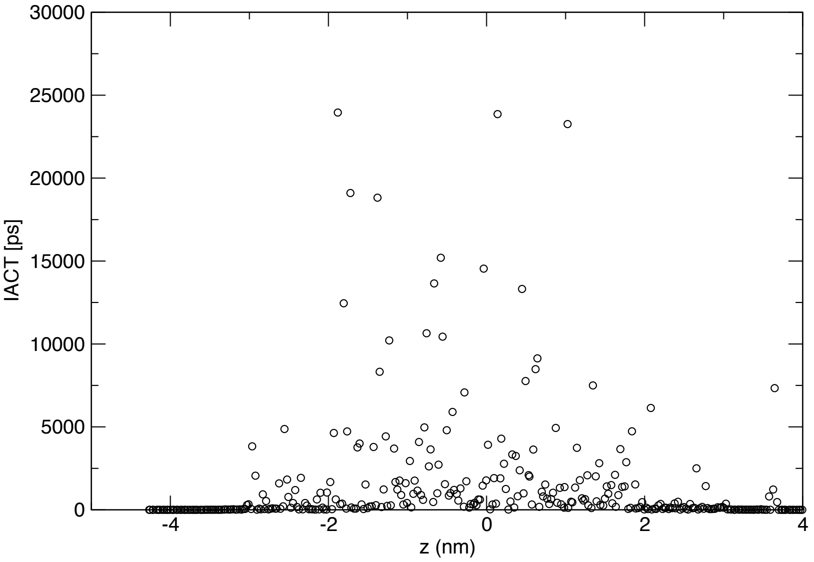
The histograms demonstrate extensive coverage of the principle axis.



Appendix A

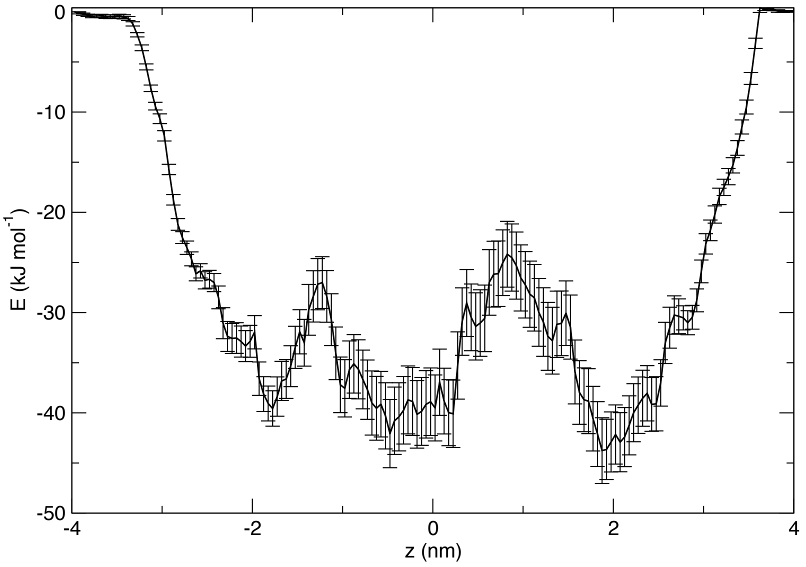
Autocorrelation times

The highest autocorrelation time is below 25 ns, with 13 frames where it is above 10 ns and 26 above 5. These are larger than for the base as expected, but only by a factor of 2, which is surprising as the sampling is much reduced comparatively.



Bootstrap Analysis

The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 3.4 kJ.mol^{-1} in the protein

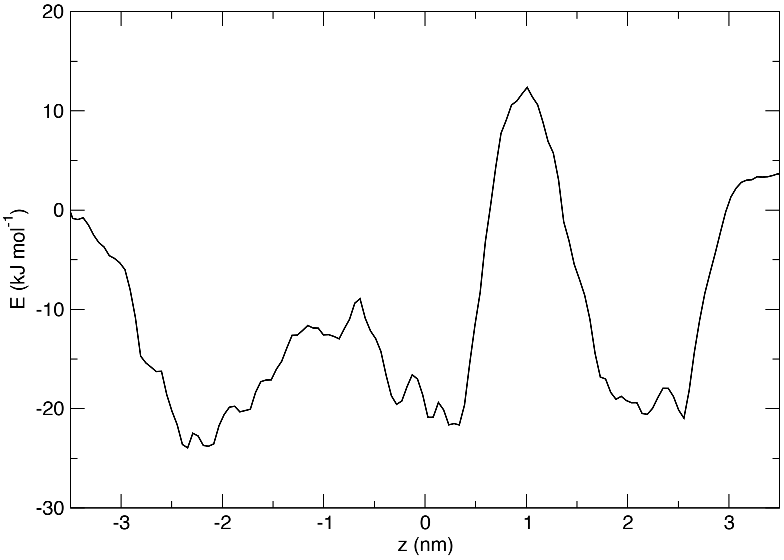


Appendix B PMFs through MspA additional data

A96R Adenine

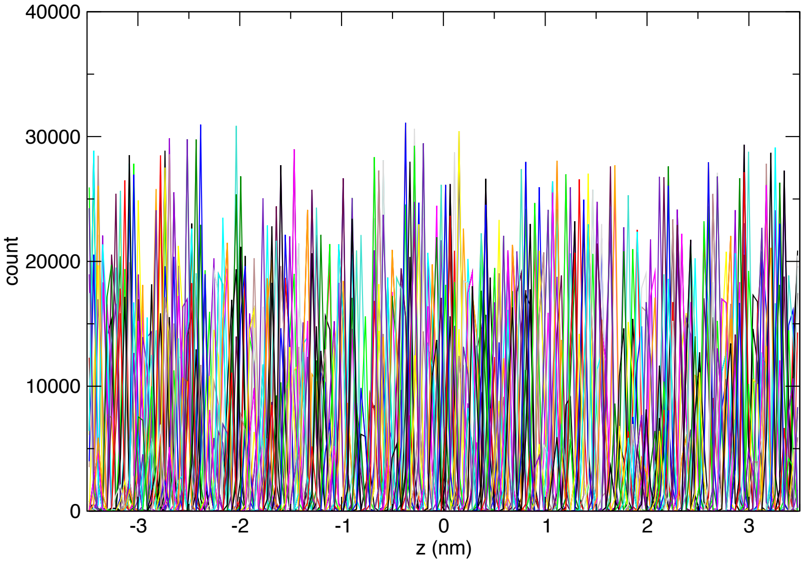
Non-cyclic profile

The difference in bulk water energies is approximately 3.7 kJ.mol^{-1}.



Histograms

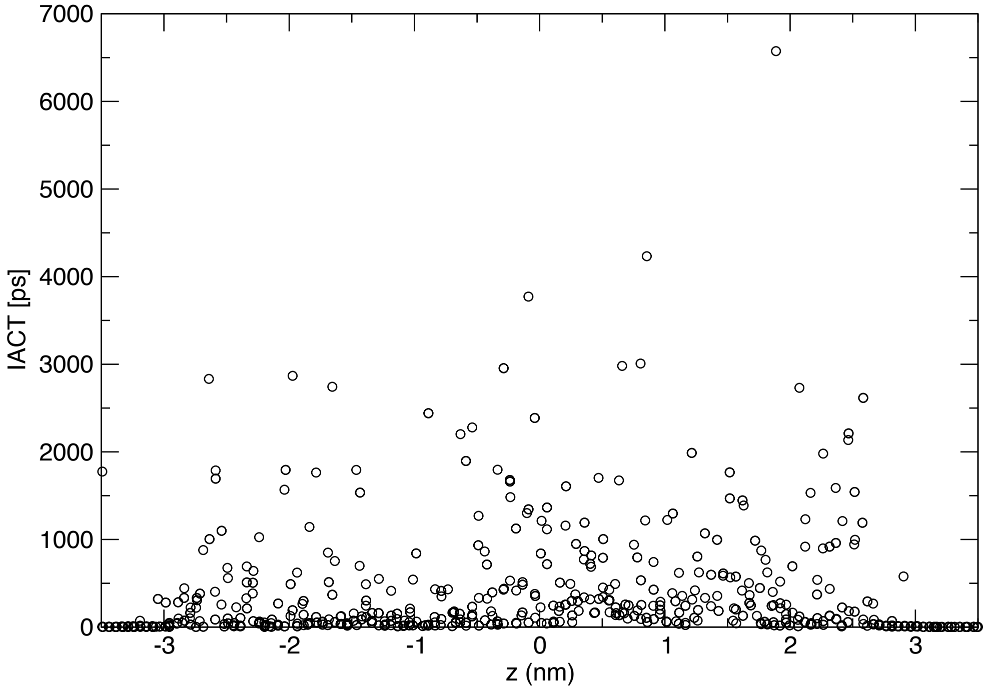
The histograms demonstrate extensive coverage of the principle axis.



Appendix E

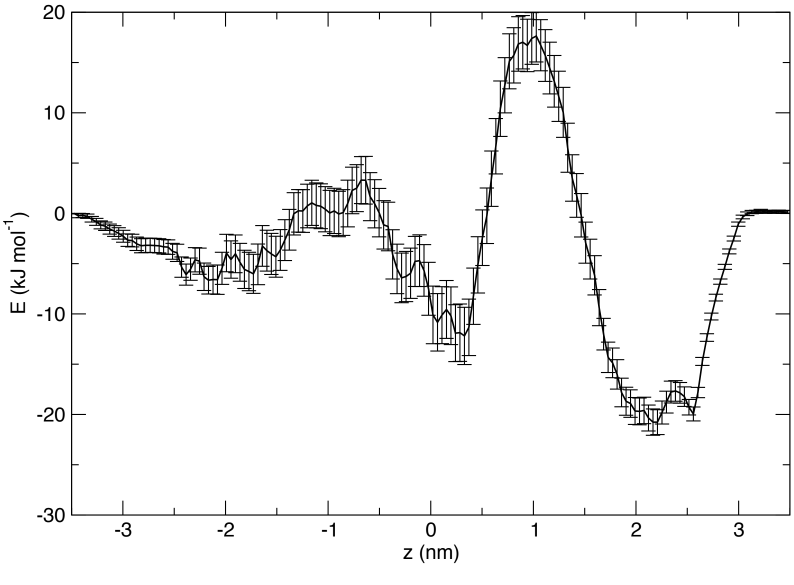
Autocorrelation times

The highest autocorrelation time is below 7 ns, with only 1 frame where it is above 5 ns, as the simulation length is 75 ns these have all been covered at least 10 times.



Bootstrap Analysis

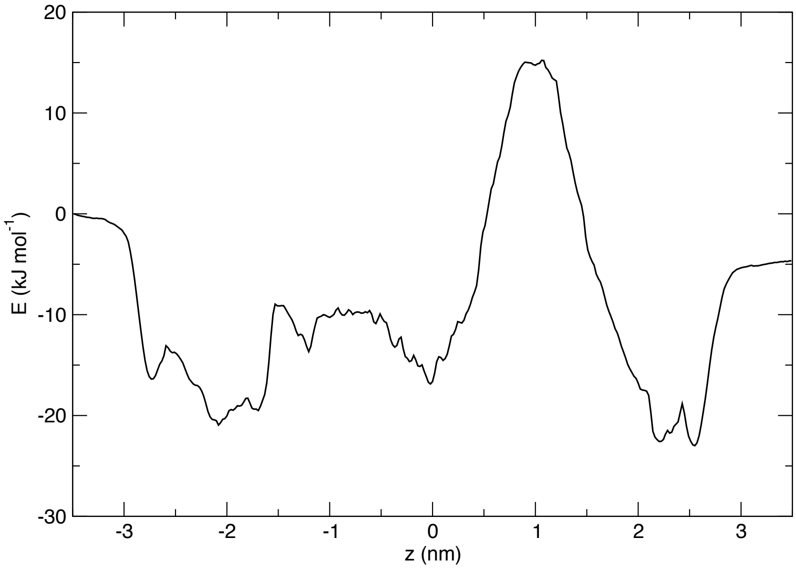
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 2.25 kJ.mol^{-1} in the protein



A96R Cytosine

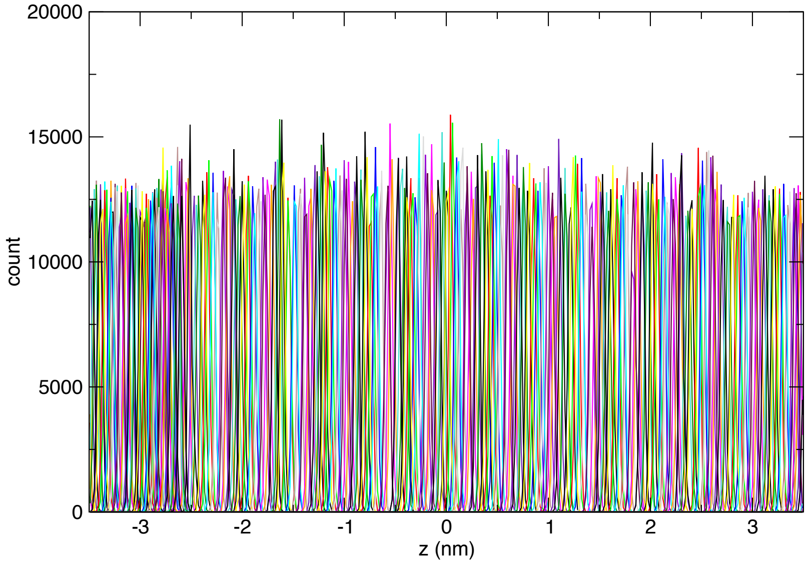
Non-cyclic profile

The difference in bulk water energies is approximately 5 kJ.mol^{-1}.



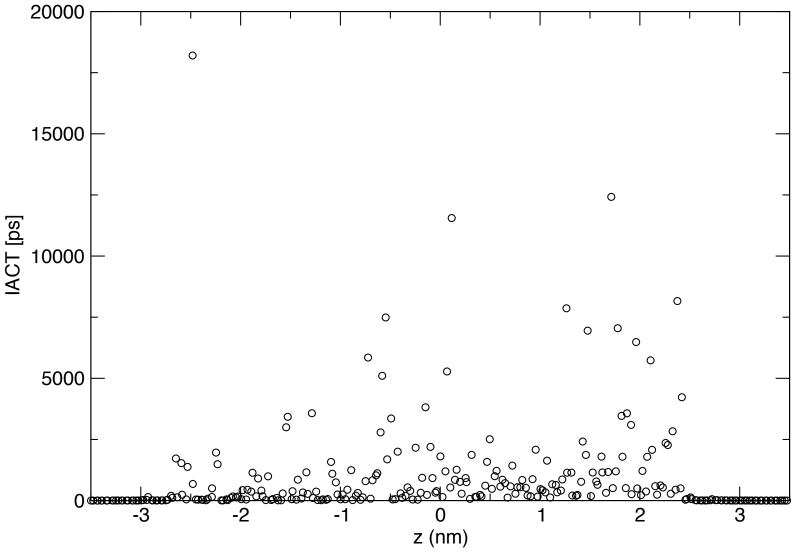
Histograms

The histograms demonstrate extensive coverage of the principle axis.



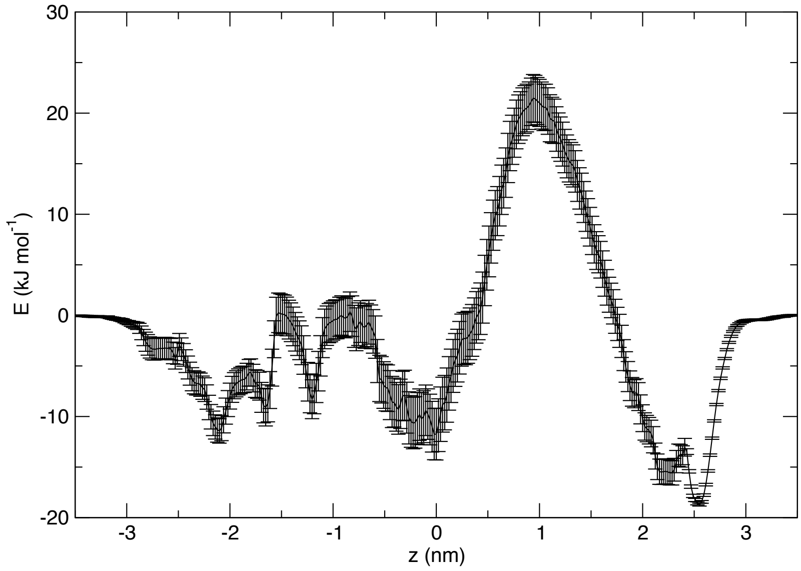
Autocorrelation times

The highest autocorrelation time is below 20 ns, with only 3 frames where it is above 10 ns, as the simulation length is 250 ns these have all been covered at least 35 times.



Bootstrap Analysis

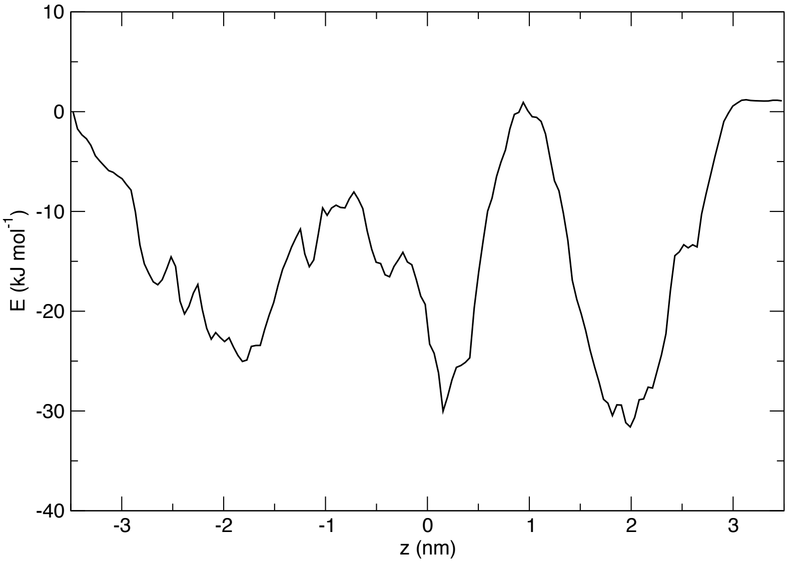
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 2.7 kJ.mol^{-1} in the protein



A96R Guanine

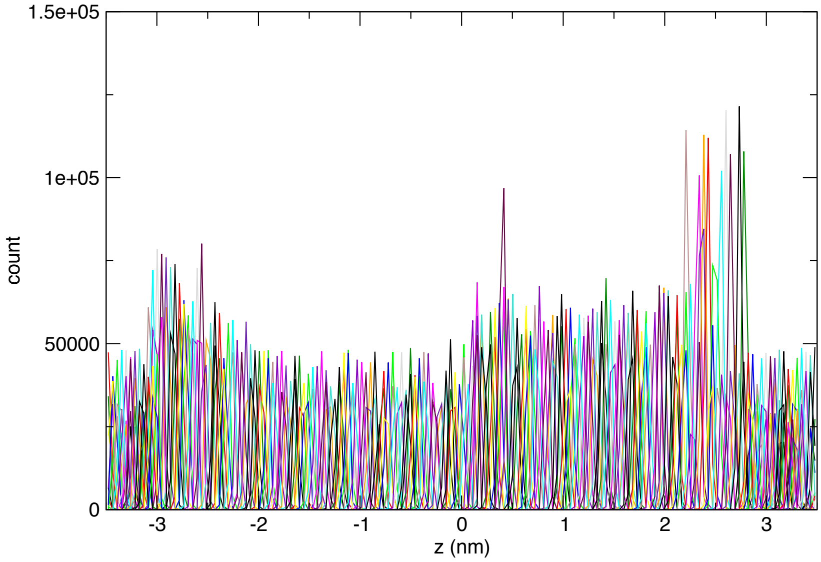
Non-cyclic profile

The difference in bulk water energies is approximately 1.2 kJ.mol^{-1}.



Histograms

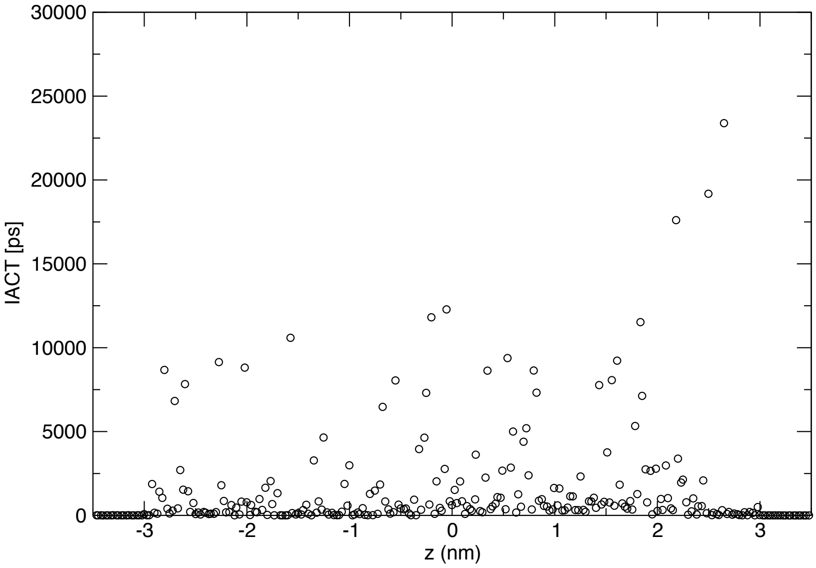
The histograms demonstrate extensive coverage of the principle axis.



Appendix E

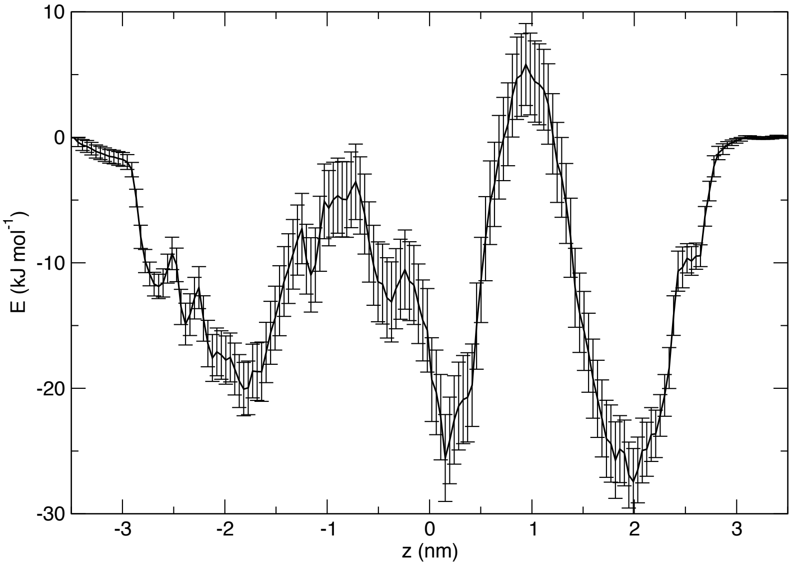
Autocorrelation times

The highest autocorrelation time is below 6.5 ns, with only 1 frame where it is above 5 ns, as the simulation length is 75 ns these have all been covered at least 35 times.



Bootstrap Analysis

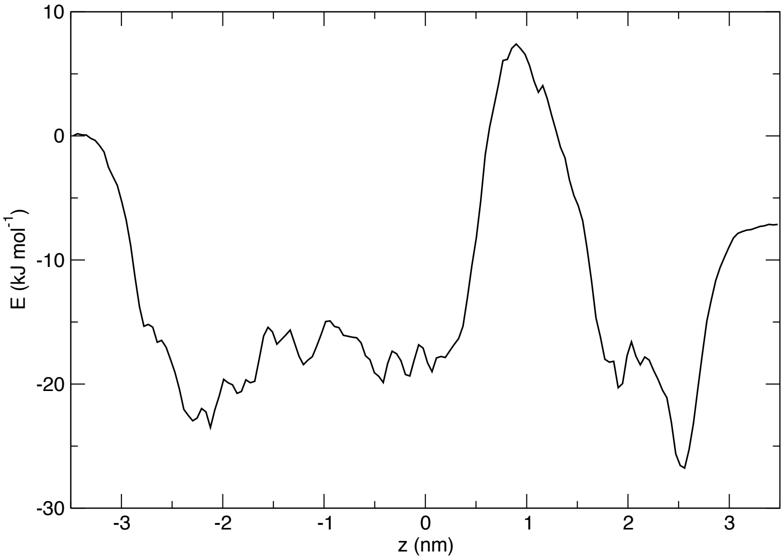
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 3.3 kJ.mol^{-1} in the protein



A96R Thymine

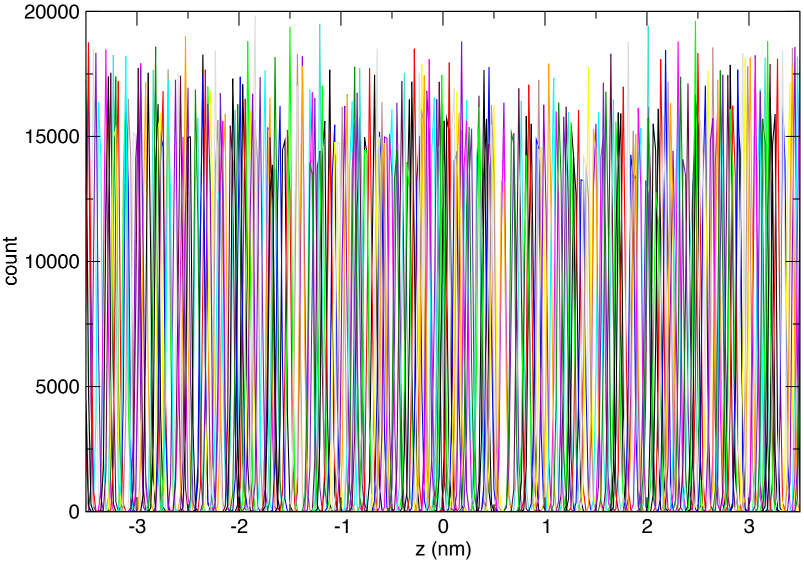
Non-cyclic profile

The difference in bulk water energies is approximately 8.5 kJ.mol^{-1}.



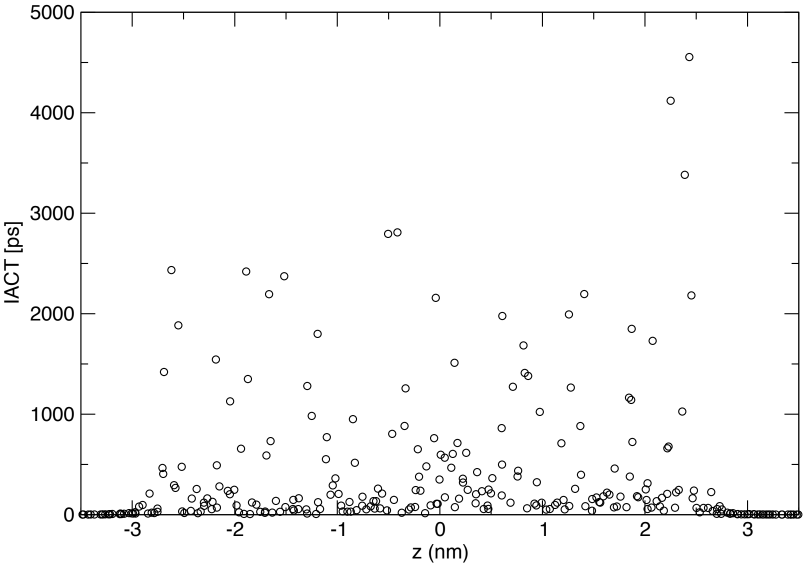
Histograms

The histograms demonstrate extensive coverage of the principle axis.



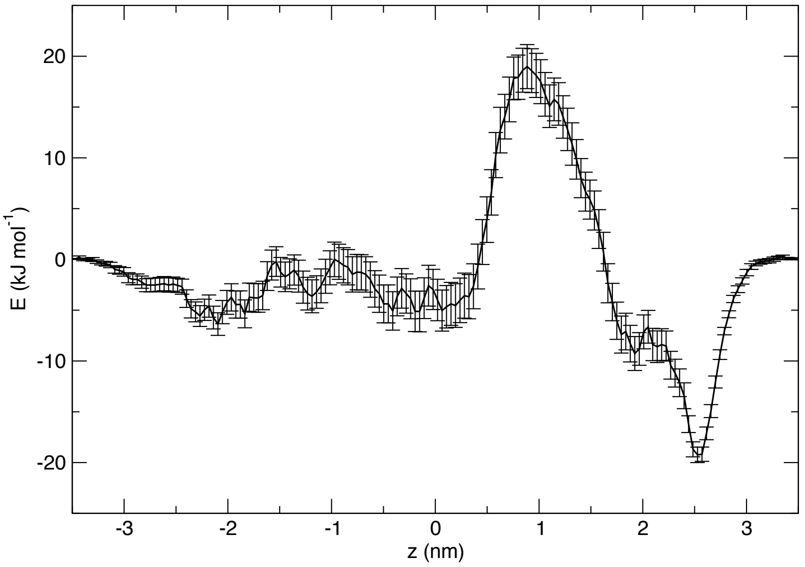
Autocorrelation times

The highest autocorrelation time is below 5 ns, as the simulation length is 75 ns these have all been covered at least 15 times.



Bootstrap Analysis

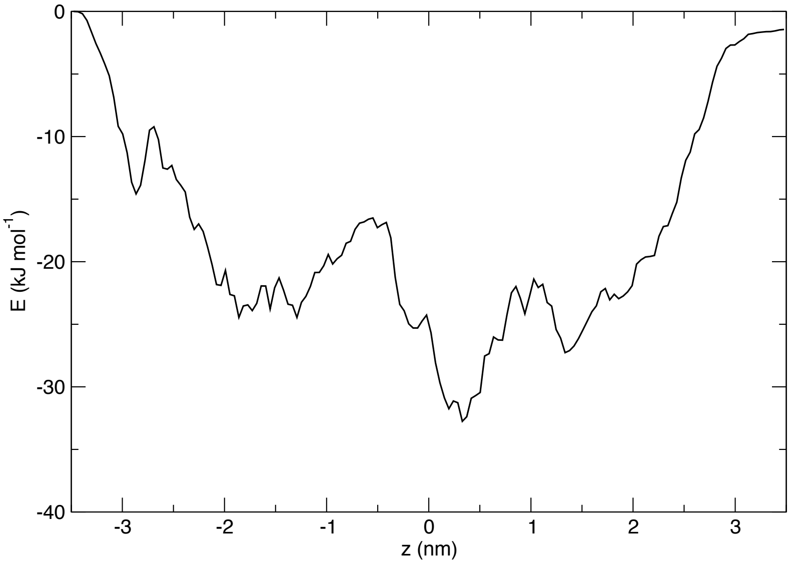
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 2 kJ.mol^{-1} in the protein



M2 Adenine

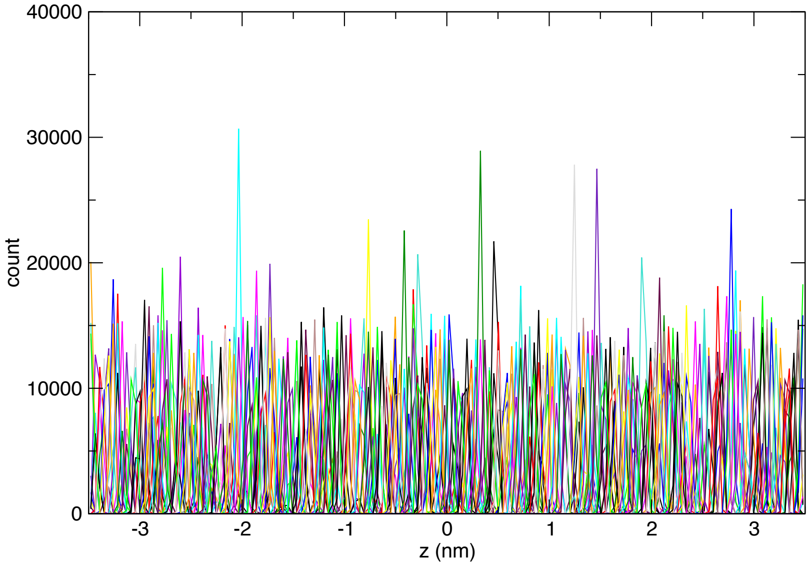
Non-cyclic profile

The difference in bulk water energies is approximately 1.4 kJ.mol^{-1}.



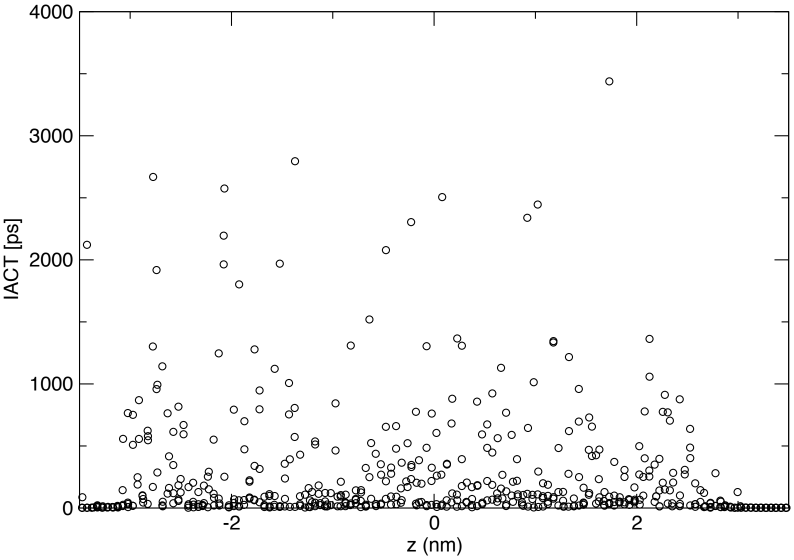
Histograms

The histograms demonstrate extensive coverage of the principle axis.



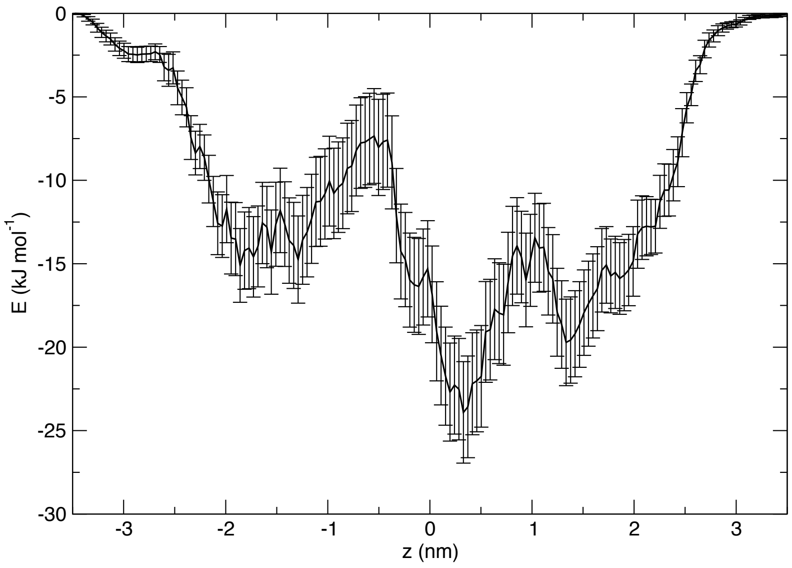
Autocorrelation times

The highest autocorrelation time is below 5 ns, as the simulation length is 75 ns these have all been covered at least 15 times.



Bootstrap Analysis

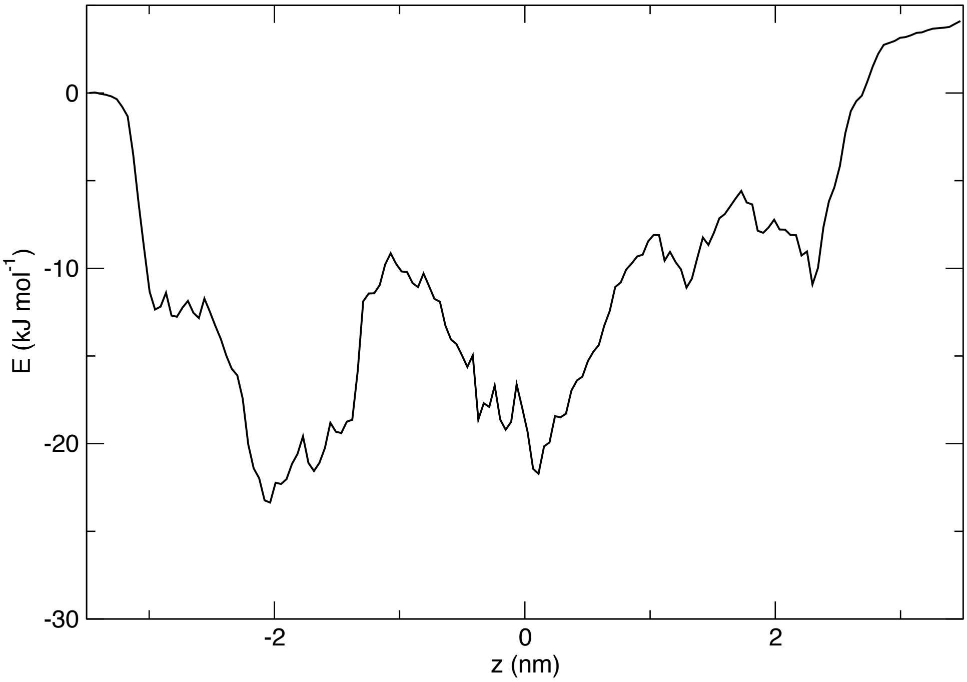
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 3.1 kJ.mol^{-1} in the protein



M2 Cytosine

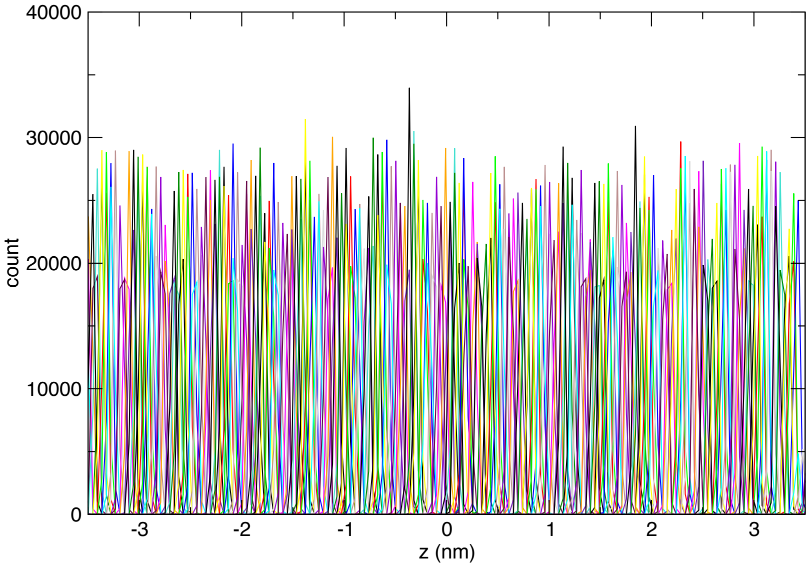
Non-cyclic profile

The difference in bulk water energies is approximately 3.5 kJ.mol^{-1}.



Histograms

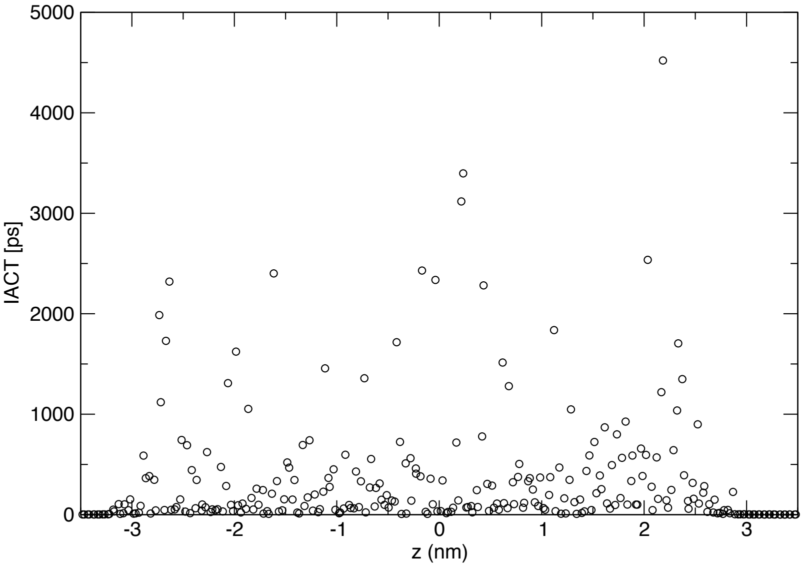
The histograms demonstrate extensive coverage of the principle axis.



Appendix E

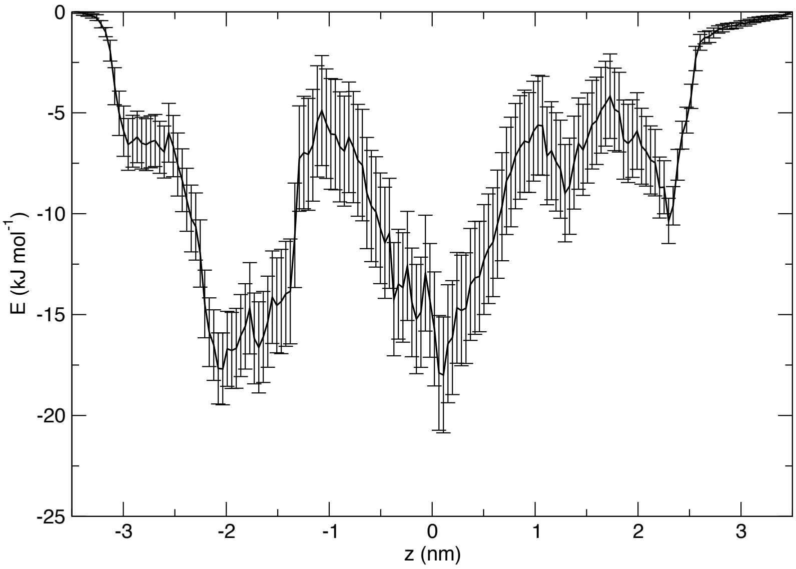
Autocorrelation times

The highest autocorrelation time is below 5 ns, as the simulation length is 75 ns these have all been covered at least 15 times.



Bootstrap Analysis

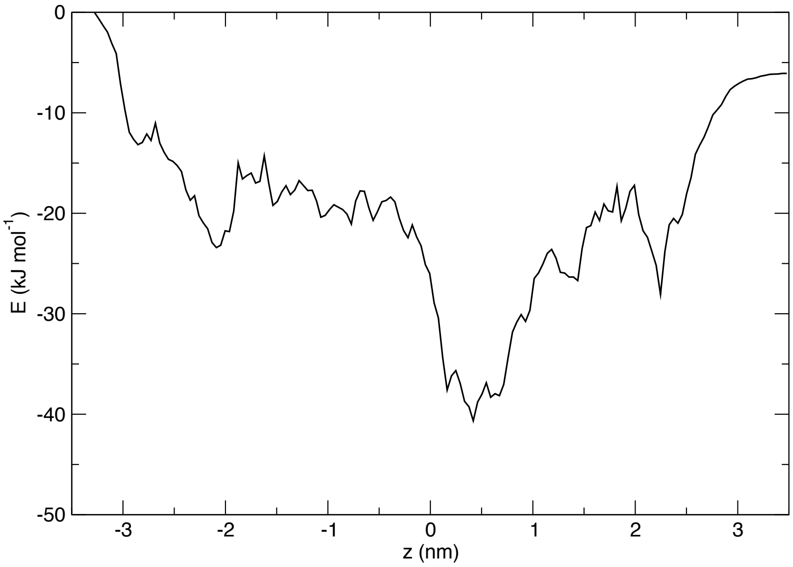
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 2.8 kJ.mol^{-1} in the protein



M2 Guanine

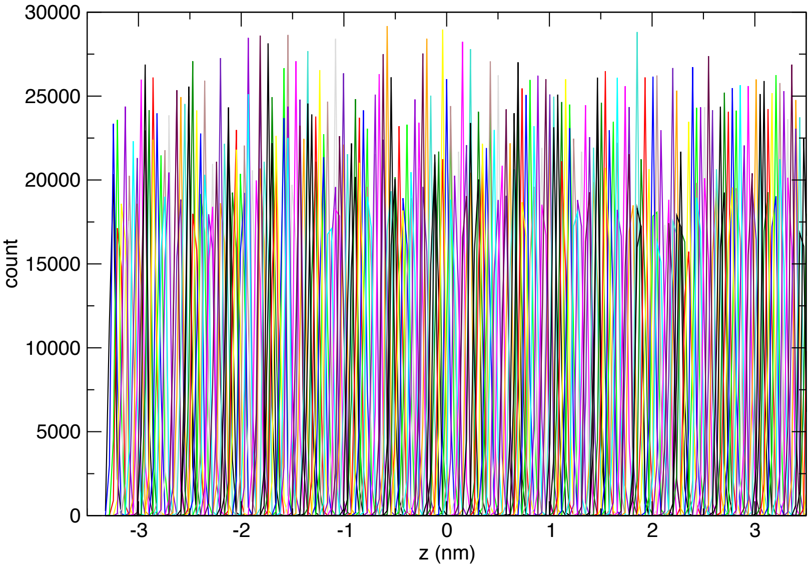
Non-cyclic profile

The difference in bulk water energies is approximately X kJ.mol^{-1}.



Histograms

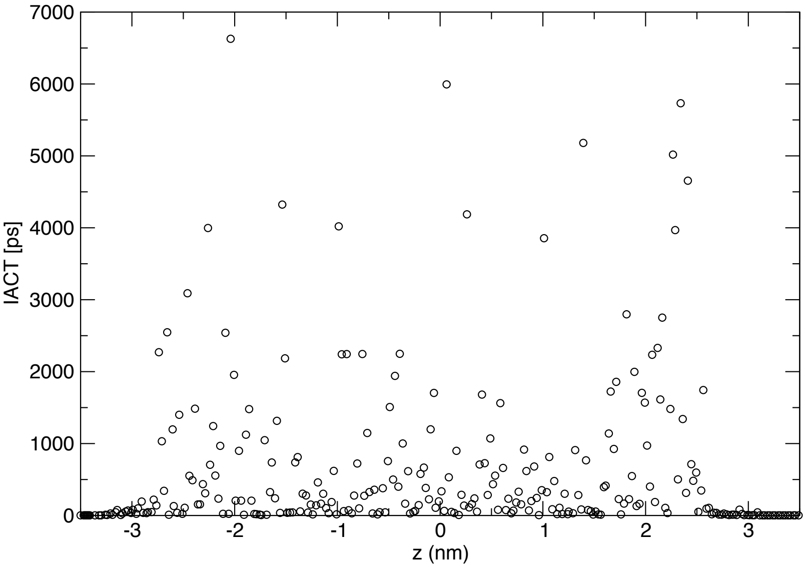
The histograms demonstrate extensive coverage of the principle axis.



Appendix E

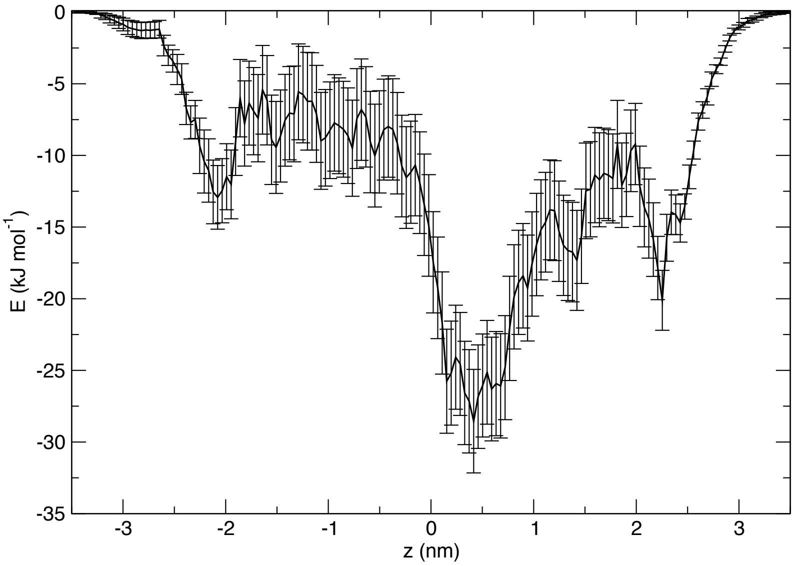
Autocorrelation times

The highest autocorrelation time is below 6.5 ns, with only 1 frame where it is above 5 ns, as the simulation length is 75 ns these have all been covered at least 35 times.



Bootstrap Analysis

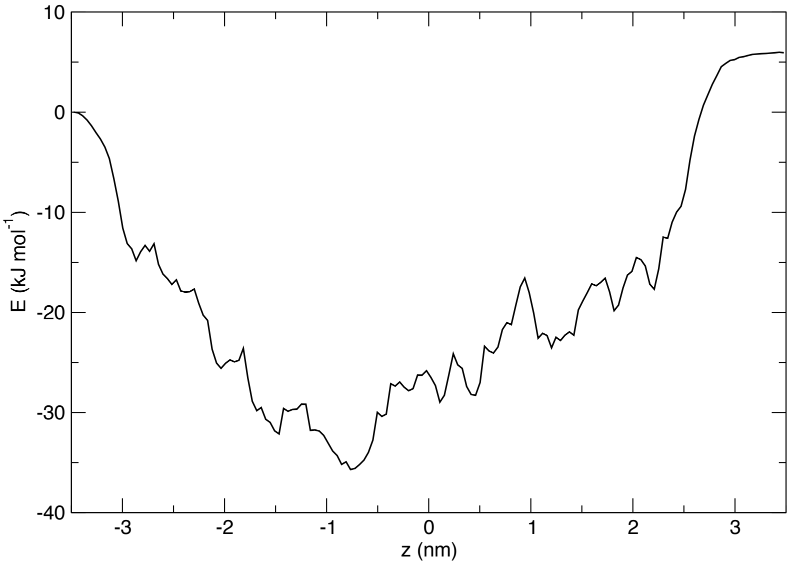
The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 3.4 kJ.mol^{-1} in the protein



M2 Thymine

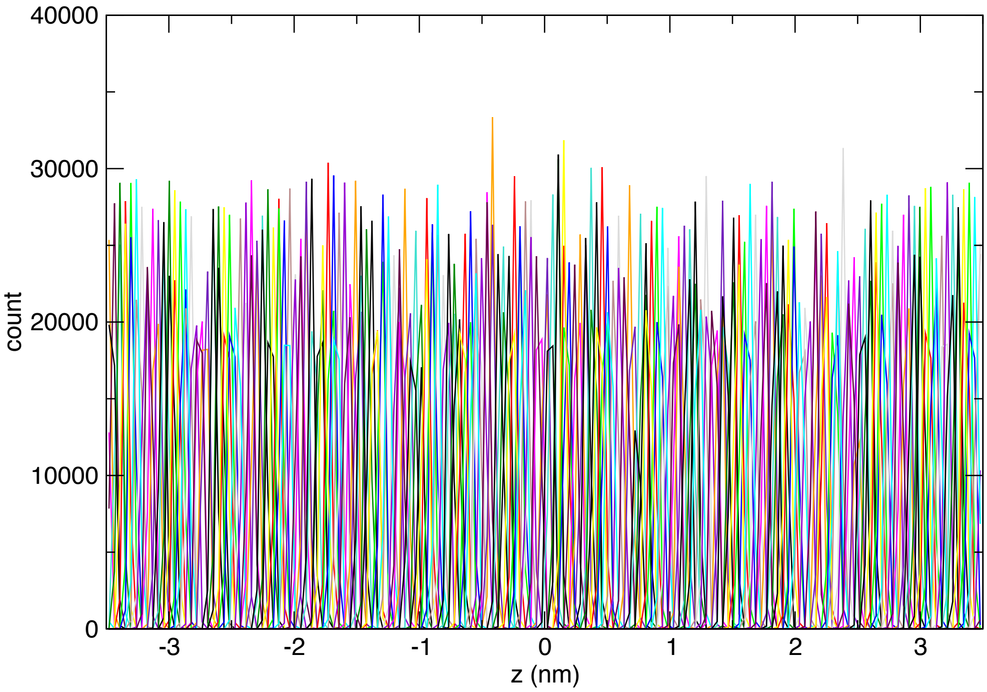
Non-cyclic profile

The difference in bulk water energies is approximately 6 kJ.mol^{-1}.



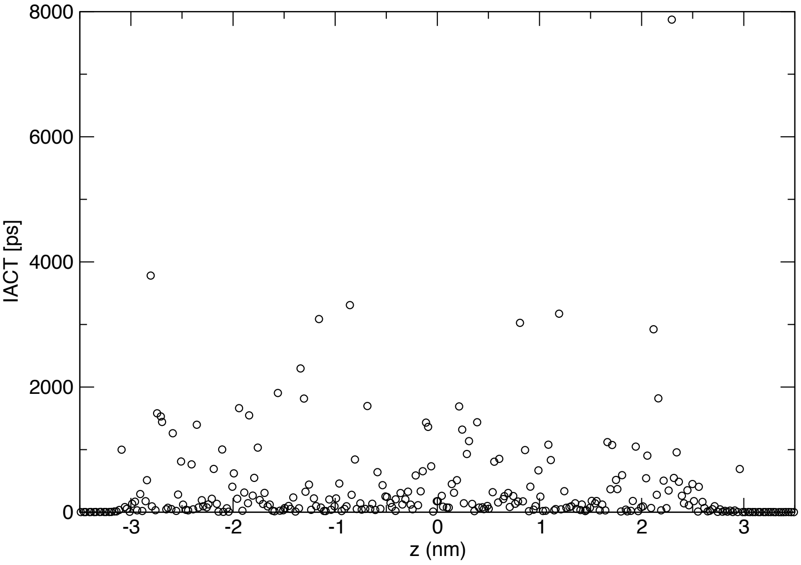
Histograms

The histograms demonstrate extensive coverage of the principle axis.



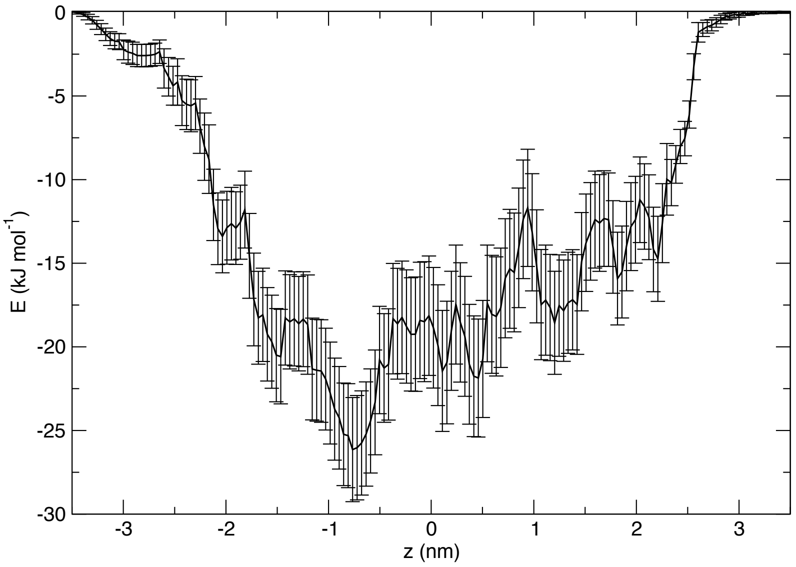
Autocorrelation times

The highest autocorrelation time is below 8 ns, with only 1 frame where it is above 5 ns, as the simulation length is 75 ns these have all been covered at least 9 times.



Bootstrap Analysis

The average of the 500 bootstraps are shown below. There are negligible errors in the bulk water, and 3.6 kJ.mol^{-1} in the protein



Appendix C Water analysis

Part 1: Calculation of approximate density of 1M NaCl

Mass of NaCl = 58.5 g.Mol^{-1}

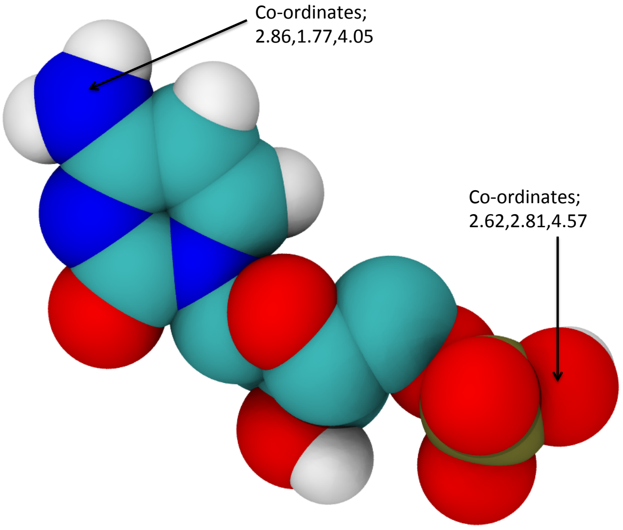
Mass of water = 1000 kg.m^{-3}

1 m^{3} of water = 1000 L

Mass of NaCl for 1000 L of 1 M solution = 1000 x 58.5 g.Mol^{-1} = 58.5 kg.mol^{-1}

Mass of 1 m^{3} of 1 M NaCl solution = 1058.5 kg.m^{-3} (Assuming negligible change in volume)

Part 2: Length of a CMP molecule



Difference in co-ordinates = 0.24 nm, 1.04 nm, 0.52 nm

From Pythagoras’s theorem, the distance between the two atoms is 1.13 nm.

Appendix D Infinite DNA

Length of a CMP molecule

Strands were downloaded from an online web server; http://structure.usc.edu/make-na/server.html

The files were then edited to be readable by the GROMOS set of forcefields; X e.g. A, T, C, G were replaced with DX. The 3' and 5' residues were labelled DX3 and DX5. In addition the C7 atom in thymine was replaced with C5M.

Appendix E Published work

Free-energy calculations reveal the subtle differences in the interactions of DNA bases with alpha-hemolysin.

Richard M.A. Manara[1], Andrew T. Guy[1], E. Jayne Wallace[2], Syma Khalid[1]*.

[1]Chemistry, Faculty of Natural and Environmental Sciences, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

[2]Oxford Nanopore Technologies Ltd, Edmund Cartwright House, 4 Robert Robinson Avenue, Oxford Science Park, Oxford, OX4 4GA, U.K.

*Email address for corresponding author: S.Khalid@soton.ac.uk

ABSTRACT: Next generation DNA sequencing methods that utilize protein nanopores have the potential to revolutionize this area of Biotechnology. While the technique is underpinned by simple physics, the wildtype protein pores do not have all of the desired properties for efficient and accurate DNA sequencing. Much of the research efforts have focused on protein nanopores, such as alpha-hemolysin from *Staphylococcus aureus.* However the speed of DNA translocation has historically been an issue, hampered in part by incomplete knowledge of the energetics of translocation. Here we have utilized atomistic molecular dynamics simulations of nucleotide fragments in order to calculate the potential of mean force (PMF) through α-hemolysin. Our results reveal specific regions within the pore that play a key role in the interaction with DNA. In particular, charged residues such as D121 and K137 provide stabilising interactions with the anionic DNA and therefore are likely to reduce the speed of translocation. These regions provide rational targets for pore optimization. Furthermore, we show that the energetic contributions to the protein-DNA interactions are a complex combination of electrostatics and short-range interactions, often mediated by water molecules.

KEYWORDS alpha-hemolysin, free energy, DNA, nanopores, molecular dynamics, next-generation sequencing

Appendix E

INTRODUCTION

In recent years, DNA sequencing using nanopores has received much scientific scrutiny due to its promise in next-generation sequencing.[1] The early promise has been realized by Oxford Nanopore Technologies in 2011-2012 and initial devices continue to be upgraded and improved to push the boundaries of what this technology can achieve. Nanopore based sequencing is based upon the same principles as stochastic biosensors:[2] an engineered protein with a nano-scale aperture is placed in a membrane that separates two chambers containing ionic solution. A voltage is then applied across the chambers, inducing the movement of ions through the pore. This ionic movement is detected as an electrical current. Any charged analyte molecules in the solution will also move through the pore, according to the electric field. In doing so, they cause partial blockage of the current. The extent of the current blockage can be used to identify the molecules, as the level of blocking is specific to a given analyte. Furthermore, it is also possible to calculate the concentration of the analyte by the frequency of blocking events, and also the strength of binding from the duration of the event.[3] This technique provides information at the single-molecule level and is even sensitive enough to discriminate between enantiomers of the same molecule. [4] Due to their polyanionc nature, DNA strands can also be driven electrophoretically through a nanopore by applying a potential difference. Each of the four DNA bases blocks the current by a different amount, allowing the sequence of the strand to be determined[1c]. While this is the basic principle, for strand sequencing additional complexity arises due to the base currents being context dependent.[1g]

Both proteinaceous and solid-state nanopores are currently being developed for sequencing DNA.[1a, 1e, 1f, 5] To date, much of the research on protein-based nanopores has focused on alpha-hemolysin, (aHL), a pore forming toxin from *Staphylococcus aureus.* aHL is a homo-heptameric protein composed of two main domains; a 14-stranded transmembrane beta-barrel and a large extramembranous cap with an internal vestibule.[6]

While aHL is a robust protein that is resistant to changes in pH and temperature over practical ranges, issues related to the high voltage thresholds required for DNA translocation render wild type aHL unsuitable for incorporation into a sequencing device.[7] To optimize aHL for DNA sequencing it is essential to understand on a molecular level how DNA behaves within the pore. Previously we have shown that DNA within both wild-type and mutant aHL pores can adopt a range of conformations. These conformations may sometimes lead to non-linear DNA translocation[8] i.e. the DNA bases do not necessarily exit the pore in the same order in which they entered. This phenomenon was primarily observed when charged residues were introduced into the barrel lumen. One way to overcome this potential problem is to ratchet the DNA into the pore with an enzyme, so it is always fully threaded through the pore. [1a] In other studies, it has been demonstrated that mutation of the lumen-lining residues of the aHL beta-barrel can have a marked effect on the rate of DNA translocation.[1d]

In this investigation we seek to identify the main protein-DNA interaction sites in the transmembrane region of wild-type aHL, and to calculate the energetic difference between the interactions of the protein with the four DNA bases. We use molecular dynamics (MD) simulations, since they enable us to explore the energetic relationships between the structure and dynamics of molecular systems at atomistic resolution. For example, MD has allowed investigation into protein and membrane dynamics, and ligands binding to receptors.[9] A major strength of MD is that it enables the direct study of energetics through techniques such as BAR[10] and the Weighted Histogram Analysis Method (WHAM).[11] Free energy techniques have previously been used by Coveney and co-workers to investigate polynucleotide translocation through aHL, using alternative techniques to those used here.[12] The authors focused upon the free energy change involved in translocation, as opposed to characterizing all of the interaction sites. They did however predict that the ring of K147 residues at the constriction site of aHL forms the dominant binding site.

In order to calculate the energetic difference between the interactions of the protein with the DNA bases we perform free energy calculations of nucleic acid fragments through aHL, using umbrella sampling and WHAM.[11]

METHODS

Simulations were performed in the GROMACS[13] package version 4.5.5,[14] using the GROMOS 53a6 forcefield[15] and the SPC water model.[16]

We simulate aHL using the pore model described previously.[17] We truncate the protein to only include the transmembrane beta-barrel, place the pore in a methane slab then solvate in neutralizing 1 M NaCl solution. Importantly the constriction site identified both by Coveney and Bayley is retained in the truncated barrel.[12, 18] Our model pore reduces the total atom count (including water and ions) from hundreds of thousands of atoms, to approximately twenty thousand atoms. Hence, this reduction in the system size leads to a decrease in the required simulation time, rendering our extensive free energy calculations feasible. As the system is a conformationally stable beta-barrel, we do not expect the use of restraints on the protein to influence the PMF. The model system has previously been validated.[17] The system was equilibrated in the NPT ensemble for 100 ns, using the Nosé-Hoover thermostat[19] and Parrinello-Rahman barostat[20] to 310 K and 1 Bar prior to the umbrella sampling frame setup. Once the substrates were inserted, the systems were energetically minimized using the steepest descent algorithm. The simulations were run between 150 ns to 250 ns with the first nanosecond discarded for pressure and temperature equilibration. Constraints were used for bond lengths using the LINCS algorithm. [21] Non-bonded interactions were treated with a cut-off of 1.2 nm and the Particle Mesh Ewald method was used for long-range electrostatics. [22]

We model the DNA molecule by fragmenting the mononucleotide into a phosphate and the bases, shown in figure 1. Fragmenting the mononucleotides in this way enables us to overcome the issue of long autocorrelation times expected for molecules as complex as mononucleotides, enabling well-resolved, and converged PMFs to be obtained.

The phosphate group was modeled as the single deprotonated form of phosphoric acid, given that 70% of the molecule is in this form at neutral pH. Furthermore, this form carries a net charge of -1, similar to the nucleic acid backbone. The structures of the nucleic acid fragments were constructed using the nucleotide parameters from GROMOS 53a6 as a reference and the additional hydrogen charge was adjusted so that the total charge was zero for the DNA bases and -1 for the phosphate.

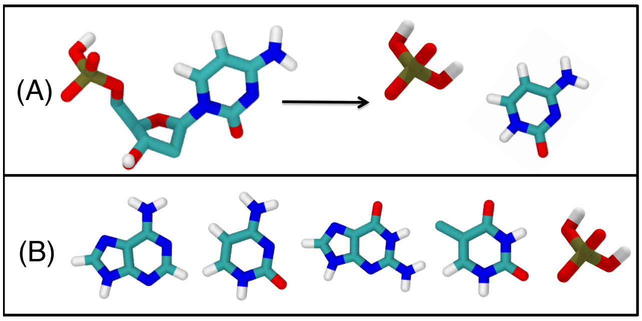


Figure 1: A. Fragmentation of a mononucleotide into a phosphate and base. B. The nucleotide fragments studied in our simulations. Bottom, from left to right: Adenine, Cytosine, Guanine, Thymine and Phosphate fragments. The molecules are colored as follows, oxygen in red, carbon in cyan, nitrogen in blue, hydrogen in white and phosphate in brown.

In house scripts were used to set up approximately 160 umbrella sampling windows separated by 5 Å along the principal axis, corresponding to the center of the aHL barrel, as shown in figure 2. Each window was simulated up to a maximum of 250 ns in duration, giving a maximum simulation time of 40 µs for each nucleotide fragment. The fragments were restrained along the principle axis of the protein, using a force of 1000 kJ.mol^{-1}.nm^{-1} remaining free to move laterally within the pore. The bias of the restraining force was removed using the weighted histogram analysis method (WHAM) *via* the GROMACS program g_WHAM.[23] WHAM was used to construct the cyclic potential of mean force (PMF) curves. For WHAM,[11] the number of bins was equal to the number of simulation windows. Autocorrelation times were also calculated using g_WHAM. Bootstrap analysis was performed using the Bayesian bootstrap method with an average of 500 bootstraps. The protein pore dimensions were analyzed using HOLE.[24] Visualization was performed with the VMD software package.[25]

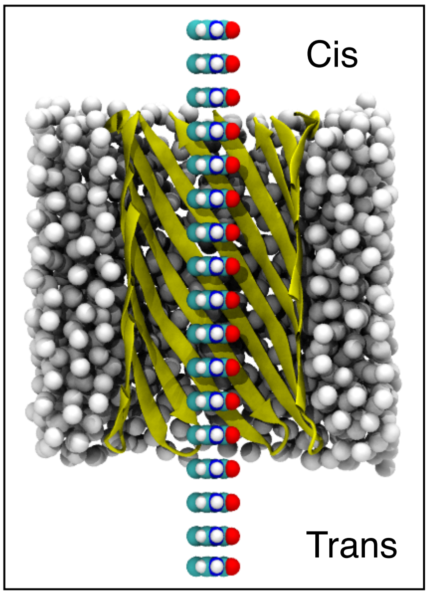


Figure 2: Thymine umbrella sampling windows showing how the position of the base is varied along the principal (z dimension) axis of the protein barrel. Only a selection of windows is shown for clarity and water and ions are omitted. The methane slab is shown as white spheres and yellow ribbons represent the protein backbone.

RESULTS AND DISCUSSION

Before discussing the energetic profiles of the various DNA fragments, it is useful to clarify the locations of the residues with sidechains pointing into the lumen of the β-barrel, these are shown in figure 3. In subsequent PMF profiles, the Gibbs free energy is plotted against the z coordinate. The cis and trans entrance to the β barrel correspond to z=2.4 nm and -2.6 nm respectively. The center of the pore (z=0.0 nm) corresponds to the location of the C-alpha atoms of G119.

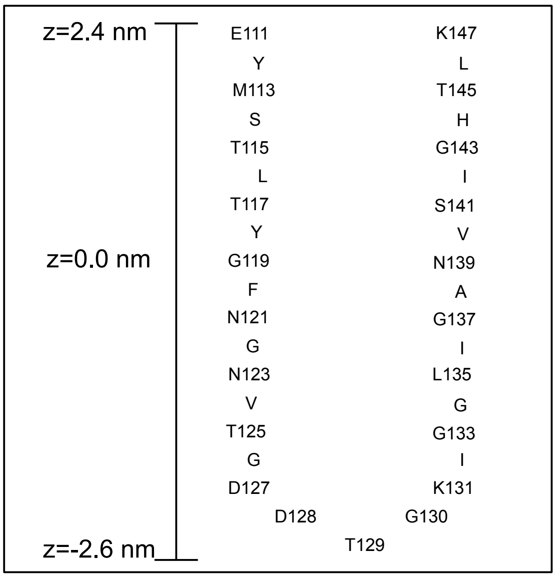


Figure 3: Residues within the aHL model pore, shown for one monomer in the beta-barrel. Numbered residues are solvated, and their sidechains point into the interior of the barrel. The approximate z coordinate of the residues is shown.

GUANINE PERMEATION ENERGETICS AND PORE DIMENSIONS

The pore radius profile provides an intuitive starting point when considering barriers to translocation. Here we will compare the pore radius against the PMF profile for guanine (figure 4).

We know that the narrowest constriction within the pore comprises the residues K147 and E111. This constriction has been identified as the major base recognition point, although other recognition points have also been identified.[1g]

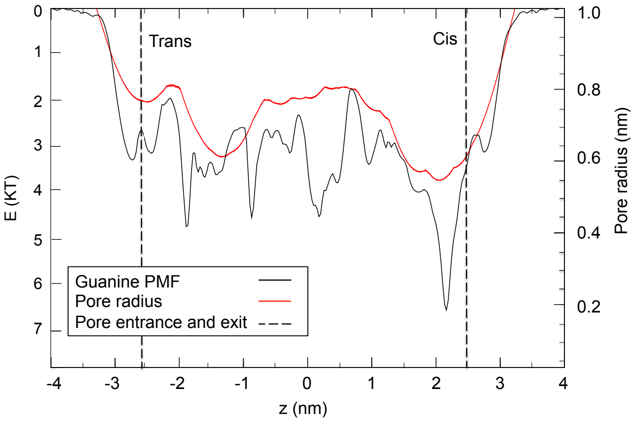


Figure 4: aHL pore radius profile overlayed on the guanine PMF profile.

The energetic peaks and troughs, in general, correspond to regions in which the pore is wider and narrower respectively. Specifically, the cis mouth of the pore provides an energetically favorable region for the bases; this is the narrowest region of the pore. The widest region of the pore is at z=0.6 to 0.9 nm, where T117 residues are flanked predominantly by glycine residues. This region is energetically less favorable presumably due in part to the lack of dispersion interactions between the base and the pore-lining residues. While the pore shape does not account for all of the features in the PMF profile it does show the importance of the contribution of the van der Waals interactions and solvation to the energetic profiles.

GUANINE INTERACTION WITH AHL

Closer inspection of the features observed in the PMF profile provides additional details of their molecular origins. First, we discuss the guanine PMF in detail, before performing a comparison with the other three bases of DNA. The PMF profile reveals no significant barrier to permeation along the entire beta-barrel (figure 2). Instead, there are a number of points that correspond to regions of particularly favorable energetics. We shall discuss each of these regions in turn.

z=-2.0 nm corresponds to the location of residues D127 and D128. We find that there are two distinct modes of interaction available for D127, either D127 forms a salt bridge with residue K131, located approximately 0.6 nm from the trans exit of the pore, or deprotonated D127 can form hydrogen bonds to guanine (figure 5). The base-protein interaction (where an interaction is defined as base-protein distance ≤ 0.3 nm) between guanine and either D127 or D128 is present in 75% of the 250 ns simulation. The lifetime of each hydrogen-bonding interaction has duration of less than 15 ns, with 1-4 hydrogen bonds existing simultaneously at any given time (figure 5). In this region there are on average ~20 water molecules within 0.4 nm of the base, these water molecules that freely interchange between hydrogen bonding with the base, the protein and other water molecules. Thus guanine is stabilized in this region through hydrogen bonding with D127, D128 and water molecules within the pore.

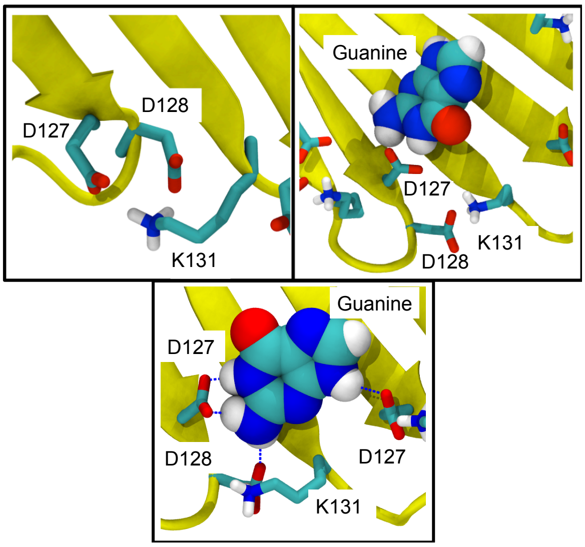


Figure 5: Alternative binding modes available to D127. Both aspartic acid residues D127 and D128 forming salt bridges with K131 (Top left) and D127 hydrogen bonding with guanine while D128 is still engaged in a salt bridge with K131 (Top right). Guanine is able to form 4 hydrogen bonds simultaneously with two flanking D127 residues and a lower D128 (Bottom). The yellow ribbons represent the protein backbone, the guanine is shown as VDW spheres

and the protein sidechains are shown in the stick representation.

At z=-0.9 nm, the sidechains of residues N123, N121 and L135 form a pocket in which the 5 membered ring of guanine can fit (figure 6). Importantly, hydrogen bonds are not formed with either the protein or water whilst the base is within this pocket. Thus the energetic stabilization of the base in this region is a result of dispersion interactions due to the shape complementarity of the guanine and the three-residue pocket.

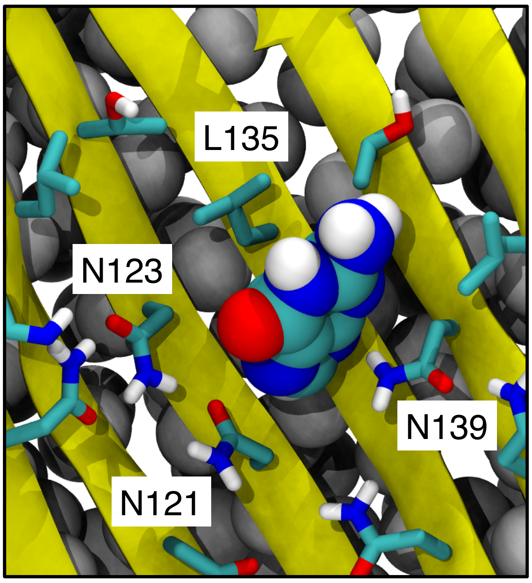


Figure 6: Shape complementarity between the guanine and protein sidechains provides favorable energetics in the region z=-0.9 nm.

At z=0.0 nm, guanine is able to form hydrogen bonds with residues N121 and N139. These residues are located on adjacent strands. The lateral distribution of guanine in the umbrella sampling calculation shows a distinct heptameric distribution, indicating the base is not stuck between any two specific beta-strands of the protein, but samples all seven beta-strands (figure 7). The base shows a preference for being located between the sidechains of N139, orientated such that the long axis of the molecule is parallel to the N139 sidechains and the oxygen is pointing into the lumen of the pore. Guanine is stabilized in this region by forming hydrogen bonds with N139. Occasionally guanine forms hydrogen bonds with two N139 residues from different monomers simultaneously, see figure 7. In this arrangement it can also form water mediated hydrogen bonds with sidechains of N121. Alternatively it can form direct hydrogen bonds to N121, the latter are not however generally observed simultaneously with hydrogen bonds to N139. In the region z=0.0-0.6 nm guanine forms hydrogen bonds with residues T117 and S141 which are located on adjacent strands.

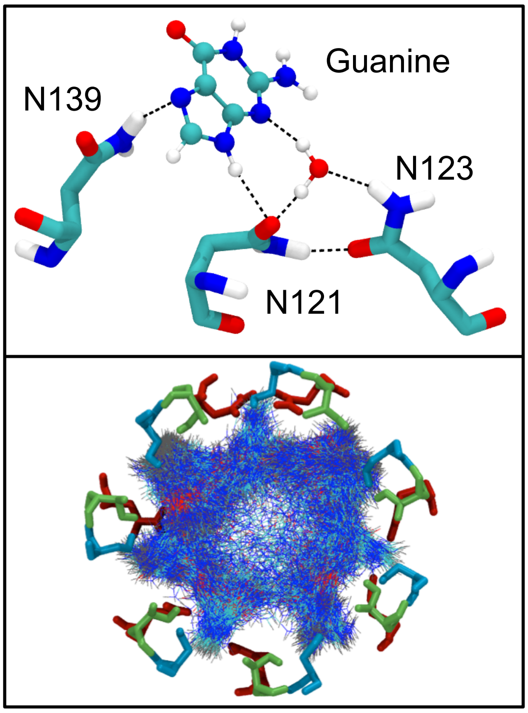


Figure 7: Top: Alternative hydrogen-bonding modes available to guanine at z=0.0 nm. Direct hydrogen bonds formed simultaneously to N139 and N121 residues are shown as well as a water-bridged hydrogen bonding interaction to N123. Bottom: guanine locations across the trajectory superimposed on each other (viewed down the protein principle axis from the cis entrance) shows good sampling of the heptamer, the sticks are asparagine residues, red for N139, green for N121 and blue for N123.

At z=2.4 nm, K147 and E111 form the narrowest constriction in the pore. We have previously shown that this region provides a barrier for translocation of DNA single strands owing to the hydrogen-bonding and permanent electrostatic interactions of the DNA strands and the sidechains of the protein residues.[8] Here our calculations show that the adjacent position of the acidic and basic sidechains can lead to a mimicking of the Watson-Crick hydrogen bonding arrangement with guanine, see figure 8. It is useful to point out here that for the guanine nucleotide, K147 may preferentially form a salt bridge with the phosphate group rather than the oxygen of guanine, but due to the homo-heptameric architecture of aHL, both interactions may be possible simultaneously, see the SI for evidence of sampling the heptamer. Alternative binding modes are also observed, with the amine group of guanine hydrogen bonding to E111. We note that Stoddart *et al*. showed that mutation of residues E111 and K147 to asparagine leads to weakened recognition of bases in this region,[1g] and also demonstrated the importance of mutation of M113.[26] They surmised that this is likely due to reduced interaction between the pore and the DNA when the charged residues are mutated. Our analysis of the energetics of DNA-pore interactions in this region support their view. Maglia *et al*. have shown the impact on ssDNA translocation by charged residues within the aHL barrel, our results are in qualitative agreement with their observations as charged residues lead to features in the PMF curves.[7]

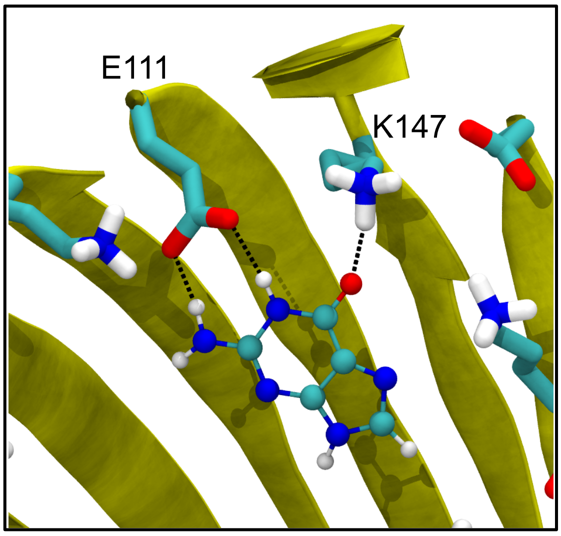


Figure 8: Hydrogen bonding between guanine and K147 and E111, at the narrowest point of the aHL pore.

COMPARATIVE FREE ENERGY PROFILES OF THE DNA BASES

In order to understand the molecular origins of the ability of aHL to discriminate between the bases of DNA, it is important to compare their individual energetic profiles. In addition to guanine, we have also constructed the PMF profiles of adenine, cytosine, thymine, and the phosphate moiety, see figure 9. In the supplementary information we also provide a short discussion of the full cytosine mononucleotide, and show that the profiles is not altered in most regions, by addition of the sugar moeity.

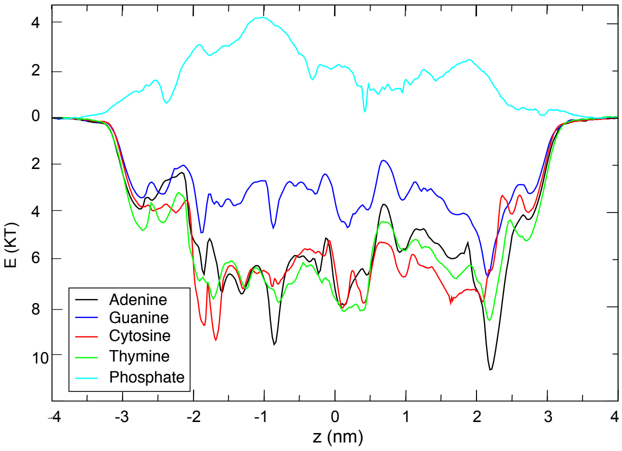


Figure 9: The PMF profiles for all four DNA bases and the phosphate group.

The PMF profiles of all four bases share a similar shape profile; the most notable difference is that the guanine PMF is at higher energy than the other bases in all regions within the barrel. The other 3 bases are generally within thermal error (1 kT is approx. 2.5 kJmol^{-1} at 310 K). Given their size and chemical similarities it is reasonable to expect the energetic profiles of the purines to be comparable to each other and likewise for the pyrimidine molecules, which is generally the case. The shapes of the free energy curves for adenine and guanine are similar to each other and likewise those of thymine and cytosine are similar to each other.

The most striking difference between the purine and pyrimidine profiles is in the region z=-0.9 nm. As previously explained when discussing the guanine profile, the purine profiles feature a trough in this region that arises from the shape complementarity of the 5-membered ring and the pocket formed from the sidechains of N139, N121 and S141. This energetic well is not observed in the PMF profiles of the pyrimidines, presumably due to the lack of a 5-membered ring in cytosine and thymine.

It is interesting to note that at the narrowest constriction within the protein pore, near E111/K147 (at z=2.0-2.1 nm), the energetics for the adenine-protein interaction are more favorable than for the other bases. We hypothesize that this may be the molecular origin of the slower translocation speed of polyA DNA strands compared to polyC strands, reported from previous experimental and computational studies.[27]

POTENTIAL OF MEAN FORCE PROFILE OF THE PHOSPHATE GROUP

For DNA sequencing applications, either DNA strands or individual nucleotides will be detected. Hence, it is useful to consider the energetics of the charge-carrying phosphate group, H2PO4- (figure 9). As expected, the energetic profile for phosphate generally follows the opposite trend to the hydrophobic bases. The phosphate experiences an energetic barrier throughout the length of the barrel.

The peak in the PMF profile occurs at z=-1.0 nm. This corresponds to a leucine (L135) and an asparagine (N123) residue, and is at a narrow region in the barrel. Given the hydrophobic nature of leucine, there are fewer protein-phosphate hydrogen bonds and no electrostatic interactions possible here.

Interestingly at z=2.4 nm where the pore narrows at the cis mouth, there is no appreciable well or barrier. This is most likely due to a cancellation of electrostatic attraction/repulsion with the charged residues K147 and E111 respectively.

Our data shows very different energetics experienced by phosphate in aHL compared to the phosphate transporter OprP, which is lined with arginine residues.[28] The phosphate group alone is generally at unfavorable energies throughout the hemolysin barrel, compared with favorable energetics within OprP.

CONVERGENCE AND ERROR ANALYSES

It is important to evaluate the convergence of the simulations and to estimate errors associated with the PMF curves in order to be confident about their validity. Figure 10 shows bootstrap analysis of the adenine PMF curve. Encouragingly the errors are negligible in the bulk water regions and have a maximum of ~±1 KT^{-1} in the protein pore. These values compare favourably with similar studies of free energies reported in the literature._ENREF_37[29] The histograms, see the supplementary information (SI), show overlap between all of the umbrella sampling windows, indicating that none of the regions along the principal axis of the pore are unsampled. Autocorrelation times of the energies (SI) are small (below 5 ns on average) compared to simulation length (150-250 ns) and have therefore been covered extensively. The non cyclised PMFs, i.e. the raw data, are presented in the SI.

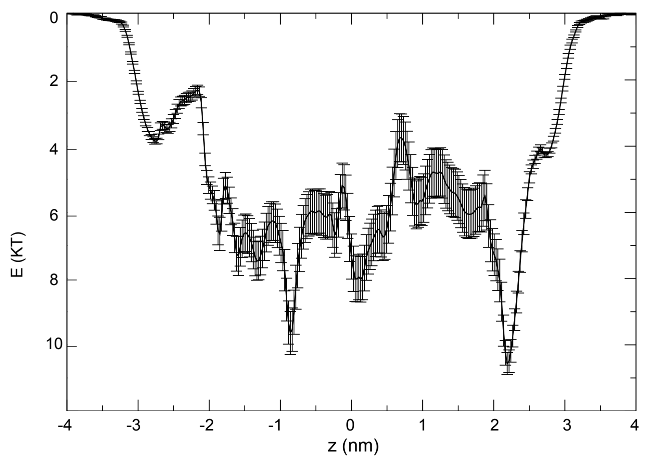


Figure 10: Bootstrap analysis of the adenine fragment PMF. The profile is an average of 500 bootstraps and shown with an associated error estimate.

It is important to place the results into context by considering the limitations of the present study. These arise primarily due to the use of a model pore system, in which only the beta-barrel region of the protein has been sampled. However given that base discrimination occurs within the barrel, for the purposes of this study neglecting the vestibule is reasonable. For a consideration of the energetics of base entry into the pore, it would be essential to incorporate a representation of the pore vestibule into the model.

CONCLUSIONS

In conclusion, we have shown that all the DNA bases experience favorable energetics throughout the beta-barrel. The trough in the PMF profiles occurs at the cis entrance, the narrowest region of the barrel. At the cis entrance, there is a ring of glutamate and lysine residues that provide not only a steric constriction, but also the possibility for various hydrogen-bonding and electrostatic interactions. In addition, there is potential for discrimination of purines by exploiting their size and shape. For example, our simulations have revealed a region in which the purines are stabilized by virtue of the shape complementarity of the pocket, which is formed by the sidechains pointing into the lumen of the barrel and the five-membered ring of the bases. Interestingly, the PMF profiles reveal that the permeation of guanine throughout the pore is less energetically favorable than the other bases. Unfortunately, the energetics of guanine rich regions are not straightforward to study experimentally, due to the formation of the G_quadruplex[23,24]. Future work will include a consideration of the guanine monophosphate nucleotide to further explore the origins of the free-energy differences exhibited by this base compared to the others.

We expect the energetics of a strand of ssDNA to be similar to the predicted PMFs presented here. However due to the increase in flexibility compared to the fragments we predict phase space would not be as thoroughly explored within the same simulation timescales.

Overall, the free energy landscape for DNA base permeation through the transmembrane region of aHL is rather more complex than, for example, a phosphate ion within the phosphate specific channel OprP.[28] In general, the dimensions of the aHL barrel are such that the narrower regions provide greater stability for the DNA within the pore. Although there are no energetic barriers to base permeation, there are a number of regions that are more favorable than others. We predict that DNA translocation will be slightly slower in these regions, due to stabilizing protein-DNA interactions. Importantly, optimization of the aHL barrel for improved base discrimination is most likely to be achieved by mutations in these regions to improve the differential 'binding' of the four bases. Our simulations have shown that the combinations of protein-base interactions present in this system are difficult to predict *a priori* as they are a complex mix of steric and electrostatic interactions, often also involving contribution from local water molecules. Also, given the inherent flexibility of ssDNA and the multiple binding modes observed in this study, it is conceivable for strand sequencing that multiple binding modes may exist simultaneously, this is particularly likely for DNA that is ratcheted in with an enzyme and fully threaded across the pore. Consequently free energy characterization *via* molecular simulation provides an invaluable tool for facilitating modification and optimization of nanopores for sequencing applications.

SUPPORTING INFORMATION

The autocorrelation times; bootstrap analysis; histograms; and non-cyclized profiles are available at in the SI. This material is available free of charge *via* the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

Syma Khalid is the corresponding author.

Author Contributions

All authors have given approval to the final version of the manuscript

Funding Sources

Richard Manara and Andrew Guy are funded by Oxford Nanopore Technologies.

ACKNOWLEDGMENT

We acknowledge the use of the University of Southampton high performance computing resources, Iridis3 and Iridis4. We would also like to thank J. Parkin and H. Bayley for helpful discussions.

ABBREVIATIONS

aHL, alpha-hemolysin; MD, Molecular Dynamics, PMF, Potential of mean force; WHAM, weighted histogram analysis method.

REFERENCES

1. (a) Schneider, G. F.; Dekker, C., DNA sequencing with nanopores. *Nat. Biotechnol.* **2012,** *30* (4), 326-328; (b) Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W., Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. USA* **1996,** *93* (24), 13770-13773; (c) Clarke, J.; Wu, H. C.; Jayasinghe, L.; Patel, A.; Reid, S.; Bayley, H., Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **2009,** *4* (4), 265-270; (d) Rincon-Restrepo, M.; Milthallova, E.; Bayley, H.; Maglia, G., Controlled Translocation of Individual DNA Molecules through Protein Nanopores with Engineered Molecular Brakes. *Nano Lett.* **2011,** *11* (2), 746-750; (e) Venta, K.; Shemer, G.; Puster, M.; Rodriguez-Manzo, J. A.; Balan, A.; Rosenstein, J. K.; Shepard, K.; Drndic, M., Differentiation of Short, Single-Stranded DNA Homopolymers in Solid-State Nanopores. *Acs Nano* **2013,** *7* (5), 4629-4636; (f) Wallace, E. V. B.; Stoddart, D.; Heron, A. J.; Mikhailova, E.; Maglia, G.; Donohoe, T. J.; Bayley, H., Identification of epigenetic DNA modifications with a protein nanopore. *Chem. Commun* **2010,** *46* (43), 8195-8197; (g) Stoddart, D.; Maglia, G.; Mikhailova, E.; Heron, A. J.; Bayley, H., Multiple Base-Recognition Sites in a Biological Nanopore: Two Heads are Better than One. *Angew Chem. Int Edit.* **2010,** *49* (3), 556-559.

2. Bayley, H.; Cremer, P. S., Stochastic sensors inspired by biology. *Nature* **2001,** *413* (6852), 226-230.

3. Dekker, C., Solid-state nanopores. *Nat. Nanotechnol.* **2007,** *2* (4), 209-215.

4. (a) Boersma, A. J.; Bayley, H., Continuous Stochastic Detection of Amino Acid Enantiomers with a Protein Nanopore. *Angew Chem. Int. Edit.* **2012,** *51* (38), 9606-9609; (b) Kang, X. F.; Cheley, S.; Guan, X. Y.; Bayley, H., Stochastic detection of enantiomers. *J. Am. Chem. Soc* **2006,** *128* (33), 10684-10685.

5. Aksimentiev, A.; Heng, J. B.; Timp, G.; Schulten, K., Microscopic kinetics of DNA

translocation through synthetic nanopores. *Biophys. J.* **2004,** *87* (3), 2086-2097.

6. Song, L. Z.; Hobaugh, M. R.; Shustak, C.; Cheley, S.; Bayley, H.; Gouaux, J. E., Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **1996,** *274* (5294), 1859-1866.

7. Maglia, G.; Restrepo, M. R.; Mikhailova, E.; Bayley, H., Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. *Proc. Natl. Acad. Sci. USA* **2008,** *105* (50), 19720-19725.

8. Guy, A. T.; Piggot, T. J.; Khalid, S., Single-Stranded DNA within Nanopores: Conformational Dynamics and Implications for Sequencing; a Molecular Dynamics Simulation Study. *Biophys. J.* **2012,** *103* (5), 1028-1036.

9. Durrant, J. D.; McCammon, J. A., Molecular dynamics simulations and drug discovery. *Bmc Biol.* **2011,** *9.*

10. Bennett, C. H., Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. *J. Comput. Phys.* **1976,** *22* (2), 245-268.

11. Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M., The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. *J. Comput. Chem.* **1992,** *13* (8), 1011-1021.

12. Martin, H. S. C.; Jha, S.; Howorka, S.; Coveney, P. V., Determination of Free Energy Profiles for the Translocation of Polynucleotides through alpha-Hemolysin Nanopores using Non-Equilibrium Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2009,** *5* (8), 2135-2148.

13. Bekker, H.; Berendsen, H. J. C.; Dijkstra, E. J.; Achterop, S.; Vondrumen, R.; Vanderspoel, D.; Sijbers, A.; Keegstra, H.; Reitsma, B.; Renardus, M. K. R., Gromacs - a Parallel Computer for Molecular-Dynamics Simulations. *Physics Computing '92* **1993,** 252-256.

14. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-54.

15. Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F., Validation of the 53A6 GROMOS force field. *Eur. Biophys. J. Biophy.* **2005,** *34* (4), 273-284.

16. H. J. C. Berendsen, J. P. M. P., W.F. van Gunsteren, J. Hermans, SPC. *Intermol. Forces* **1981**, p. 331.

17. Bond, P. J.; Guy, A. T.; Heron, A. J.; Bayley, H.; Khalid, S., Molecular Dynamics Simulations of DNA within a Nanopore: Arginine-Phosphate Tethering and a Binding/Sliding Mechanism for Translocation. *Biochemistry-Us* **2011,** *50* (18), 3777-3783.

18. Stoddart, D.; Heron, A. J.; Mikhailova, E.; Maglia, G.; Bayley, H., Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl. Acad. Sci. USA* **2009,** *106* (19), 7702-7707.

19. Cheng, A. L.; Merz, K. M., Application of the Nose-Hoover chain algorithm to the study of protein dynamics. *J. Phys Chem-Us* **1996,** *100* (5), 1927-1937.

20. Nosé, S.; Klein, M. L., Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* **1983,** *50* (5), 1055-1076.

21. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M., LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997,** *18* (12), 1463-1472.

22. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys* **1993,** *98* (12), 10089-10092.

23. Hub, J. S.; de Groot, B. L.; van der Spoel, D., g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010,** *6* (12), 3713-3720.

24. Smart, O. S.; Neduvelil, J. G.; Wang, X.; Wallace, B. A.; Sansom, M. S., HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. graphics* **1996,** *14* (6), 354-60, 376.

25. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J. Mol. graphics* **1996,** *14* (1), 33-8, 27-8.

26. Stoddart, D.; Heron, A. J.; Klingelhoefer, J.; Mikhailova, E.; Maglia, G.; Bayley, H., Nucleobase Recognition in ssDNA at the Central Constriction of the alpha-Hemolysin Pore. *Nano Lett.* **2010,** *10* (9), 3633-3637.

27. (a) Meller, A.; Nivon, L.; Brandin, E.; Golovchenko, J.; Branton, D., Rapid nanopore discrimination between single polynucleotide molecules. *Proc. Natl. Acad. Sci. USA* **2000,** *97* (3), 1079-1084; (b) Wells, D. B.; Abramkina, V.; Aksimentiev, A., Exploring transmembrane transport through alpha-hemolysin with grid-steered molecular dynamics. *J. Chem. Phys* **2007,** *127* (12).

28. Pongprayoon, P.; Beckstein, O.; Wee, C. L.; Sansom, M. S. P., Simulations of anion transport through OprP reveal the molecular basis for high affinity and selectivity for phosphate. *Proc. Natl. Acad. Sci. USA* **2009,** *106* (51), 21614-21618.

29. Yang, C.; Kim, E.; Pak, Y., Potential of Mean Force Simulation by Pulling a DNA Aptamer in Complex with Thrombin. *B Korean Chem. Soc* **2012,** *33* (11), 3597-3600.

Nanomaterials **2014**, *4*, 1-x; doi: 10.3390/nano40x000x

OPEN ACCESS

nanomaterials

ISSN 2079-4991

www.mdpi.com/journal/nanomaterials

Communication

The nucleotide capture region of alpha hemolysin: insights into nanopore design for DNA sequencing from Molecular Dynamics simulations

Richard M. A. Manara [1], Susanna Tomasio [1], Syma Khalid [1*]

[1] Building 27, School of Chemistry, University Rd, University of Southampton, Highfield Campus, Southampton SO17 1BJ; E-Mail: rm16g09@soton.ac.uk, susana@cresset-group.com, S.Khalid@soton.ac.uk

* Correspondence should be addressed to Dr Khalid; E-Mail: S.Khalid@soton.ac.uk;

Tel.: +44-2380-594-176 (ext. 123); Fax: +44 23 8059 3781

.

External editor:

Published:

Abstract: Nanopore technology for DNA sequencing is constantly being refined and improved. In strand sequencing a single strand of DNA is fed through a nanopore and subsequent fluctuations in the current are measured. A major hurdle is that the DNA is translocated through the pore at a rate that is too fast for the current measurement systems. An alternative approach is 'exonuclease sequencing', in which an exonuclease is attached to the nanopore that is able to process the strand, cleaving off one base at a time. The bases then flow through the nanopore and the current is measured. This method has the advantage of potentially solving the translocation rate problem, as the speed is controlled by the exonuclease. Here we consider the practical details of exonuclease attachment to the protein alpha hemolysin. We employ molecular dynamics simulations to determine the ideal (a) distance from alpha hemolysin, and (b) the orientation of the monophosphate nucleotides upon release from the exonuclease such that they will enter the protein. Our results indicate an almost linear decrease in the probability of entry into the protein with increasing distance of nucleotide release. The nucleotide orientation is less significant for entry into the protein

Keywords: Alpha-hemolysin; Exonuclease sequencing; Molecular Dynamics; Nanopore sequencing;

1. Introduction

DNA sequencing using nanopores is a method that allows for direct analysis of genomic DNA [1-8]. Benefits of nanopore-based sequencing include detection of epigenetic markers [9,10], which is not possible using chemical methods as well as reducing costs compared to more traditional methods. The early promise has been realized by Oxford Nanopore Technologies in 2011-2012, however the initial devices continue to be upgraded and improved to push the boundaries of what this technology can achieve.

The pore forming toxin, alpha-hemolysin from *S. aureus* [11] is perhaps the best studied example of a proteinaceous nanopore for DNA sequencing [12]. This protein has two domains, the vestibule or cap domain and the transmembrane pore domain (Figure 1) [11]. While aHL is a robust protein that is resistant to changes in pH and temperature over practical ranges, issues related to the high voltage thresholds required for DNA translocation render wild type aHL unsuitable for incorporation into a sequencing device. However engineered mutants have shown considerable success for DNA sequencing, although they too are continuously being improved for more efficient DNA sequencing. For example, methods to improve the accuracy of the device by decreasing the rate of translocation of ssDNA [2,13] through the protein pores are being investigated. It has been proposed that incorporation of an exonuclease enzyme, which would split the ssDNA into individual mononucleotides before they enter the nanopore, may be an efficient way to reduce the translocation rate [10,14]. Selection of an exonuclease enzyme for incorporation into an alpha-hemolysin based sequencing device is dependent upon practical considerations such as temperature and pH dependence, size and conformational flexibility of the enzyme. Furthermore it is imperative that the enzyme is positioned relative to the alpha-hemolysin such that there is a practical level of confidence mononucleotides exiting the enzyme will enter the alpha-hemolysin and not simply diffuse away in solution. This information is vital for the practical design and construction of the chimera protein.

Molecular dynamics simulations provide a route to study the translocation of biological molecules through nanoscale pores at levels of detail that are difficult to achieve with experimental methods alone [15-19]. Here we employ atomistic molecular dynamics simulations to characterize the 'nucleotide capture area' of the wildtype alpha-hemolysin protein. Capture of the cytosine monophosphate (CMP), was investigated by embedding alpha-hemolysin in a 1,2-dimyristoyl-sn-glycero-3-phosphocholine (DMPC) bilayer and solvating in 1M NaCl, resulting in a simulation system composed of 273,000 atoms. Only one of the four mononucleotides was simulated due to the similarity in their masses with respect to the protein. The distance of CMP from the protein was varied. Specifically, CMP was positioned above the protein on the vestibule side such that it was located centrally, above the entrance to the protein (where central is defined here as the average co-ordinates of the C-alpha atoms of K8 residues which are in the vestibule domain of the protein) as shown in Figure 1. Simulations were performed with the nucleotide initially positioned at 10, 15, 20, 30 and 40 angstroms above the ring of K8 residues, along the principal axis of the protein. Additional systems were set up in which, at a distance of 30 angstroms from the K8 residues, the nucleotides were displaced in the plane parallel to the membrane, by 5, 10, 15 and 20 angstroms. Two sets of simulations were performed for each location of CMP, in which the CMP was orientated with either the phosphate moiety furthest or closest to the vestibule entrance, which we define as phosphate 'up' or 'down' orientations respectively, shown in the inset of Figure 1. For each location, 20 independent simulations of 5 ns duration were performed at 310 K, with an applied electric field of magnitude 0.1 V.nm^{-1} equivalent to approximately 350 mV across the membrane. Simulations were performed using the GROMACS [20] software package, version 4.5.5 [20,21] with the GROMOS53a6 [22] forcefield and the SPC model of water [23]. Further details of the methodology are provided in the Supporting Information (SI).

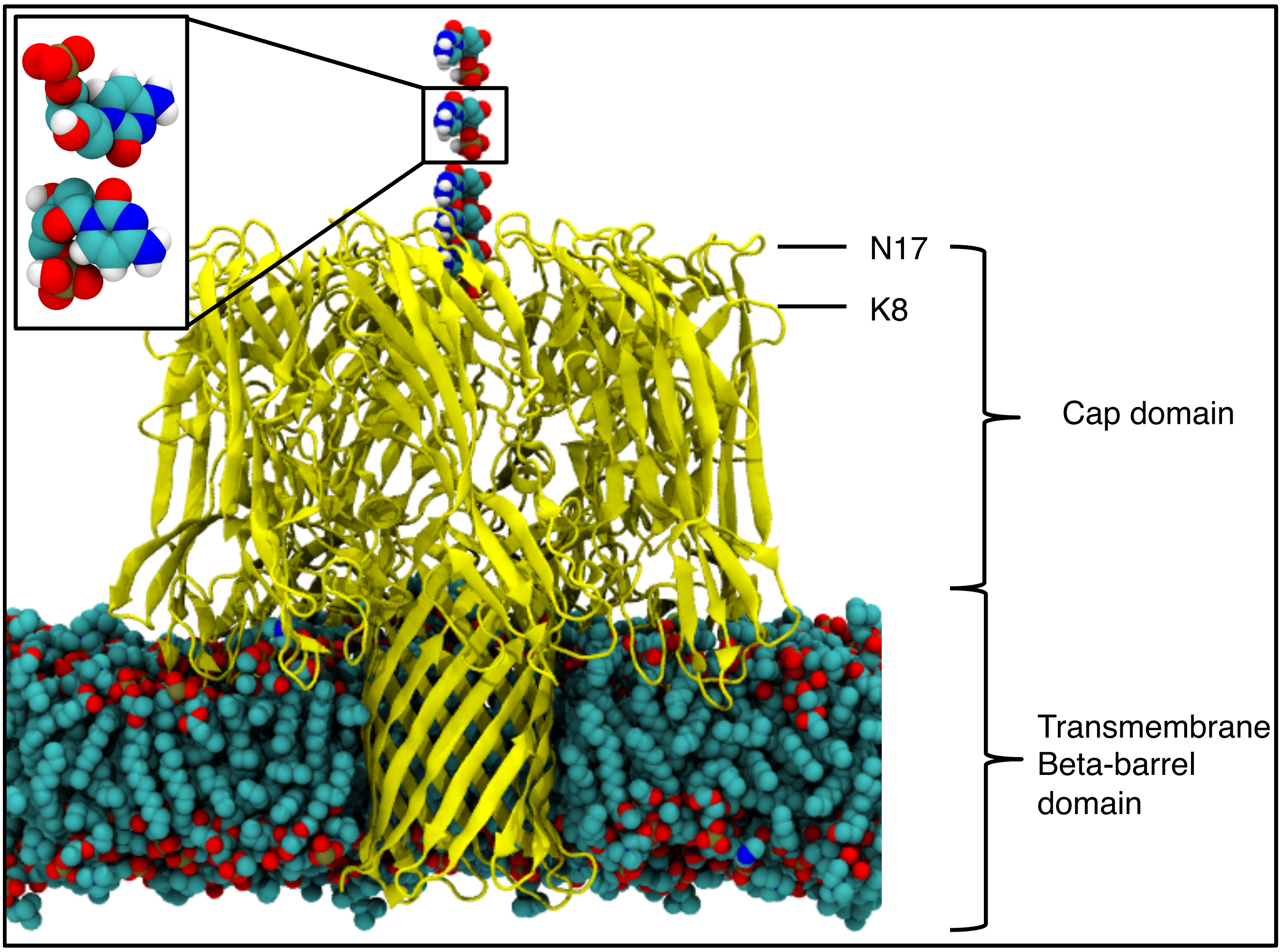


Figure 1. The aHL protein in a DMPC bilayer, with the mononucleotide positioned above the vestibule entrance, where each mononucleotide represents an individual simulation. The protein is represented by yellow ribbons and for other atoms: carbon is shown in cyan, oxygen in red, nitrogen in blue, phosphorus in brown and hydrogen in white. The waters and ions are excluded for clarity. **(Inset)** The phosphate orientations used, termed the 'up' and 'down' orientations, shown top and bottom respectively.

2. Results and Discussion

For the purposes of clarity when reporting our results, we use two terms for the behaviour of the nucleotide with respect to the protein: 'capture' and 'possible capture'. These are defined as follows: when the entire CMP is below the ring of C-alpha atoms of N17, it is regarded as captured as over all simulations we observed no examples of exit from the vestibule entrance once interacting with the protein below this region. In contrast, if the CMP is entirely above this ring, it is deemed not captured. The final alternative is when simultaneously parts of the CMP are above and below the N17 ring, which we refer to as possible capture, as it was observed that CMP in this region, which we describe as the 'edge' of the vestibule entrance, is capable of either entering the vestibule or diffusing into solution.

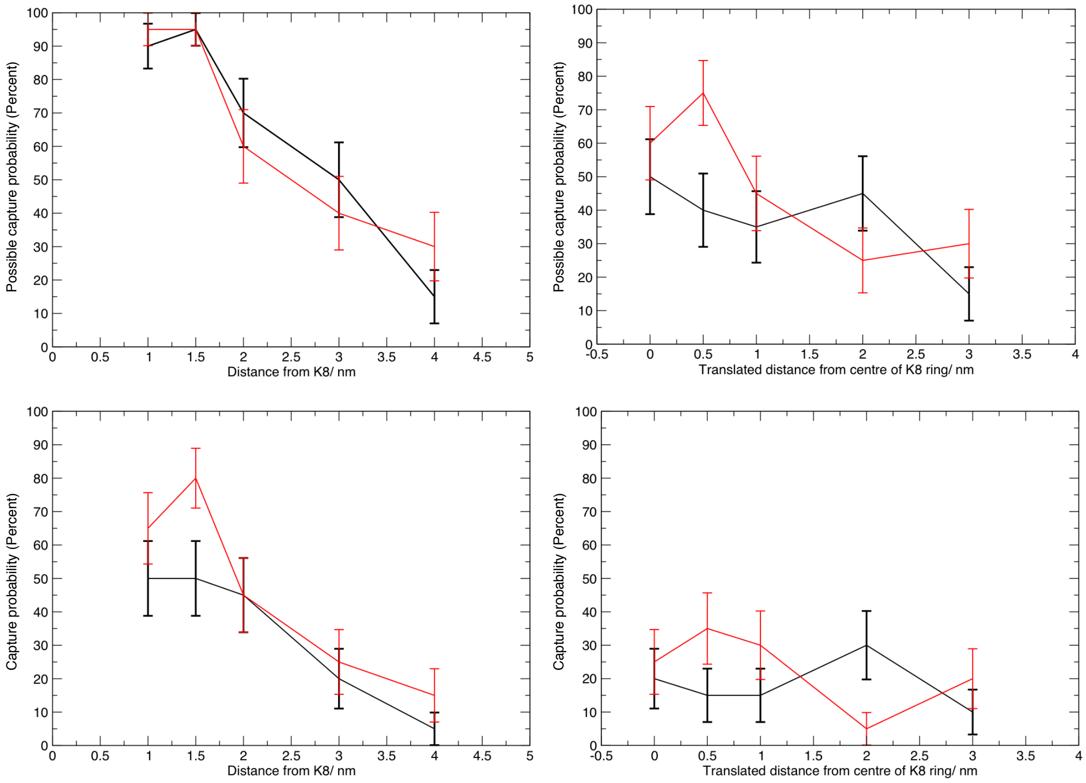


Figure 2. Plots showing the relationship between release location and probability of capture. **(Top)** The highest possible probability of entry for **(Left)** the distance study and **(Right)**, the displacement study, e.g. these are the simulations where either capture or possible capture is observed. **(Bottom)** The lowest possible probability of entry for **(Left)** the distance study and **(Right)**, the displacement study, e.g. these are the simulations where capture is observed. The black lines are for the phosphate up orientation and the red lines are the phosphate down orientation. Error bars were calculated by using the standard error, treating the data as a binomial distribution.

As the distance from which the mononucleotide is released from the protein is increased, the probability of entry into the vestibule decreases almost linearly, this reduction in the likelihood of entry occurs at approximately the same rate regardless of phosphate orientation. The probability of capture is higher with the phosphate in the down orientation, by an average of 10%. The phosphate orientation has negligible impact on the probability of 'possible capture'.

As the CMP is displaced away from the centre of the vestibule entrance at a distance of 30 angstroms, in general, the probability of entry decreases, however we note that there is an increased probability of entry with translation of 0.5 nm. We theorize that this is due to the CMP now being released above the edge of the vestibule entrance instead of above the centre, therefore less lateral diffusion is required to interact with the protein. Here we note a higher correlation between the phosphate orientation and entry, with the phosphate down orientation demonstrating an average of 15% higher entry probability. The minimum probability of entry is generally higher by 10% with the CMP in the down orientation, and the rate of decrease in the capture probability as the translation away from the centre of the vestibule entrance increases, is lower than observed for the up orientation.

We also note that unlike OccD1 [17] or OprP [24] the orientation of the molecule, beyond release, does not have a noticeable impact on entry into the vestibule entrance or the translocation process, as the protein freely interacts with the various moieties on the substrate, e.g. the nucleobase, the hydroxyl group on the sugar and the phosphate group.

An in-house script was used with the graphics software VMD [25] to calculate the frequency and duration of Van der Waals contacts between protein residues and CMP. For simulations in which entry to the vestibule was observed, the CMP had extended interactions (an average of more than 5% over all simulations) with: D2, S3, K8, T9, D13 and N293 residues, which are located predominately in the vestibule. Whilst for simulations in which possible entry was observed, interactions were noted with residues A1, N6, K8, T9, G10, T11, D13, G15, S16, D17, T18, T19, V20, D45, K46 and N47, these residues predominately lie on the edge of the entrance to the protein. For those simulations where entry to the vestibule was not observed we noticed the majority simply didn't interact with the protein at all, or when they did, it was with residues K8, I16, N17, T18, K46 and D47. This data and the location of these residues is summarized in figure 3. Based on these observations we propose that the D45, K46 and N47 residues reduce the probability of entry to the vestibule *via* alternative favourable interactions with CMP outside the vestibule, on the surface of the cap region. The protein-nucleobase interactions that dominate here are largely hydrogen bonding of the amino acid side chains to the base and the sugar hydroxyl groups. The phosphate moiety is typically excluded from these interactions with the only exception being transient interactions with the side chain of K46 (for durations of less than 100 ps). An example interaction between CMP and the 3 residues is shown inset in figure 3.

Over the 360 simulations performed, translocation through the entire alpha-hemolysin protein was observed in only 2 simulations, as the remainder of captured CMP substrates found various binding sites in the vestibule or beta barrel.

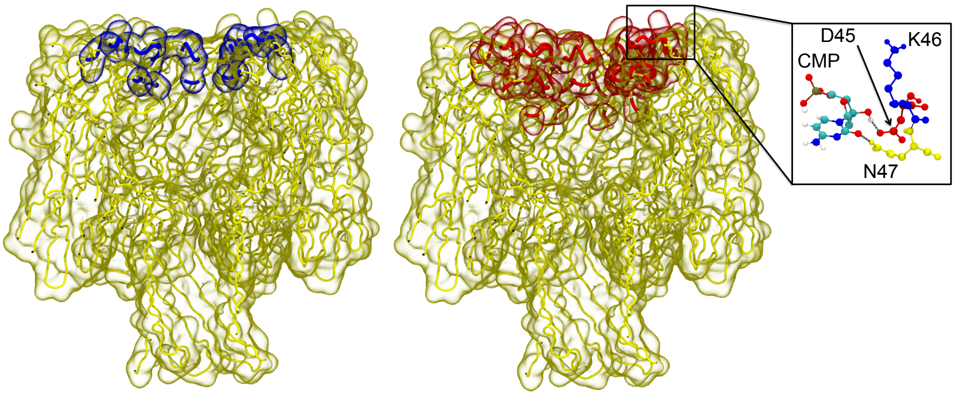


Figure 3. The residues with the highest propensity to interact with the nucleotide for: failed capture **(Left)** and possible capture, **(right)**. The blue highlighted residues corresponding to failed capture are on the edge of the vestibule entrance and on the surface of the vestibule. The red highlighted residues associated with possible capture are predominantly on the edge of the entrance to the vestibule. **(Inset)** The most frequently observed binding mode between CMP and the amino acid triplet, D45, K46 and N47. Hydrogen bonds (dashed lines) are observed between the hydroxyl group of the sugar and the side-chain of D45, as well as the nucleobase and the side-chain of N47. These hydrogen bonds are stable and are present for extended periods of time (greater than 1 ns). Residues are coloured for clarity; D45 in red K46 in blue, and N47 in yellow.

2.1. Sum effects to CMP entry into the vestibule

Root mean squares linear regression was used to calculate the equations of the lines for the relationship between distance away from the vestibule (variable labelled 'height' in equations) and probability of entry into the vestibule, shown below in equation 1, phosphate up and equation 2, phosphate down. It was observed that with the nucleotide at the position of K8, or zero distance, the probability of capture is 72 to 93%, demonstrating that nucleotides can diffuse rapidly under the conditions used, failing to remain captured even when released below residue N17.

$$Probability\ of\ possible\ capture = \ 72\% - (16.5\%.nm\ height^{-1}\ ) \tag{1}$$

$$Probability\ of\ possible\ capture = 93\% - (20.5\%.nm\ height^{-1}\ ) \tag{2}$$

The data from the simulations in which the CMP had been translated in the xy plane, had the effect of distance removed by setting the probability at zero translation equal to the predicted probability from the previous equations. From this data an equation was calculated using the previous method, which provides the total effect of nucleotide release location on the probability of entry into the vestibule, see equations 3, phosphate up, and 4, phosphate down, below.

$$Probability\ of\ possible\ capture = 72\% - (16.5\%.nm^{-1}\ height) - (2.3\%.nm\ translated^{-1}) \tag{3}$$

$$Probability\ of\ possible\ capture = 93\% - (20.5\%.nm^{-1}\ height) - (6.2\%.nm\ translated^{-1}) \tag{4}$$

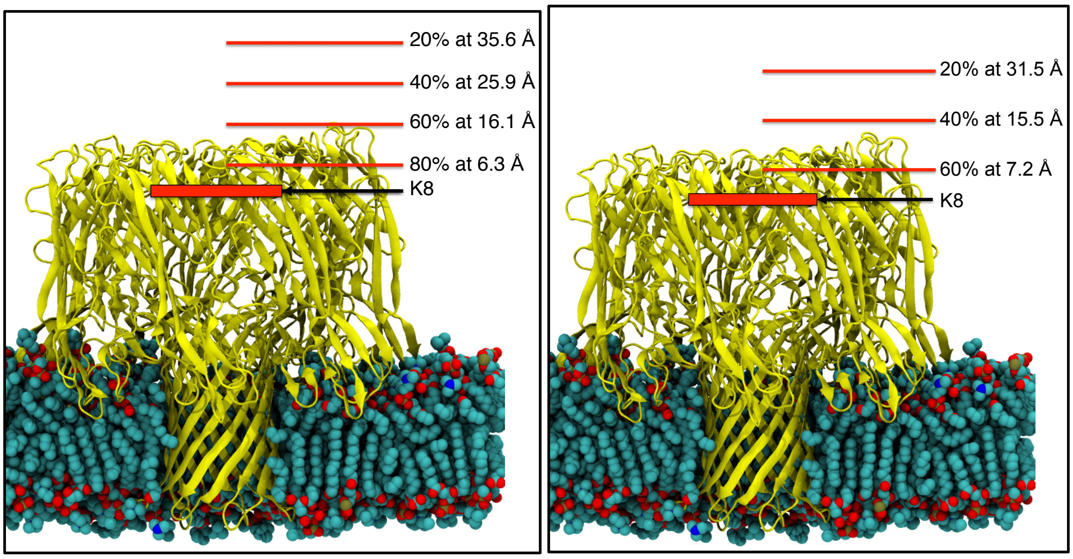


Figure 4. The probabilities of entry for given release points relative to the ring of K8 residues for **(Left)** the phosphate down orientation and **(Right)** the phosphate up orientation. Simulations performed with the phosphate in the down orientation are closer than the corresponding phosphate up simulations for the same capture probability.

4. Conclusions

In conclusion, our simulations predict that optimisation of alpha-hemolysin for nanopore sequencing, which incorporates an exonuclease enzyme for cleaving nucleotides from a strand of DNA, must consider the protein-exonuclease distance of nucleotide release. We show that for successful capture of the nucleotide by the protein, the point of nucleotide release above the protein is more important than the lateral displacement of the nucleotide with respect to the dimensions of the entrance to the protein. In other words it is more important to release the nucleotide closer to the mouth of the vestibule, than it is to ensure that it is released directly above the centre of the mouth. Furthermore, our simulations reveal that the orientation of the nucleotide is also only likely to have negligible impact on the probability of entry into the protein.

Closer inspection of the wildtype alpha hemolysin revealed that residues D45, K46 and N47 play a role in interacting with CMP in instances of failed capture, therefore we recommend mutational studies to optimise this region. Of course, it must be noted that due to the number of simulations appropriate for this study, we have been unable to vary the effects of forces on the nucleotides due to the electro-osmotic flow of ions through the pore such as have been described in [26]. For a fuller picture of nucleotide dynamics in and around nanopores, in future work we plan longer simulations to include these effects and also to study all four nucleotides.

Acknowledgments

We acknowledge the use of both University of Southampton's computing cluster IRIDIS 4 and the ARCHER national facility.

Author Contributions

S.K. conceived the experiments. S.T. performed preliminary simulations. R.M. performed the simulations. R.M and S.K analysed the simulations. R.M. and S.K. wrote the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Schneider, G.F.; Dekker, C. DNA sequencing with nanopores. *Nat Biotechnol* **2012**, *30*, 326-328.
2. Rincon-Restrepo, M.; Milthallova, E.; Bayley, H.; Maglia, G. Controlled translocation of individual DNA molecules through protein nanopores with engineered molecular brakes. *Nano Lett* **2011**, *11*, 746-750.
3. Derrington, I.M.; Butler, T.Z.; Collins, M.D.; Manrao, E.; Pavlenok, M.; Niederweis, M.; Gundlach, J.H. Nanopore DNA sequencing with mspa. *P Natl Acad Sci USA* **2010**, *107*, 16060-16065.
4. Wanunu, M. Nanopores: A journey towards DNA sequencing. *Phys Life Rev* **2012**, *9*, 125-158.
5. Dekker, C. Solid-state nanopores. *Nat Nanotechnol* **2007**, *2*, 209-215.
6. Venkatesan, B.M.; Bashir, R. Nanopore sensors for nucleic acid analysis. *Nat Nanotechnol* **2011**, *6*, 615-624.

7. Fyta, M.; Melchionna, S.; Succi, S. Translocation of biomolecules through solid-state nanopores: Theory meets experiments. *J Polym Sci Pol Phys* **2011**, *49*, 985-1011.

8. Scheicher, R.H.; Grigoriev, A.; Ahuja, R. DNA sequencing with nanopores from an ab initio perspective. *J Mater Sci* **2012**, *47*, 7439-7446.

9. Laszlo, A.H.; Derrington, I.M.; Brinkerhoff, H.; Langford, K.W.; Nova, I.C.; Samson, J.M.; Bartlett, J.J.; Pavlenok, M.; Gundlach, J.H. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore mspa. *P Natl Acad Sci USA* **2013**, *110*, 18904-18909.

10. Clarke, J.; Wu, H.C.; Jayasinghe, L.; Patel, A.; Reid, S.; Bayley, H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **2009**, *4*, 265-270.

11. Song, L.Z.; Hobaugh, M.R.; Shustak, C.; Cheley, S.; Bayley, H.; Gouaux, J.E. Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **1996**, *274*, 1859-1866.

12. Stoddart, D.; Maglia, G.; Mikhailova, E.; Heron, A.J.; Bayley, H. Multiple base-recognition sites in a biological nanopore: Two heads are better than one. *Angew Chem Int Edit* **2010**, *49*, 556-559.

13. Branton, D.; Deamer, D.W.; Marziali, A.; Bayley, H.; Benner, S.A.; Butler, T.; Di Ventra, M.; Garaj, S.; Hibbs, A.; Huang, X.H., *et al.* The potential and challenges of nanopore sequencing. *Nat Biotechnol* **2008**, *26*, 1146-1153.

14. Reiner, J.E.; Balijepalli, A.; Robertson, J.W.F.; Drown, B.S.; Burden, D.L.; Kasianowicz, J.J. The effects of diffusion on an exonuclease/nanopore-based DNA sequencing engine. *J Chem Phys* **2012**, *137*.

15. Bond, P.J.; Guy, A.T.; Heron, A.J.; Bayley, H.; Khalid, S. Molecular dynamics simulations of DNA within a nanopore: Arginine-phosphate tethering and a binding/sliding mechanism for translocation. *Biochemistry-Us* **2011**, *50*, 3777-3783.

16. Guy, A.T.; Piggot, T.J.; Khalid, S. Single-stranded DNA within nanopores: Conformational dynamics and implications for sequencing; a molecular dynamics simulation study. *Biophys J* **2012**, *103*, 1028-1036.

17. Parkin, J.; Khalid, S. Atomistic molecular-dynamics simulations enable prediction of the arginine permeation pathway through occd1/oprd from pseudomonas aeruginosa. *Biophys J* **2014**, *107*, 1853-1861.

18. Bhattacharya, S.; Derrington, I.M.; Pavlenok, M.; Niederweis, M.; Gundlach, J.H.; Aksimentiev, A. Molecular dynamics study of mspa arginine mutants predicts slow DNA translocations and ion current blockades indicative of DNA sequence. *Acs Nano* **2012**, *6*, 6960-6968.

19. Haider, S.; Neidle, S. Molecular modeling and simulation of g-quadruplexes and quadruplex-ligand complexes. *Methods in molecular biology* **2010**, *608*, 17-37.

20. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M.R.; Smith, J.C.; Kasson, P.M.; van der Spoel, D., *et al.* Gromacs 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845-854.

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

22. Oostenbrink, C.; Villa, A.; Mark, A.E.; Van Gunsteren, W.F. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6. *J Comput Chem* **2004**, *25*, 1656-1676.

23. H. J. C. Berendsen, J.P.M.P., W.F. van Gunsteren, J. Hermans. Spc. *Intermolecular Forces* **1981**, p. 331.

24. Pongprayoon, P.; Beckstein, O.; Wee, C.L.; Sansom, M.S.P. Simulations of anion transport through oprp reveal the molecular basis for high affinity and selectivity for phosphate. *P Natl Acad Sci USA* **2009**, *106*, 21614-21618.

25. Humphrey, W.; Dalke, A.; Schulten, K. Vmd: Visual molecular dynamics. *Journal of molecular graphics* **1996**, *14*, 33-38, 27-38.

26. He, Y.H.; Tsutsui, M.; Scheicher, R.H.; Fan, C.; Taniguchi, M.; Kawai, T. Mechanism of how salt-gradient-induced charges affect the translocation of DNA molecules through a nanopore. *Biophys J* **2013**, *105*, 776-782.

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

List of References

1. Mendel, G., Versuche uber pflanzen hybriden. Verhandlugen des Naturforschenden Vareines in Brünn, 1866. 3: p. 3-47.

2. Watson, J.D. and F.H. Crick, Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 1953. 171(4356): p. 737-8.

3. Sanger, F., S. Nicklen, and A.R. Coulson, DNA Sequencing with Chain-Terminating Inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 1977. 74(12): p. 5463-5467.

4. Hayes, J.J., D.J. Clark, and A.P. Wolffe, Histone Contributions to the Structure of DNA in the Nucleosome. Proceedings of the National Academy of Sciences of the United States of America, 1991. 88(15): p. 6829-6833.

5. Paulson, J.R. and U.K. Laemmli, The structure of histone-depleted metaphase chromosomes. Cell, 1977. 12(3): p. 817-28.

6. Poater, J., et al., B-DNA structure and stability: the role of hydrogen bonding, pi-pi stacking interactions, twist-angle, and solvation. Organic & Biomolecular Chemistry, 2014. 12(26): p. 4691-4700.

7. Richmond, T.J. and C.A. Davey, The structure of DNA in the nucleosome core. Nature, 2003. 423(6936): p. 145-150.

8. Egli, M., et al., Crystal-Structure of an Okazaki Fragment at 2-a Resolution. Proceedings of the National Academy of Sciences of the United States of America, 1992. 89(2): p. 534-538.

9. Rich, A. and S.G. Zhang, Z-DNA: the long road to biological function. Nature Reviews Genetics, 2003. 4(7): p. 566-572.

10. Singleton, C.K., et al., Left-Handed Z-DNA Is Induced by Supercoiling in Physiological Ionic Conditions. Nature, 1982. 299(5881): p. 312-316.

11. Narayana, N. and M.A. Weiss, Crystallographic Analysis of a Sex-Specific Enhancer Element: Sequence-Dependent DNA Structure, Hydration, and Dynamics. Journal of Molecular Biology, 2009. 385(2): p. 469-490.

12. Stroud, J. Nucleic Acid builder server. 2004 [cited 2012-2014; Available from: http://structure.usc.edu/make-na/server.html.

13. Wang, A.H.J., et al., Molecular-Structure of a Left-Handed Double Helical DNA Fragment at Atomic Resolution. Nature, 1979. 282(5740): p. 680-686.

14. Lodish H, B.A., Zipursky SL, et al., Molecular Cell Biology. . 4th edition ed. 2000.

15. Luger, K., et al., Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. Nature, 1997. 389(6648): p. 251-260.

16. Thanbichler, M. and L. Shapiro, Chromosome organization and segregation in bacteria. Journal of Structural Biology, 2006. 156(2): p. 292-303.

17. Levens, D. and C.J. Benham, DNA stress and strain, in silico, in vitro and in vivo. Physical Biology, 2011. 8(3): p. 035011.

18. Early, T.A., et al., Base-Pairing Structure in Poly-D(G-T) Double Helix - Wobble Base-Pairs. Nucleic Acids Research, 1978. 5(6): p. 1955-1970.

19. Abrescia, N.G.A., et al., Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. Proceedings of the National Academy of Sciences of the United States of America, 2002. 99(5): p. 2806-2811.

20. Lam, E.Y.N., et al., G-quadruplex structures are stable and detectable in human genomic DNA. Nature Communications, 2013. 4.

21. Storz, G., An expanding universe of noncoding RNAs. Science, 2002. 296(5571): p. 1260-1263.

22. Dykxhoorn, D.M., C.D. Novina, and P.A. Sharp, Killing the messenger: Short RNAs that silence gene expression. Nature Reviews Molecular Cell Biology, 2003. 4(6): p. 457-467.

23. Dupont, C., D.R. Armant, and C.A. Brenner, Epigenetics: Definition, Mechanisms and Clinical Perspective. Seminars in Reproductive Medicine, 2009. 27(5): p. 351-357.

24. Scherr, M., M.A. Morgan, and M. Eder, Gene silencing mediated by small interfering RNAs in mammalian cells. Current Medicinal Chemistry, 2003. 10(3): p. 245-256.

25. Fossey, A., Epigenetics: beyond genes. Southern Forests, 2009. 71(2): p. 121-124.

26. Wallace, E.V.B., et al., Identification of epigenetic DNA modifications with a protein nanopore. Chemical Communications, 2010. 46(43): p. 8195-8197.

27. Jay, E., et al., Nucleotide-Sequence Analysis of DNA .13. DNA Sequence-Analysis - a General, Simple and Rapid Method for Sequencing Large Oligodeoxyribonucleotide Fragments by Mapping. Nucleic Acids Research, 1974. 1(3): p. 331-353.

28. Min Jou, W., et al., Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature, 1972. 237(5350): p. 82-8.

29. Fiers, W., et al., Complete Nucleotide-Sequence of Bacteriophage Ms2-Rna - Primary and Secondary Structure of Replicase Gene. Nature, 1976. 260(5551): p. 500-507.

30. Maxam, A.M. and W. Gilbert, New Method for Sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America, 1977. 74(2): p. 560-564.

31. Bresler, G., M. Bresler, and D. Tse, Optimal assembly for high throughput shotgun sequencing. BMC Bioinformatics, 2013. 14 Suppl 5: p. S18.

32. Collins, F.S., et al., Finishing the euchromatic sequence of the human genome. Nature, 2004. 431(7011): p. 931-945.

33. Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001. 409(6822): p. 860-921.

34. Lockhart, D.J. and E.A. Winzeler, Genomics, gene expression and DNA arrays. Nature, 2000. 405(6788): p. 827-836.

35. Crick, F., Central Dogma of Molecular Biology. Nature, 1970. 227(5258): p. 561-&.

36. Gambetti, P., et al., Molecular biology and pathology of prion strains in sporadic human prion diseases. Acta Neuropathologica, 2011. 121(1): p. 79-90.

37. Kandel, E.R., The molecular biology of memory storage: A dialog between genes and synapses. Bioscience Reports, 2001. 21(5): p. 565-611.

38. Laszlo, A.H., et al., Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. Proceedings of the National Academy of Sciences of the United States of America, 2013. 110(47): p. 18904-18909.

39. Parker, G.A. and J.M. Smith, Optimality Theory in Evolutionary Biology. Nature, 1990. 348(6296): p. 27-33.

40. Hudson, M.E., Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources, 2008. 8(1): p. 3-17.

41. Lorenz, P. and J. Eck, Metagenomics and industrial applications. Nature Reviews Microbiology, 2005. 3(6): p. 510-516.

42. Daniel, R., The metagenomics of soil. Nature Reviews Microbiology, 2005. 3(6): p. 470-478.

43. Tringe, S.G. and E.M. Rubin, Metagenomics: DNA sequencing of environmental samples. Nature Reviews Genetics, 2005. 6(11): p. 805-814.

44. Jakupciak, J.P. and R.R. Colwell, Biological agent detection technologies. Molecular Ecology Resources, 2009. 9: p. 51-57.

45. Howorka, S., S. Cheley, and H. Bayley, Sequence-specific detection of individual DNA strands using engineered nanopores. Nature Biotechnology, 2001. 19(7): p. 636-639.

46. Kuppuswamy, M.N., et al., Single Nucleotide Primer Extension to Detect Genetic-Diseases - Experimental Application to Hemophilia-B (Factor-Ix) and Cystic-Fibrosis Genes. Proceedings of the National Academy of Sciences of the United States of America, 1991. 88(4): p. 1143-1147.

47. Pembrey, M. and A.S. Team, The Avon Longitudinal Study of Parents and Children (ALSPAC): a resource for genetic epidemiology. European Journal of Endocrinology, 2004. 151: p. U125-U129.

48. Jones, R.W., et al., A new human genetic resource: a DNA bank established as part of the Avon Longitudinal Study of Pregnancy and Childhood (ALSPAC). European Journal of Human Genetics, 2000. 8(9): p. 653-660.

49. Kedes, L. and G. Campany, The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE Competition. Nature Genetics, 2011. 43(11): p. 1055-1058.

50. Mardis, E.R., Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 2008. 9: p. 387-402.

51. McPherson, J.D., A defining decade in DNA sequencing. Nature Methods, 2014. 11(10): p. 1003-1005.

52. von Bubnoff, A., Next-generation sequencing: The race is on. Cell, 2008. 132(5): p. 721-723.

53. Bayley, H. and P.S. Cremer, Stochastic sensors inspired by biology. Nature, 2001. 413(6852): p. 226-230.

54. Movileanu, L., Interrogating single proteins through nanopores: challenges and opportunities. Trends in Biotechnology, 2009. 27(6): p. 333-341.

55. Heins, E.A., et al., Detecting single porphyrin molecules in a conically shaped synthetic nanopore. Nano Letters, 2005. 5(9): p. 1824-1829.

56. Wei, R.S., et al., Stochastic sensing of proteins with receptor-modified solid-state nanopores. Nature Nanotechnology, 2012. 7(4): p. 257-263.

57. Deamer, D.W. and M. Akeson, Nanopores and nucleic acids: prospects for ultrarapid sequencing. Trends in Biotechnology, 2000. 18(4): p. 147-151.

58. Schneider, G.F. and C. Dekker, DNA sequencing with nanopores. Nature Biotechnology, 2012. 30(4): p. 326-328.

59. Vercoutere, W. and M. Akeson, Biosensors for DNA sequence detection. Current Opinion in Chemical Biology, 2002. 6(6): p. 816-822.

60. Aksimentiev, A., et al., Microscopic kinetics of DNA translocation through synthetic nanopores. Biophysical Journal, 2004. 87(3): p. 2086-2097.

61. Wanunu, M., Nanopores: A journey towards DNA sequencing. Physics of Life Reviews, 2012. 9(2): p. 125-158.

62. Shendure, J. and H.L. Ji, Next-generation DNA sequencing. Nature Biotechnology, 2008. 26(10): p. 1135-1145.

63. Venkatesan, B.M. and R. Bashir, Nanopore sensors for nucleic acid analysis. Nature Nanotechnology, 2011. 6(10): p. 615-624.

64. Boersma, A.J. and H. Bayley, Continuous Stochastic Detection of Amino Acid Enantiomers with a Protein Nanopore. Angewandte Chemie-International Edition, 2012. 51(38): p. 9606-9609.

65. Kang, X.F., et al., Stochastic detection of enantiomers. Journal of the American Chemical Society, 2006. 128(33): p. 10684-10685.

66. Dekker, C., Solid-state nanopores. Nature Nanotechnology, 2007. 2(4): p. 209-215.

67. Sint, K., B. Wang, and P. Kral, Selective Ion Passage through Functionalized Graphene Nanopores. Journal of the American Chemical Society, 2008. 130(49): p. 16448-+.

68. Fologea, D., et al., Slowing DNA translocation in a solid-state nanopore. Nano Letters, 2005. 5(9): p. 1734-1737.

69. Venta, K., et al., Differentiation of Short, Single-Stranded DNA Homopolymers in Solid-State Nanopores. Acs Nano, 2013. 7(5): p. 4629-4636.

70. Storm, A.J., et al., Translocation of double-strand DNA through a silicon oxide nanopore. Physical Review E, 2005. 71(5).

71. Shim, J., et al., Detection and Quantification of Methylation in DNA using Solid-State Nanopores. Scientific Reports, 2013. 3.

72. Bayley, H., Nanotechnology Holes with an Edge. Nature, 2010. 467(7312): p. 164-165.

73. Wells, D.B., et al., Assessing Graphene Nanopores for Sequencing DNA. Nano Letters, 2012. 12(8): p. 4117-4123.

74. Merchant, C.A., et al., DNA Translocation through Graphene Nanopores. Nano Letters, 2010. 10(8): p. 2915-2921.

75. Drndic, M., Sequencing with graphene pores. Nature Nanotechnology, 2014. 9(10): p. 743.

76. Shankla, M. and A. Aksimentiev, Conformational transitions and stop-and-go nanopore transport of single-stranded DNA on charged graphene. Nature Communications, 2014. 5: p. 5171.

77. Farimani, A.B., K. Min, and N.R. Aluru, DNA Base Detection Using a Single-Layer MoS2. Acs Nano, 2014. 8(8): p. 7914-7922.

78. Liu, K., et al., Atomically thin molybdenum disulfide nanopores with high sensitivity for DNA translocation. ACS Nano, 2014. 8(3): p. 2504-11.

79. Eisenberg, D., The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. Proceedings of the National Academy of Sciences of the United States of America, 2003. 100(20): p. 11207-11210.

80. Song, L.Z., et al., Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science, 1996. 274(5294): p. 1859-1866.

81. Cheng, Z.H., et al., Crystal structure of Ski8p, a WD-repeat protein with dual roles in mRNA metabolism and meiotic recombination. Protein Science, 2004. 13(10): p. 2673-2684.

82. Akeson, M., et al., Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. Biophysical Journal, 1999. 77(6): p. 3227-3233.

83. Branton, D., et al., The potential and challenges of nanopore sequencing. Nature Biotechnology, 2008. 26(10): p. 1146-1153.

84. Butler, T.Z., et al., Single-molecule DNA detection with an engineered MspA protein nanopore. Proceedings of the National Academy of Sciences of the United States of America, 2008. 105(52): p. 20647-20652.

85. Manrao, E.A., et al., Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nature Biotechnology, 2012. 30(4): p. 349-U174.

86. Rincon-Restrepo, M., et al., Controlled Translocation of Individual DNA Molecules through Protein Nanopores with Engineered Molecular Brakes. Nano Letters, 2011. 11(2): p. 746-750.

87. Kasianowicz, J.J., et al., Characterization of individual polynucleotide molecules using a membrane channel. Proceedings of the National Academy of Sciences of the United States of America, 1996. 93(24): p. 13770-13773.

88. Subbarao, G.V. and B. van den Berg, Crystal structure of the monomeric porin OmpG. Journal of Molecular Biology, 2006. 360(4): p. 750-759.

89. Cowan, S.W., et al., The Structure of Ompf Porin in a Tetragonal Crystal Form. Structure, 1995. 3(10): p. 1041-1050.

90. Yildiz, O., et al., Structure of the monomeric outer-membrane porin OmpG in the open and closed conformation. EMBO Journal, 2006. 25(15): p. 3702-3713.

91. Chen, M., et al., Outer membrane protein G: Engineering a quiet pore for biosensing. Proceedings of the National Academy of Sciences of the United States of America, 2008. 105(17): p. 6272-6277.

92. Faller, M., M. Niederweis, and G.E. Schulz, The structure of a mycobacterial outer-membrane channel. Science, 2004. 303(5661): p. 1189-1192.

93. Stahl, C., et al., MspA provides the main hydrophilic pathway through the cell wall of Mycobacterium smegmatis. Molecular Microbiology, 2001. 40(2): p. 451-464.

94. Derrington, I.M., et al., Nanopore DNA sequencing with MspA. Proceedings of the National Academy of Sciences of the United States of America, 2010. 107(37): p. 16060-16065.

95. Manrao, E.A., et al., Nucleotide Discrimination with DNA Immobilized in the MspA Nanopore. Plos One, 2011. 6(10).

96. Rusk, N., Nanopores read long genomic DNA. Nature Methods, 2014. 11(9): p. 887-887.

97. Oxford Nanopore Technologies. https://www.nanoporetech.com/

2008-2014 [cited 2014 22 Nov 14].

98. Bond, P.J., et al., Molecular Dynamics Simulations of DNA within a Nanopore: Arginine-Phosphate Tethering and a Binding/Sliding Mechanism for Translocation. Biochemistry, 2011. 50(18): p. 3777-3783.

99. Cherf, G.M., et al., Automated forward and reverse ratcheting of DNA in a nanopore at 5-angstrom precision. Nature Biotechnology, 2012. 30(4): p. 344-348.

100. Berman, A.J., et al., Structures of phi29 DNA polymerase complexed with substrate: The mechanism of translocation in B-family polymerases. EMBO Journal, 2007. 26(14): p. 3494-3505.

101. Gu, L.Q., S. Cheley, and H. Bayley, Prolonged residence time of a noncovalent molecular adapter, beta-cyclodextrin, within the lumen of mutant alpha-hemolysin pores. Journal of General Physiology, 2001. 118(5): p. 481-493.

102. Banerjee, A., et al., Molecular bases of cyclodextrin adapter interactions with engineered protein nanopores. Proceedings of the National Academy of Sciences of the United States of America, 2010. 107(18): p. 8165-8170.

103. Wu, H.C., et al., Protein nanopores with covalently attached molecular adapters. Journal of the American Chemical Society, 2007. 129(51): p. 16142-16148.

104. Auffinger, P. and E. Westhof, Simulations of the molecular dynamics of nucleic acids. Current Opinion in Structural Biology, 1998. 8(2): p. 227-236.

105. Cheatham, T.E., et al., Molecular-Dynamics Simulations on Solvated Biomolecular Systems - the Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, Rna, and Proteins. Journal of the American Chemical Society, 1995. 117(14): p. 4193-4194.

106. York, D.M., et al., Toward the Accurate Modeling of DNA: The Importance of Long-Range Electrostatics. Journal of the American Chemical Society, 1995. 117(17): p. 5001-5002.

107. Swaminathan, S., G. Ravishanker, and D.L. Beveridge, Molecular dynamics of B-DNA including water and counterions: a 140-ps trajectory for d(CGCGAATTCGCG) based on the GROMOS force field. Journal of the American Chemical Society, 1991. 113(13): p. 5027-5040.

108. Ricci, C.G., et al., Molecular dynamics of DNA: comparison of force fields and terminal nucleotide definitions. Journal of Physical Chemistry B, 2010. 114(30): p. 9882-93.

109. Oostenbrink, C., et al., Validation of the 53A6 GROMOS force field. European Biophysics Journal with Biophysics Letters, 2005. 34(4): p. 273-284.

110. Cornell, W.D., et al., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). Journal of the American Chemical Society, 1996. 118(9): p. 2309-2309.

111. Perez, A., et al., Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophysical Journal, 2007. 92(11): p. 3817-29.

112. Beveridge, D.L., T.E. Cheatham, 3rd, and M. Mezei, The ABCs of molecular dynamics simulations on B-DNA, circa 2012. Journal of Biosciences, 2012. 37(3): p. 379-97.

113. Arora, N. and B. Jayaram, Energetics of base pairs in B-DNA in solution: An appraisal of potential functions and dielectric treatments. Journal of Physical Chemistry B, 1998. 102(31): p. 6139-6144.

114. Reddy, S.Y., F. Leclerc, and M. Karplus, DNA polymorphism: a comparison of force fields for nucleic acids. Biophysical Journal, 2003. 84(3): p. 1421-49.

115. MacKerell, A.D. and N.K. Banavali, All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. Journal of Computational Chemistry, 2000. 21(2): p. 105-120.

116. Dixit, S.B., et al., Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context

effects on the dynamical structures of the 10 unique dinucleotide steps. Biophysical Journal, 2005. 89(6): p. 3721-40.

117. Orozco, M., A. Noy, and A. Perez, Recent advances in the study of nucleic acid flexibility by molecular dynamics. Current Opinion Structural Biology, 2008. 18(2): p. 185-93.

118. Varnai, P. and K. Zakrzewska, DNA and its counterions: a molecular dynamics study. Nucleic Acids Research, 2004. 32(14): p. 4269-80.

119. Kundu, S., S. Mukherjee, and D. Bhattacharyya, Effect of temperature on DNA double helix: An insight from molecular dynamics simulation. Journal of Biosciences, 2012. 37(3): p. 445-55.

120. Haider, S. and S. Neidle, Molecular modeling and simulation of G-quadruplexes and quadruplex-ligand complexes. Methods in Molecular Biology, 2010. 608: p. 17-37.

121. Krepl, M., et al., Reference simulations of noncanonical nucleic acids with different chi variants of the AMBER force field: quadruplex DNA, quadruplex RNA and Z-DNA. Journal of Chemical Theory Computation, 2012. 8(7): p. 2506-2520.

122. Trovato, F. and V. Tozzini, Supercoiling and Local Denaturation of Plasmids with a Minimalist DNA Model. Journal of Physical Chemistry B, 2008. 112(42): p. 13197-13200.

123. Mitchell, J.S. and S.A. Harris, Thermodynamics of Writhe in DNA Minicircles from Molecular Dynamics Simulations. Physical Review Letters, 2013. 110(14).

124. Harris, S.A., C.A. Laughton, and T.B. Liverpool, Mapping the phase diagram of the writhe of DNA nanocircles using atomistic molecular dynamics simulations. Nucleic Acids Research, 2008. 36(1): p. 21-29.

125. Mitchell, J.S., C.A. Laughton, and S.A. Harris, Atomistic simulations reveal bubbles, kinks and wrinkles in supercoiled DNA. Nucleic Acids Research, 2011. 39(9): p. 3928-38.

126. Guy, A.T., T.J. Piggot, and S. Khalid, Single-Stranded DNA within Nanopores: Conformational Dynamics and Implications for Sequencing; a

Molecular Dynamics Simulation Study. Biophysical Journal, 2012. 103(5): p. 1028-1036.

127. MacKerell, A.D. and L. Nilsson, Molecular dynamics simulations of nucleic acid-protein complexes. Current Opinion in Structural Biology, 2008. 18(2): p. 194-199.

128. Rohs, R., et al., Origins of Specificity in Protein-DNA Recognition. Annual Review of Biochemistry, Vol 79, 2010. 79: p. 233-269.

129. Siggers, T. and R. Gordan, Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Research, 2014. 42(4): p. 2099-2111.

130. Menendez, D., et al., Changing the p53 master regulatory network: ELEMENTary, my dear Mr Watson. Oncogene, 2007. 26(15): p. 2191-201.

131. Wright, J.D. and C. Lim, Mechanism of DNA-binding loss upon single-point mutation in p53. Journal of Biosciences, 2007. 32(5): p. 827-839.

132. Heng, J.B., et al., The electromechanics of DNA in a synthetic nanopore. Biophysical Journal, 2006. 90(3): p. 1098-1106.

133. Martin H, J.S., Coveney P, Comparative analysis of nucleotide translocation through protein nanopores using steered molecular dynamics and an adaptive biasing force. Journal of Computational Chemistry, 2014. 35(9): p. 692-702.

134. Bhattacharya, S., et al., Molecular Dynamics Study of MspA Arginine Mutants Predicts Slow DNA Translocations and Ion Current Blockades Indicative of DNA Sequence. Acs Nano, 2012. 6(8): p. 6960-6968.

135. Martin, H.S.C., et al., Determination of Free Energy Profiles for the Translocation of Polynucleotides through alpha-Hemolysin Nanopores using Non-Equilibrium Molecular Dynamics Simulations. Journal of Chemical Theory and Computation, 2009. 5(8): p. 2135-2148.

136. Markosyan, S., et al., Effect of confinement on DNA, solvent and counterion dynamics in a model biological nanopore. Nanoscale, 2014. 6(15): p. 9006-9016.

137. Gilson, M.K. and H.X. Zhou, Calculation of protein-ligand binding affinities. Annual Review of Biophysics and Biomolecular Structure, 2007. 36: p. 21-42.

138. Skylaris, C.K., et al., Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. Journal of Chemical Physics, 2005. 122(8).

139. Clark, S.J., et al., First principles methods using CASTEP. Zeitschrift Fur Kristallographie, 2005. 220(5-6): p. 567-570.

140. Pronk, S., et al., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics, 2013. 29(7): p. 845-54.

141. Adcock, S.A. and J.A. McCammon, Molecular dynamics: Survey of methods for simulating the activity of proteins. Chemical Reviews, 2006. 106(5): p. 1589-1615.

142. Michel, J., R.D. Taylor, and J.W. Essex, Efficient generalized Born models for Monte Carlo simulations. Journal of Chemical Theory and Computation, 2006. 2(3): p. 732-739.

143. Jorgensen, W.L., D.S. Maxwell, and J. TiradoRives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. Journal of the American Chemical Society, 1996. 118(45): p. 11225-11236.

144. Lee, S., et al., CHARMM36 United Atom Chain Model for Lipids and Surfactants. Journal of Physical Chemistry B, 2014. 118(2): p. 547-556.

145. Lindorff-Larsen, K., et al., Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins-Structure Function and Bioinformatics, 2010. 78(8): p. 1950-1958.

146. Orsi, M. and J.W. Essex, The ELBA Force Field for Coarse-Grain Modeling of Lipid Membranes. Plos One, 2011. 6(12).

147. Shen, H., et al., An Anisotropic Coarse-Grained Model for Proteins Based On Gay-Berne and Electric Multipole Potentials. Journal of Chemical Theory and Computation, 2014. 10(2): p. 731-750.

148. Marrink, S.J., et al., The MARTINI force field: Coarse grained model for biomolecular simulations. Journal of Physical Chemistry B, 2007. 111(27): p. 7812-7824.

149. Ohkubo, Y.Z., et al., Accelerating Membrane Insertion of Peripheral Proteins with a Novel Membrane Mimetic Model. Biophysical Journal, 2012. 102(9): p. 2130-2139.

150. van der Kamp, M.W. and A.J. Mulholland, Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. Biochemistry, 2013. 52(16): p. 2708-2728.

151. Xiang, Z.X., Advances in homology protein structure modeling. Current Protein & Peptide Science, 2006. 7(3): p. 217-227.

152. Eswar, N., et al., Tools for comparative protein structure modeling and analysis. Nucleic Acids Research, 2003. 31(13): p. 3375-3380.

153. Ordog, R., Z. Szabadka, and V. Grolmusz, DECOMP: a PDB decomposition tool on the web. Bioinformation, 2009. 3(10): p. 413-4.

154. Wolf, M.G., et al., g_membed: Efficient insertion of a membrane protein into an equilibrated lipid bilayer with minimal perturbation. Journal Computational Chemistry, 2010. 31(11): p. 2169-74.

155. Reif, M.M., M. Winger, and C. Oostenbrink, Testing of the GROMOS Force-Field Parameter Set 54A8: Structural Properties of Electrolyte Solutions, Lipid Bilayers, and Proteins. Journal of Chemical Theory and Computation, 2013. 9(2): p. 1247-1264.

156. Huang, W., Z.X. Lin, and W.F. van Gunsteren, Validation of the GROMOS 54A7 Force Field with Respect to beta-Peptide Folding. Journal of Chemical Theory and Computation, 2011. 7(5): p. 1237-1243.

157. Soares, T.A., et al., Validation of the GROMOS force-field parameter set 45A3 against nuclear magnetic resonance data of hen egg lysozyme. Journal of Biomolecular Nmr, 2004. 30(4): p. 407-422.

158. Oostenbrink, C., et al., A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets

53A5 and 53A6. Journal of Computational Chemistry, 2004. 25(13): p. 1656-1676.

159. Darden, T., D. York, and L. Pedersen, Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. Journal of Chemical Physics, 1993. 98(12): p. 10089-10092.

160. Berendsen, H.J.C., et al., Molecular-Dynamics with Coupling to an External Bath. Journal of Chemical Physics, 1984. 81(8): p. 3684-3690.

161. Bussi, G., D. Donadio, and M. Parrinello, Canonical sampling through velocity rescaling. Journal of Chemical Physics, 2007. 126(1).

162. Cheng, A.L. and K.M. Merz, Application of the Nose-Hoover chain algorithm to the study of protein dynamics. Journal of Physical Chemistry, 1996. 100(5): p. 1927-1937.

163. Nose, S., A Unified Formulation of the Constant Temperature Molecular-Dynamics Methods. Journal of Chemical Physics, 1984. 81(1): p. 511-519.

164. Nosé, S. and M.L. Klein, Constant pressure molecular dynamics for molecular systems. Molecular Physics, 1983. 50(5): p. 1055-1076.

165. Parrinello, M. and A. Rahman, Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method. Journal of Applied Physics, 1981. 52(12): p. 7182-7190.

166. Patel, R.Y. and P.V. Balaji, Effect of the choice of the pressure coupling method on the spontaneous aggregation of DPPC molecules. Journal of Physical Chemistry B, 2005. 109(30): p. 14667-14674.

167. Clarke, J., et al., Continuous base identification for single-molecule nanopore DNA sequencing. Nature Nanotechnology, 2009. 4(4): p. 265-270.

168. Korada, S.K.C., et al., Crystal structures of Escherichia coli exonuclease I in complex with single-stranded DNA provide insights into the mechanism of processive digestion. Nucleic Acids Research, 2013. 41(11): p. 5887-5897.

169. Stoddart, D., et al., Multiple Base-Recognition Sites in a Biological Nanopore: Two Heads are Better than One. Angewandte Chemie-International Edition, 2010. 49(3): p. 556-559.

170. Parkin, J. and S. Khalid, Atomistic Molecular-Dynamics Simulations Enable Prediction of the Arginine Permeation Pathway through OccD1/OprD from Pseudomonas aeruginosa. Biophysical Journal, 2014. 107(8): p. 1853-1861.

171. Pongprayoon, P., et al., Simulations of anion transport through OprP reveal the molecular basis for high affinity and selectivity for phosphate. Proceedings of the National Academy of Sciences of the United States of America, 2009. 106(51): p. 21614-21618.

172. Humphrey, W., A. Dalke, and K. Schulten, VMD: visual molecular dynamics. Journal of Molecular Graphics, 1996. 14(1): p. 33-8, 27-8.

173. Eswar, N., et al., Comparative protein structure modeling using Modeller. Current Protocols in Bioinformatics, 2006. Chapter 5: p. Unit 5 6.

174. Kumar, S., et al., The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. Journal of Computational Chemistry, 1992. 13(8): p. 1011-1021.

175. Meng, F.C. and W.R. Xu, Drug permeability prediction using PMF method. Journal of Molecular Modeling, 2013. 19(3): p. 991-997.

176. Bemporad, D., J.W. Essex, and C. Luttmann, Permeation of small molecules through a lipid bilayer: A computer simulation study. Journal of Physical Chemistry B, 2004. 108(15): p. 4875-4884.

177. Bastug, T., et al., Potential of mean force calculations of ligand binding to ion channels from Jarzynski's equality and umbrella sampling. Journal of Chemical Physics, 2008. 128(15).

178. Yang, C., E. Kim, and Y. Pak, Potential of Mean Force Simulation by Pulling a DNA Aptamer in Complex with Thrombin. Bulletin of the Korean Chemical Society, 2012. 33(11): p. 3597-3600.

179. Neale, C., et al., Statistical Convergence of Equilibrium Properties in Simulations of Molecular Solutes Embedded in Lipid Bilayers. Journal of Chemical Theory and Computation, 2011. 7(12): p. 4175-4188.

180. Zhu, F.Q. and G. Hummer, Convergence and error estimation in free energy calculations using the weighted histogram analysis method. Journal of Computational Chemistry, 2012. 33(4): p. 453-465.

181. Hub, J.S., B.L. de Groot, and D. van der Spoel, g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. Journal of Chemical Theory and Computation, 2010. 6(12): p. 3713-3720.

182. Smart, O.S., et al., HOLE: a program for the analysis of the pore dimensions of ion channel structural models. Journal of Molecular Graphics, 1996. 14(6): p. 354-60, 376.

183. Stoddart, D., et al., Nucleobase Recognition in ssDNA at the Central Constriction of the alpha-Hemolysin Pore. Nano Letters, 2010. 10(9): p. 3633-3637.

184. Meller, A., et al., Rapid nanopore discrimination between single polynucleotide molecules. Proceedings of the National Academy of Sciences of the United States of America, 2000. 97(3): p. 1079-1084.

185. Murata, K., Y. Sugita, and Y. Okamoto, Free energy calculations for DNA base stacking by replica-exchange umbrella sampling. Chemical Physics Letters, 2004. 385(1-2): p. 1-7.

186. Perera, A.S., et al., Nanoscopic surfactant behavior of the porin MspA in aqueous media. Beilstein Journal of Nanotechnology, 2013. 4: p. 278-284.

187. Jayaram, B. and T. Jain, The role of water in protein-DNA recognition. Annual Review of Biophysics and Biomolecular Structure, 2004. 33: p. 343-361.

188. Kosztin, D., T.C. Bishop, and K. Schulten, Binding of the estrogen receptor to DNA. The role of waters. Biophysical Journal, 1997. 73(2): p. 557-570.

189. Homans, S.W., Water, water everywhere--except where it matters? Drug Discovery Today, 2007. 12(13-14): p. 534-9.

190. Huggins, D.J., W. Sherman, and B. Tidor, Rational approaches to improving selectivity in drug design. Journal of Medicinal Chemistry, 2012. 55(4): p. 1424-44.

191. Michel, J., J. Tirado-Rives, and W.L. Jorgensen, Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand

Optimization. Journal of the American Chemical Society, 2009. 131(42): p. 15403-15411.

192. Sanschagrin, P.C. and L.A. Kuhn, Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. Protein Science, 1998. 7(10): p. 2054-2064.

193. Huang, H.C., et al., Cluster analysis of hydration waters around the active sites of bacterial alanine racemase using a 2-ns MD simulation. Biopolymers, 2008. 89(3): p. 210-219.

194. Shirts, M.R. and V.S. Pande, Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. Journal of Chemical Physics, 2005. 122(14).

195. Mitchell, M.J. and J.A. Mccammon, Free-Energy Difference Calculations by Thermodynamic Integration - Difficulties in Obtaining a Precise Value. Journal of Computational Chemistry, 1991. 12(2): p. 271-275.

196. Bennett, C.H., Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. Journal of Computational Physics, 1976. 22(2): p. 245-268.

197. Jiao, D., et al., Simulation of Ca2+ and Mg2+ solvation using polarizable atomic multipole potential. Journal of Physical Chemistry B, 2006. 110(37): p. 18553-18559.

198. Fujitani, H., et al., Direct calculation of the binding free energies of FKBP ligands. Journal of Chemical Physics, 2005. 123(8).

199. Doyle, D.A., et al., The structure of the potassium channel: Molecular basis of K+ conduction and selectivity. Science, 1998. 280(5360): p. 69-77.

200. Maglia, G., et al., Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. Proceedings of the National Academy of Sciences of the United States of America, 2008. 105(50): p. 19720-19725.

201. Cracknell, J.A., D. Japrung, and H. Bayley, Translocating Kilobase RNA through the Staphylococcal alpha-Hemolysin Nanopore. Nano Letters, 2013. 13(6): p. 2500-2505.

202. Frankland, S.J.V., et al., Molecular simulation of the influence of chemical cross-links on the shear strength of carbon nanotube-polymer interfaces. Journal of Physical Chemistry B, 2002. 106(12): p. 3046-3048.

203. Stroud, J. make-na server. 2004, 2011 [cited 2012.]

204. Brooks, B.R., et al., Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. Journal of Computational Chemistry, 1983. 4(2): p. 187-217.

205. Kale, L., et al., NAMD2: Greater scalability for parallel molecular dynamics. Journal of Computational Physics, 1999. 151(1): p. 283-312.