UNIVERSITY OF SOUTHAMPTON

An Investigation into ssDNA Translocation Through Protein Nanopores

by

Taylor Haynes

A thesis submitted in partial fulfillment for the
degree of Master of Philosophy

in the
Faculty of Engineering and Physical Sciences
School of Chemistry

July 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
SCHOOL OF CHEMISTRY

Master of Philosophy

AN INVESTIGATION INTO ssDNA TRANSLOCATION THROUGH PROTEIN NANOPORES

Taylor Haynes

DNA sequencing is of great significance to molecular biology with myriad applications in fields such as medical research, and yet for the majority of its history the only commercially available method has been Sanger sequencing, which is still a costly and slow process despite recent improvements. Protein nanopores present an exciting direction for DNA sequencing, with the first commercial devices becoming available in 2015. Current research into sequencing with protein nanopores is mainly focussed on slowing down DNA translocation through the pore and improving the read resolution between bases.

This work focusses on investigating the suitability of protein nanopores incorporating a hydrophobic constriction region for DNA sequencing applications, with two directions of approach. An investigation into the conformational behaviour and translocation speed through the protein nanopores, and a subsequent investigation into the ability of the protein nanopores to discern between individual nucleotides during translocation.

It is shown that ssDNA translocates through these pores in an extended linear conformation, without the formation of secondary structures that can reduce the accuracy of DNA sequencing. The ability of these pores to distinguish between different nucleotides during translocation remains unclear, however the easily modifiable nature of the hydrophobic constrictions leaves myriad possible routes to improve this aspect of ssDNA sequencing.

Contents

List of Figures

List of Tables

DECLARATION OF AUTHORSHIP

I, Taylor Haynes, declare that this thesis titled *"An Investigation into ssDNA Translocation Through Protein Nanopores"* and the work presented in it are my own. I confirm that:

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

- Parts of this work have been published as:
  "Electric-Field-Driven Translocation of ssDNA Through Hydrophobic Nanopores" T. Haynes, I. P. S. Smith, E. J. Wallace, J. L. Trick, M. S. P. Sansom and S. Khalid, *ACS Nano*, 2018, **12**, 8208-8213.

Signed:

Date:

Acknowledgements

I would like to acknowledge all members of the Khalid research group throughout the course of my studies for their help and company. Particular mentions for James Graham, for his help with the programming side of things, and both Graham Saunders and Alister Boags for their discussions, often helpful and always appreciated. Prof. Syma Khalid deserves a great thanks for her advice and guidance throughout.

Regarding funding, thanks to Oxford Nanopore Technologies for sponsoring this work.

Of the many friends and family I would like to thank for their support, a special mention goes to my parents for reminding me that taking breaks is just as important as writing.

Abbreviations and Definitions

DNA: Deoxyribonucleic acid
ssDNA: Single-stranded DNA
ddNTP: 2,3-dideoxynucleotides
α-HL: Alpha-Haemolysin
MspA: *Mycobacterium smegmatis* porin A
CsgG: Curli-specific gene G protein
QM: Quantum Mechanics
MM: Molecular Mechanics
MD: Molecular Dynamics
PDB: Protein Data Bank
DPPC: 1,2-dipalmitoyl-sn-glycero-3-phosphocholine

Chapter 1

Introduction

1.1 DNA and Molecular Biology

The field of molecular biology was originally a concept rather than a defined discipline, the idea of explaining life in chemical and physical terms. In 1952 the discovery was made that DNA was the genetic material[1], followed by the characterisation of its structure as the double helix the following year[2]. These discoveries, among others, led to the establishment of the central dogma of molecular biology, which lays out the flow of genetic information in a biological system. Simply put; genetic information flows from nucleic acids to proteins, and cannot be transferred back from a protein. DNA is central to the field of molecular biology as the origin of biological molecules, and as such it is of immense importance.

1.2 DNA Sequencing

The sequencing of the first full DNA genome, that of bacteriophage ϕX174, used a chain-termination technique more commonly known as Sanger sequencing[3]. Sanger sequencing was used to successfully sequence a full human genome in 2007[4]. Chain termination sequencing uses a DNA polymerase enzyme to replicate a DNA sequence in the presence of unmodified nucleotides and labelled 'dideoxynucleotides' (ddNTPs) that are missing the 3' hydroxyl group. Whenever a ddNTP is incorporated into a DNA strand, that strand is terminated. After sufficient amplification, a ddNTP will have been incorporated at every position of the sequence, which can then be separated by strand length using electrophoresis and thus the DNA sequence obtained. It is worth mentioning that Sanger sequencing and its derivatives were the most widely used DNA sequencing techniques for over 30 years, during which the improvements in sequencing technology progressed at a rate unseen in any other technology, including that of

computer chips[5]. Current generation DNA sequencing still uses methods based on Sanger-sequencing, with the market leader Illumina using a reversible chain termination technique with fluorescent markers to sequence genomes in a rapid fashion. This technique, just like Sanger sequencing, is a sequence by synthesis method and therefore requires time consuming amplification steps before sequencing can be carried out.

1.3 Nanopore sequencing

Nanopores are nanometer-scale apertures in a surface. They occur naturally in some membrane proteins, and have been produced synthetically using various methods[6]. Unlike most of the sequencing methods in use today, nanopore sequencing does not require any expensive DNA amplification steps and as such has a theoretically lower cost.

The principle of nanopore sequencing lies in the separation of two chambers filled with an ionic liquid by a nanopore. Applying a potential difference across the barrier causes the flow of ionic current, the magnitude of which is proportional to the size of the aperture, or nanopore. When a species of interest passes through the pore the observed current is reduced, as shown in figure 1.1. This process is sensitive enough to distinguish between different DNA bases as they pass through the pore[7, 8]. Due to the nature of the process only charged substrates are able to be used and thus DNA, with it's negatively charged backbone, is an attractive target.

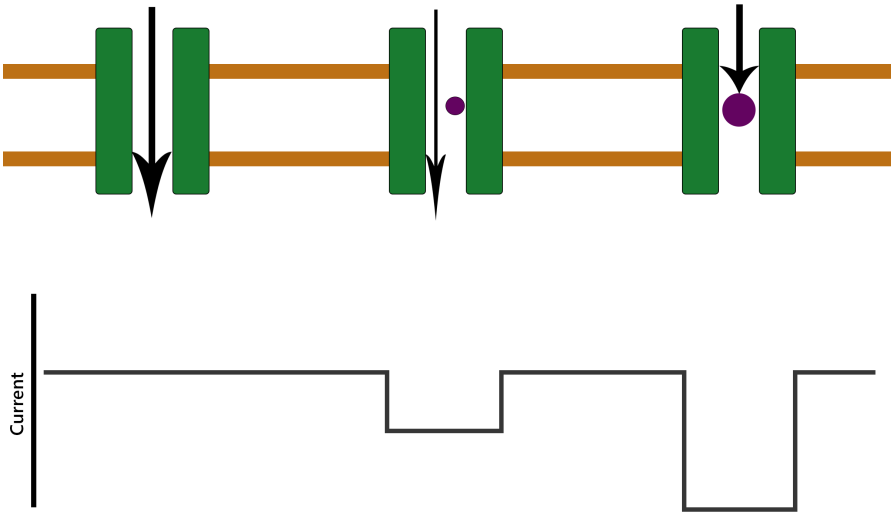


FIGURE 1.1: A demonstration of the principles behind nanopore sequencing. Each pore state (open, partially occluded and fully occluded from left to right) corresponds to the current change depicted beneath.

Nanopores being investigated for use in sequencing technology can be separated into two general categories, solid-state and biological.

1.3.1 Solid-state nanopores

Solid state nanopores are simple holes created in a substrate. Early solid-state pores used ion beam sculpting, in which an initial hole is bored using an ion beam, and then closed to the desired width by stimulating the movement of surface atoms[6]. Another technique used to create solid-state nanopores of narrow diameter is cold ion beam sculpting[9], in which surface diffusivity is reduced to the point that pores open rather than close under the effect of an ion beam. A wide variety of substrates have been used for the creation of nanopores, notably silicon based constructs, aluminium oxide surfaces[10] and graphene[11] have all been investigated. One of the advantages that solid state nanopores have over biological nanopores is the ability to combine substrates in order to combine their properties. As an example, single atom thick graphene nanopores tend to show significant noise levels in their ionic current, which have been partially mitigated by layering over a thicker silicon nitride support with a larger aperture, so the detection is still carried out by the graphene nanopore[11]

The advantages of solid state pores are that they are very stable and are able to have finely controlled nanopore dimensions, but they lack the flexibility of protein nanopores when it comes to functionalisation. Solid-state nanopores show great promise for use in future sequencing technologies but are not yet mature enough for use in the present, lacking the vital capability of discrimination between single nucleotides as they pass through the nanopore.

1.3.2 Protein nanopores

Biological nanopores are commonly found throughout nature in the form of membrane proteins. Whilst more sensitive to environmental factors such as temperature and pH than solid state nanopores, some proteins have nevertheless been found to be stable enough to find use in nanopore sequencing applications. The advantages that protein nanopores offer are their prior existence and relatively easy adaptability through mutation or direct chemical modification.

The protein nanopore sequencing process can be broken down into three distinct regions, shown in figure 1.2, each of which is a prime target for optimisation:

1. The capture of the DNA from solution
2. The transmission of DNA through the vestibule to the pore entrance
3. The translocation of DNA through the sequencing region of the pore

In this work, I will be focussing on the translocation of DNA through the sequencing region of the pore.

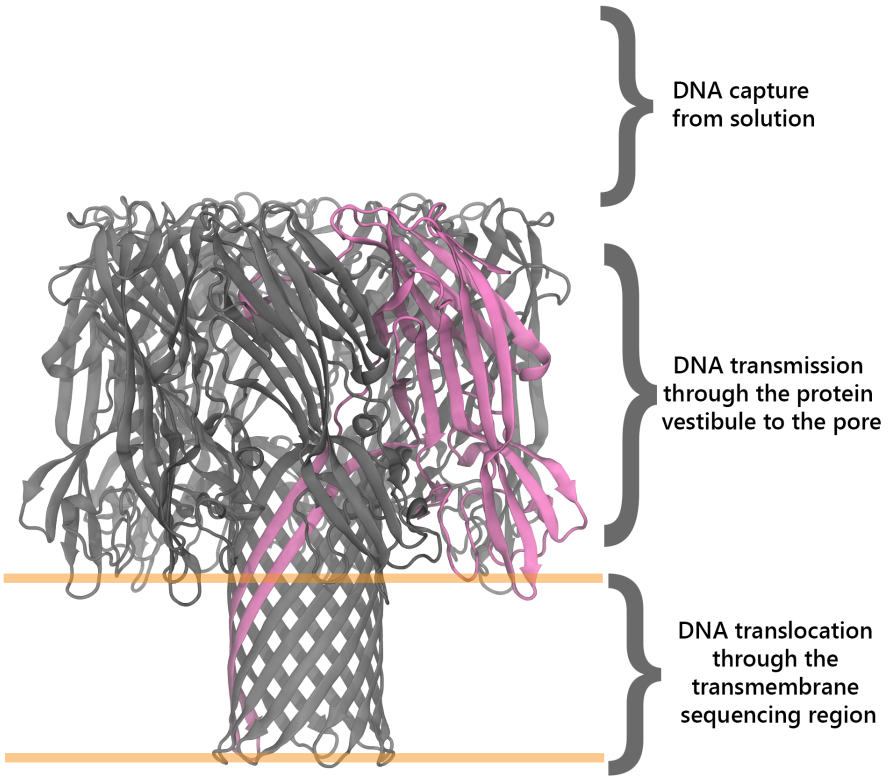


FIGURE 1.2: A breakdown of the regions for potential optimisation of nanopores for DNA sequencing.

Specific proteins that have been subject to much scrutiny for use in DNA sequencing include α-Hemolysin, MspA and CsgG.

1.3.2.1 α-Hemolysin

α-HL was the first reported nucleotide detector in 1996[13], following which it has been the subject of extensive study as both a DNA sequencing pore[14] and a detector for other small molecules[15]. α-HL itself is a mushroom-shaped heptameric protein complex from *S. aureus* of approximately 240 kDa. The protein can be divided into a 5nm transmembrane beta-barrel with varying diameter that reaches 1.4nm at its narrowest point, and a large capping vestibule. When occurring naturally α-HL is excreted as a pore-forming toxin[16].

1.3.2.2 MspA

Mycobacterium smegmatis main porin A (MspA) is a goblet-shaped porin that has been found to have similar stability and suitability for DNA sequencing to α-HL[18], with the notable advantage of having only a single point of constriction instead of a long transmembrane region with multiple constriction points. The transmembrane section

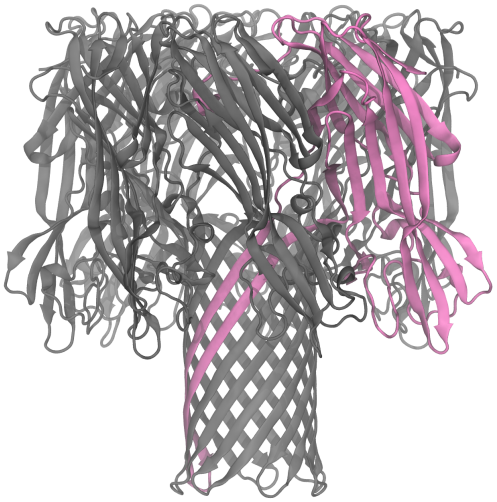


FIGURE 1.3: The α-hemolysin nanopore with the monomer shown in pink (Taken from PDB 3ANZ [12])

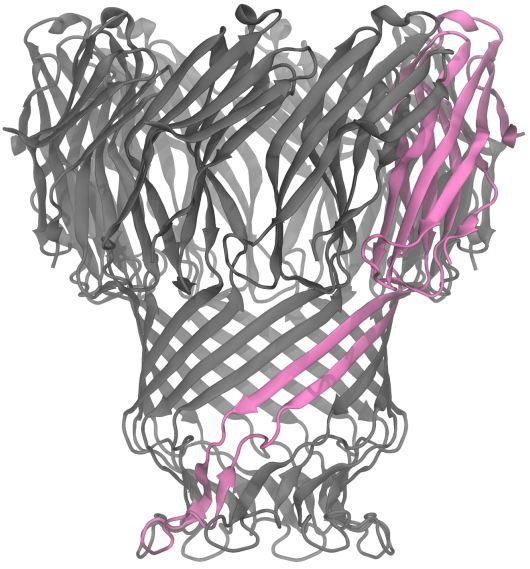
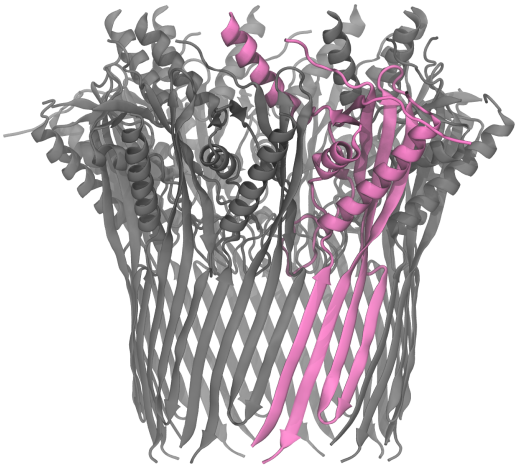


FIGURE 1.4: The *Mycobacterium smegmatis* main porin A with the monomer shown in pink (Taken from PDB 1UUN [17])

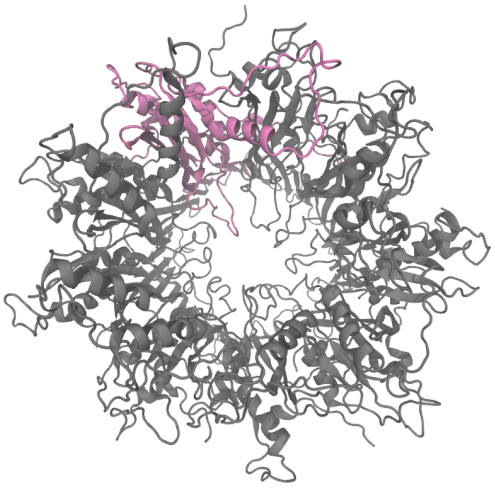
of MspA consists of two β barrels separated by a proline residue, the lower of which provides the constriction point of diameter 1.2nm and length 2nm. The dimensions of the constriction point are significant as they are the limiting site in ionic current and therefore in DNA sequencing.

1.3.2.3 CsgG

CsgG (Curli specific gene G) is a 262 residue lipoprotein present in the outer membrane of *Escheria coli*[20] that has recently come to light as being the active pore used in the



(a) A side view of the CsgG pore



(b) A top-down view of the CsgG pore, showing the constriction region at the base of the vestibule

FIGURE 1.5: The Curli specific gene G (CsgG) nanopore with the monomer shown in pink (Taken from PDB 3X2R [19])

commercial nanopore sequencing devices produced by Oxford Nanopore Technologies. The transmembrane region of CsgG is a β barrel of 4nm diameter that is linked to the periplasmic vestibule by a consticton site of 9Å diameter.

1.3.3 Oxford Nanopore Technologies

Oxford nanopore technologies is a company founded in 2005 and is the industry leader in single molecule, long read nanopore sequencing. They are the industrial sponsors of this work. Several sequencing platforms are in active development; MinION is a currently commercially available portable sequencing device, the larger PromethION platform is a benchtop device with a much greater potential throughput than minION and for which an early access program was launched in July 2015. The recently announced SmidgION platform is a mobile phone based sequencer for which details are scarce but which promises unprecedented portability in sequencing technology.

MinION is a small, portable sequencing device that has found use in various situations despite only having been commercially available since 2015. Applications to which MinION has been successfully put include sequencing of viral genomes[21], scaffolding of bacterial genomes[22], identifying the position of bacterial antibiotic markers[23] and detection of pathogens directly from clinical samples[24].

Chapter 2

Methods

2.1 Computational Chemistry

Computational chemistry is a branch of Chemistry which can be broadly divided into two applications: Firstly, it seeks to validate experimentally observed properties and provide insight into the molecular origins of said properties, and secondly can be used to predict the outcomes of as-yet unperformed wet experiments. Computational chemistry is a powerful tool for investigating interactions on the nanometer and nanosecond scales, observations which aren't possible to make in experimental systems. One area in which computational chemistry can be used to great effect is in the study of biological systems, where individual atoms or amino acid residues can't be studied experimentally, but could be responsible for the formation of a stable protein-ligand complex.

2.2 QM and MM

The intuitive way to simulate chemical reactions is to work from the ground up, modelling the nucleus and electrons of individual atoms to realistically represent a chemical system. This approach leads to a quantum mechanical (QM) representation of a system, and allows investigation of aspects such as bond formation and breakage. The downside of this approach is that it requires approximating the Schrödinger equation for each atomic centre in the system. This level of calculation makes QM simulation impractical for more than a few atoms or small molecules on a short timescale. These simulation methods are not always mutually exclusive, several examples exist where QM/MM hybrid systems have been constructed to allow study of electronic properties in specific regions of particular interest within larger systems, such as those found in ligand binding interactions[25].

To simulate larger systems for longer timescales, as is necessary for studying biological systems, accuracy must be sacrificed in order to increase computational efficiency. To this end, classical mechanics are used to describe atomic systems rather than quantum mechanics. In classical mechanical simulations, atoms are treated as spheres with a partial charge dependent on their electrostatic properties and bonds and bond angles are represented as harmonic potentials. This approximation is known as molecular mechanics (MM). The MM approach allows simulation of much larger systems at the cost of explicit electron modelling, which excludes study of any electronic movement such as bond formation or charge transfer. As a result of this, MM simulations are widely used to simulate large systems in which non-bonding interactions are dominant, such as protein folding and cell membranes.

Two main sampling methods are used for both QM and MM, Monte Carlo and molecular dynamics (MD). Monte Carlo simulations generate a number of structures based on random sampling. A trial of one of these structures (a "move") is randomly generated and compared to a selection criteria, on which it is either accepted or rejected. Monte Carlo simulations are time independent and cannot be used to track dynamic processes. MD applies Newton's laws of motion to calculate new positions for each particle based on position and starting velocities, and can therefore be used to model time-dependent processes.

2.2.1 Molecular Dynamics

Molecular dynamics relies on classical mechanics to update the positions of atoms after each timestep, which requires the force acting on the atom to be calculated from the derivative of the potential energy, i.e.

$$\vec{\mathbf{F}} = -\frac{d\vec{\mathbf{U}}}{dr} \tag{2.1}$$

Where $\vec{\mathbf{F}}$ is the force on the atom, $d\vec{\mathbf{U}}$ is the derivative of the potential energy and r is the derivative of the distance. The timestep for which the positions are updated is chosen individually for each system, but must not allow enough movement that the atomic radii overlap on the following timestep. If this does happen, the large forces created due to steric repulsion cause greater movement which cascades to eventually cause the simulation to fail.

With MD simulations, further approximations can be made to increase the potential system size and simulation time if necessary. This is accomplished by reducing the number of particles simulated to a greater degree.

The most straightforward MM model is at fully atomistic resolution, which models each atom individually as described previously. This is the most computationally expensive

MM model to simulate. A common method taken to increase the speed of these simulation is to use the united atom model, which combines non-polar hydrogen atoms with their parent atom, e.g. each CH_2 group in a hydrocarbon chain would be reduced from three particles to one. For systems with long carbon chains, such as lipids, this model provides a large increase in efficiency for relatively little loss of information[26].

Further merging of atoms into single particles results in coarse-grained (CG) models, in which several heavy atoms are grouped together in a single particle or 'bead'. Each bead carries the properties of its constituent atoms such as charge and polarity. A common CG model used in protein and membrane simulations is the Martini CG model[27], which uses beads containing 4 atoms. CG models suffer from more drawbacks than the united atom model, notably that the movement of individual atoms is lost. Similar to QM/MM hybrid simulations, these different resolutions can be merged to provide more detail in the specific area of study.

2.3 Forcefields

Forcefields are collections of the atomic information that is required to calculate the potential energy of the system, and thus the forces acting on them as discussed earlier, and the equations that describe the interactions between them. There is a wide variety of forcefields available, each of which is generally designed with a specific use in mind. For biological molecules the most commonly used force field families are AMBER[28], CHARMM[29] and GROMOS[30].

All forcefields differ in their specific form, but essentially they all calculate two aspects of the potential energy: the bonded and non-bonded terms. The bonded term includes bond stretching, bond angle and dihedral (torsional) elements. The nonbonded term encompasses electrostatic and van der Waals type interactions.

$$U_{pot} = \Sigma_{Bonded} + \Sigma_{Non-Bonded} \tag{2.2}$$

2.3.1 Bonded term

The bond stretching element describes the oscillation of a bond between two atoms. This is represented as a harmonic potential;

$$PE = \frac{1}{2}k(b - b_0)^2 \tag{2.3}$$

where k is the spring constant, b is the bond length and b_0 is the equilibrium bond length.

Bond angle bending requires 3 atomic centres and represents the changing angle between two bonds. This is also represented as a harmonic potential around an equilibrium angle.

The torsional terms measure properties across groups of 4 atoms, and are split into two categories; proper and improper dihedrals. Proper dihedrals arise from 4 atoms rotating around a bond, and due to the possibility of 360 degree rotation are modelled as a cosine potential rather than the harmonic potential used for the bond stretching and bond angle terms. This can be illustrated by the function

$$PE = k[1 + cos(n\theta - \theta_0)] \tag{2.4}$$

Improper dihedrals fulfill a different role to proper dihedrals. Instead of allowing full rotation about a bond, improper dihedrals restrict rotation in order to keep 4 atoms in the same relative configuration. The main uses of this are in planar molecules such as aromatic rings and on tetrahedral centres. The improper dihedral function is a return to the harmonic potential seen previously.

2.3.2 Nonbonded term

Nonbonded interactions are, as indicated by the name, interactions between two unconnected atoms. This term covers two seperate forces, electrostatics and Lennard-Jones interactions. The electrostatic interactions of an atom can be modelled by the Coulomb potential:

$$V = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r} \tag{2.5}$$

where ε_0 is the permittivity of free space, q_1 and q_2 are the charges on each atom and r is the distance between them. Coulomb's potential can be either attractive or repulsive dependant on the signs of the charges.

The Lennard-Jones potential has the form

$$V = 4\varepsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}] \tag{2.6}$$

in which ε and σ are specific to the particles involved in the interaction;ε is the minimum potential and σ is the distance of zero potential. This potential provides an approximation of both short-range repulsive forces (otherwise known as Pauli repulsion) and longer-range attractive van der Waals forces

2.4 Integrator

An integrator is the function responsible for updating the position of each atom at each timestep by solving the equations of motion. MD integrators typically calculate the acceleration of a particle at time t in order to update the position of the particle at time $t+dt$, with dt being the timestep. The default integrator used in the GROMACS software package is the leapfrog integrator[31], so named because it calculates the velocities at the half-timestep and the positions at the whole timestep, resulting in the calculations 'leap-frogging' one another. The Velocity Verlet algorithm [32] is also able to be used in GROMACS, although is not yet fully integrated with all options.

2.5 Periodic Boundary conditions

Simulations are limited in size and time by hardware restrictions and as such even in the largest MD simulations, which can number into the millions of atoms, boundary conditions of the system become a factor. Periodic boundary conditions avoid the problems associated with traditional boundaries by treating the simulation system as a periodic image. That is to say, each side of the simulation box is treated as being its own opposite, so an atom leaving the box simply re-enters on the opposing side.

One use that periodic boundary conditions have been put to is in the creation of 'infinite' surfaces such as graphene[33] and in polymer extension[34].

Using periodic boundary conditions necessitates consideration of force cut offs; one image of a particle must not 'see' the copy of itself in an adjacent unit cell, resulting in an infinite number of long-range interaction calculations to perform. Fortunately, the potential of long-range interactions tend to 0 at infinite distance so an approximation providing a cut off point is a reasonable solution. Simply ignoring any interactions beyond a certain distance isn't desirable, especially for electrostatic interactions, as it can lead to significant errors over time around the cut-off distance.

Two solutions to this problem exist in the shifted or switched potentials, the former of which modifies the entire length of the function so that the potential is 0 at the desired cut-off distance and the latter introduces a second threshold before the cut-off distance, between which the potential decays linearly to 0. This is sufficient for the Lennard-Jones potentials after applying dispersion correction terms to account for the longer-range interactions, but the electrostatic interactions can still be meaningful past the cut-off distances.

For long-range electrostatic calculations, the most common method is the particle mesh Ewald method [35], which uses a lattice approach to approximate the potential beyond the cut off point.

2.6 Ensembles

An ensemble is a collection of microstates from which the macroscopic properties can be derived. A collection of all possible microstates is the macrostate, so in order to estimate the properties of the macrostate a sufficiently large number of microstates need to be analysed. With respect to MD, an ensemble is the thermodynamic system within which a simulation is run. Three commonly used ensembles are the microcanonical, canonical and isothermal-isobaric ensembles, also referred to as NVE, NVT and NPT respectively. N is the number of particles, V is the system volume, E is the total energy, T is the temperature and P is the pressure.

The NVT and NPT ensembles are often used prior to running a production MD simulation in order to ensure that the temperature and pressure of the system are stable, referred to as equilibrating the system.

In this work, all systems were equilibrated using the NVT and NPT ensembles sequentially in order to bring the system to a stable state prior to full simulation runs.

2.6.1 Thermodynamic controls

In order to use ensembles one is required to keep certain variables constant, with both NVT and NPT ensembles requiring a constant temperature and NPT requiring a constant pressure as well.

Temperature at the macroscale is a property arising from the average kinetic energy of a group of particles, which arises from their velocity. At the start of a simulation these velocities are assigned according to a Boltzmann distribution around the desired temperature, but without some sort of control this distribution will drift over time. The control required comes in the form of a thermostat, which modifies particle velocities within the system. Two commonly used thermostats are the Berendsen[36] and Nose-Hoover[37, 38], which both treat the system as being attached to an external thermal reservoir.

The Berendsen thermostat relaxes a system toward a target temperature in an exponential fashion, a useful property during system equilibration, but unfortunately does not distribute the velocities correctly across all molecules. The velocity-rescaling thermostat[39] is a modification to the Berendsen thermostat that adds an additional term to allow fluctuations in temperature, resulting in a correct modelling of the NVT ensemble. the Nose-Hoover thermostat is more suited to systems that are already equilibrated to a stable temperature, as it is gives a correct energy distribution after an equilibrium has been attained.

Pressure is controlled in a similar manner using a barostat, which adjusts the box size whilst connected to an external pressure reservoir. The Berendsen barostat[36], in a similar manner to the thermostat, relaxes a system quickly to the required pressure but does not generate accurate fluctuations and as such is not desirable to use for production runs. The Parrinello-Rahman barostat[40] is a more complex barostat that alters the simulation box vectors, and is suited for use on pre-equilibrated systems as it generates a correct distribution of fluctuations.

In this work, the Berendsen thermo- and barostats were used during initial system equilibration simulations, but were unsuitable for use during production runs as explained previously. The chosen thermostat and barostat for full production simulations were the velocity-rescaling thermostat, due to its aforementioned correct distribution of kinetic energies, and the Parrinello-Rahman barostat to yield the correct isothermal-isobaric ensemble in a pre-equilibrated system. This combination of thermostat and barostat has the added advantage of maintaining consistency with previous works in the area[41, 42].

Chapter 3

Single-Stranded DNA Translocation Through Hydrophobic Nanopores

3.1 Abstract

Accurately sequencing DNA using protein nanopores requires fine control over both the speed at which DNA passes through the pore and the secondary structure of the DNA strand during translocation. This work shows that hourglass-shaped protein nanopores incorporating a constriction of hydrophobic amino acid residues allow ssDNA translocation under application of an electric field. This translocation proceeds in an extended, linear conformation with little secondary structure formation within the constriction region. In addition to this, the hydrophobic constriction provides a sufficient barrier to translocation to slow the translocation of ssDNA compared to control pores.

3.2 Notes

The hydrophobic pore models used in this research were developed from model hydrophobic nanopores created by Dr Jemma Trick.

This chapter has been submitted to the journal ACS Nano under the title 'Electric-Field-Driven Translocation of ssDNA Through Hydrophobic Nanopores'.

3.3 Introduction

Hydrophobic gating is a proposed mechanism by which a pore protein can be functionally closed by having a narrow region lined with hydrophobic moieties provide a barrier to water and ions whilst not being sterically occluded. This mechanism has been suggested to exist in several naturally occurring channel proteins[43–45], as well as having been explored in a number of studies both experimental[46, 47] and computational[48].

Previous work has been carried out using simulations to investigate the viability of designing a hydrophobic barrier into β-barrel protein nanopore templates[49]. It was shown that incorporating a hydrophobic constriction region into minimalistic nanopore templates was possible whilst retaining pore stability, and that this was sufficient to create a hydrophobic gating mechanism capable of stochastically wetting and dewetting the nanopore interior.

ssDNA translocation through protein nanopores has been well studied both experimentally[50, 51] and computationally[41, 52]. Important factors affecting the quality of sequencing data obtained from passing ssDNA through a nanopore include the rate of translocation and the conformation of the ssDNA strand during translocation.

In this work the suitability of these simple hydrophobic-gated nanopores for the sequencing of ssDNA is investigated. The behaviour and stability of the nanopore models under the application of a sustained electric field is examined, and both the barriers to ssDNA translocation through a hydrophobic gate and the effect of these structures on the conformational behaviour of a ssDNA strand within a hydrophobic gate are studied.

3.4 Methods

3.4.1 Hydrophobic Pore Models

The hydrophobic nanopore models used in this work were based on those previously reported by Trick et al[49]. These nanopores were based on a 14-stranded transmembrane β-barrel, with mutations introduced to create the hydrophobic constriction regions within the pores. Modifications made (e.g. anchoring tryptophan residues to tyrosine) were performed using PyMOL[53]. The pores studied here have symmetrical linings of sequence GAVLVAG and GAGGGAG, where the bulky VLV region creates the hydrophobic constriction under investigation. A secondary mutation was made in which the bulky tryptophan residues used to anchor the pore in the membrane were mutated to tyrosine, to explore the impact of the anchoring residues to the overall function of the pore.

3.4.2 DNA models

The ssDNA model used was a 16-nucleotide strand of sequence 5'-ATATCGCGATATCGCG-3', generated using the 3DNA package[54, 55]. This sequence was chosen to allow for hairpin bend secondary structure formation if possible, as has been observed previously in ssDNA translocation[41] and identified as a potential inhibitor of accurate sequencing. The parameters for modelling the ssDNA were drawn from the gromos53a6 forcefield, as this has previously been shown to provide good conformational flexibility on the timescales required[42]. gromos53a6 does not include parameters for terminal DNA residues, so the modified terminal RNA parameters described by Guy et al.[42] were used. This has the added advantage of maintaining consistency with prior ssDNA translocation studies.

3.4.3 System Setup

The pore of interest was positioned centrally in bilayers made up of 430 1,2-dipalmitoyl-sn-glycero-3-phosphocholine (DPPC) lipids, and solvated using the SPC water model[56] and ions added to a neutralising concentration of 1.0 M. During simulation the phosphates of the lipid head groups were restrained using a 1000 $\mathrm{kJ\,nm^{-1}}$ harmonic potential in order to keep the pore in place and avoid electroporation of the membrane in all simulations except during initial equilibration of the freshly inserted protein.

In simulations in which ssDNA was incorporated, the ssDNA strand was positioned with the 5' terminus pointing in the direction of translocation and either (1) with the terminal nucleotide positioned at the pore mouth (determined using the C-αs of the anchoring TRP/TYR residues) or (2) with the ssDNA strand threaded through the pore such that the terminal nucleotide was past the three rings of residues comprising the pore constriction region..

Systems were equilibrated in the NPT ensemble for 2 ns prior to production run simulation, with a minimum of three repeats per system. ssDNA was restrained using a 1000 $\mathrm{kJ\,nm^{-1}}$ harmonic potential applied to the phosphate groups during equilibration. All simulations were run using the Parinnello-Rahman barostat[40] and the v-rescale thermostat[39] to maintain the pressure and temperature at 1 bar and 310 K respectively.

3.4.4 Simulation Protocols

All simulations were performed using the GROMACS software package[57–62], versions 5.1.4, 5.1.5 and 2016 corresponding to the release of package updates. The gromos53a6 forcefield[11, 63] with Berger lipid definitions[64] and previously described DNA modifications was used. The SPC water model was used for system solvation. Long range

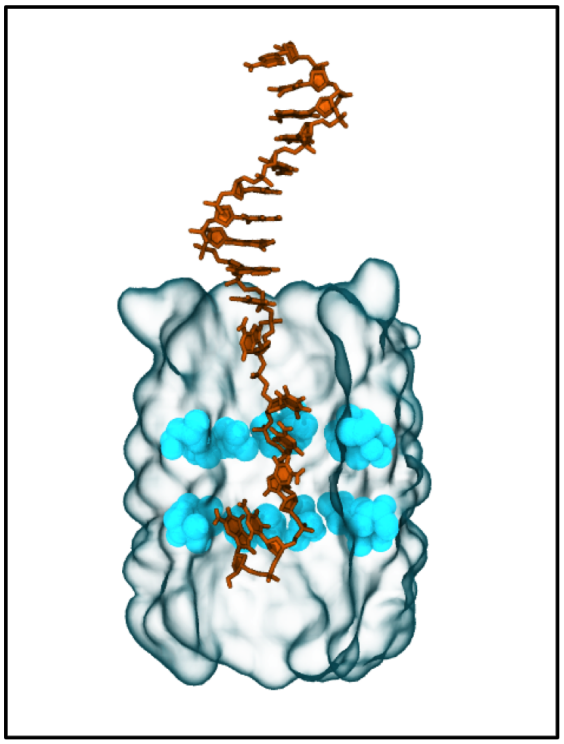


FIGURE 3.1: *Starting position of threaded ssDNA; DNA is orange, valine residues are cyan.*

TABLE 3.1: A summary of all simulations performed. 72 simulations with a cumulative duration of 1.7 µs were carried out.

Pore	**Field Strength** ($V\,nm^{-1}$)	**Initial DNA Position**		
		No DNA	Pore Mouth	Pore Constriction
GAGGGAG	0	-	-	-
	0.1	-	-	3x40 ns
	0.2	3x20 ns	6x30 ns	6x20 ns
GAVLVAG-TRP	0	1x20 ns	1x20 ns	3x20 ns
	0.1	1x20 ns	3x50 ns	3x20 ns
	0.2	1x20 ns	11x20 ns	3x20 ns
GAVLVAG-TYR	0	3x20 ns	2x20 ns	6x20 ns
	0.1	1x20 ns	3x30 ns	-
	0.2	1x20 ns	6x20 ns	6x20 ns

electrostatic interactions were treated using the Particle Mesh Ewald method[65] with a short range cutoff of 1.4 nm. Van der Waals interactions were also truncated at 1.4 nm, with long range dispersion corrections being added to the energy and pressure. During simulation the system temperature was maintained at 310 K using the v-rescale thermostat[39] and a coupling constant of 0.1 ps, and pressure was maintained semi-isotropically at 1 bar using the Parrinello-Rahman barostat[40] and a time constant of

1 ps. A timestep of 2 fs was enabled by using the LINCS algorithm[66] to contrain all bonds. Electric fields were applied as a constant voltage drop across the simulation box.

3.4.5 Analysis

Analysis was carried out using GROMACS utilities and bespoke scripts using both R and Python code to ascertain ssDNA translocation rates and end-to-end distances as a measure of conformational changes. Pore profiles were obtained using HOLE[67]. Water flux counts were obtained by defining xy planes at the pore entrances and exits, and counting a flux event only once a water molecule had crossed both planes, as was previously reported by Trick et al.[49]

3.5 Results

3.5.1 Validation of Nanopore Models

The model nanopores previously described were initially simulated in the absence of DNA in order to establish the effect of the hydrophobic constriction (VLV - pore diameter of 0.4 nm) in relation to an unconstricted analogous pore (GGG - pore diameter of 0.8 nm). As such, this second pore (lining sequence GAGGGAG) will be referred to as the reference pore. A constant electric field of 0.2 $\mathrm{V\,nm^{-1}}$ was applied to the simulation box, equivalent to approximately a 600 mV potential difference across the membrane. Under this field strength, the pore structure fluctuated slightly but not to a significant degree, showing that the pore models are stable and able to withstand an applied field of this strength (Figure 3.2).

As was expected from the prior work[49], the application of an electric field induced pore wetting. The mean bi-directional water flux through the reference pore was 69 ±23 $\mathrm{ns^{-1}}$. The mean bi-directional water flux through the TRP- and TYR- anchored pores was 61 ±16 $\mathrm{ns^{-1}}$ and 61 ±23 $\mathrm{ns^{-1}}$ respectively (Figure 3.3). Due to the similarity of the water flux and later ssDNA behaviour, results from these two pores will be combined and referred to as the *hourglass pore*.

3.5.2 DNA translocation

For describing the outcomes of these simulations, the *constriction region* of the hourglass pores refers to the three stacked rings of valine-leucine-valine residues positioned centrally in the pore. The *pore mouth* is defined by the z-position of the C-α atoms of the anchoring aromatic residues on the side of the membrane where the ssDNA strand is initially positioned, with the *pore exit* being the equivalent on the opposite side of

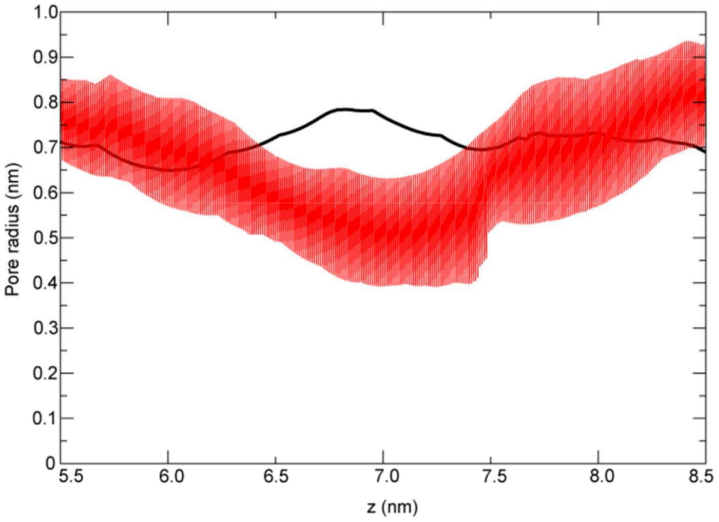


FIGURE 3.2: *Pore radius profile over a 20* ns *simulation under a 0.2* $\mathrm{V\,nm^{-1}}$ *field strength (red) of the HG pore in the absence of ssDNA, and of a single snapshot of the reference pore (black)*

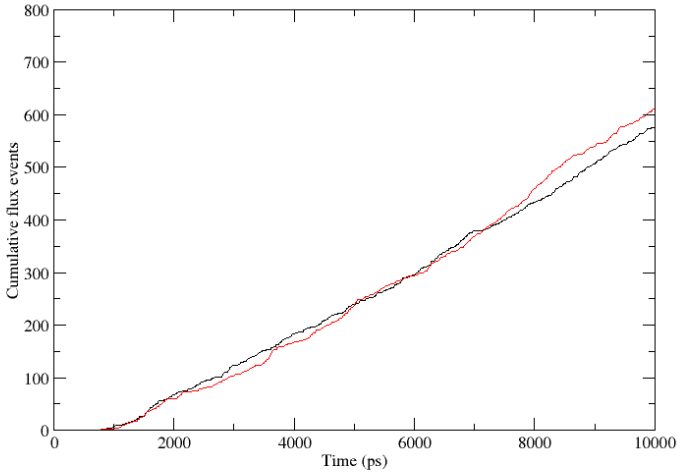


FIGURE 3.3: *Cumulative bidirectional water flux for the TRP- (black) and TYR- (red) anchored HG pores under a 0.2* $\mathrm{V\,nm^{-1}}$ *field and in the absence of ssDNA.*

the pore. *Full translocation* is defined as the phosphate group of the 3' nucleotide of the ssDNA strand passing through the pore exit.

3.5.2.1 Control simulations

Three control simulations were performed with ssDNA in close proximity to the hourglass pore entrance in the absence of any electric field. Under these conditions, the ssDNA was observed to coil around the pore entrance, interacting primarily with the anchoring

TABLE 3.2: *A summary of the outcomes of all DNA translocation simulations. Numbers in parentheses correspond to simulations run under a 0.1* $\mathrm{V\,nm^{-1}}$ *field strength, all other simulations were run under a 0.2* $\mathrm{V\,nm^{-1}}$ *field.*

Initial DNA location		**Final DNA location**		
		Pore Entrance	Pore Constriction	Pore Exit
Hourglass Pore (TRP anchored)	Pore Entrance	5 (2)	3 (1)	3
	Threaded	0	(3)	3
Hourglass Pore (TYR anchored)	Pore Entrance	3	3	4
	Threaded	0	0	6
Reference Pore	Pore Entrance	5	1	-
	Threaded	0	(3)	6

aromatic residues as shown in figure 3.4. At no point did the ssDNA strand enter the pore interior, as expected from the highly charged DNA backbone and the hydrophobic pore lining.

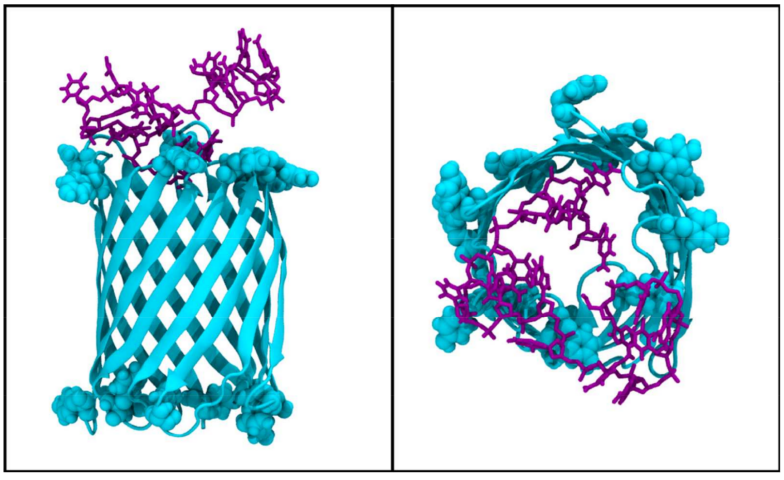


FIGURE 3.4: DNA as observed to coil around the pore entrance in the absence of any electric field, showing close preferential association to the bulky anchoring residues

3.5.2.2 ssDNA translocation behaviour

36 individual simulations were performed with the hourglass pores, the initial and final ssDNA positions of which are summarised in table 3.2. In all simulations, the hourglass pores remained stable and showed the expected bi-directional water flux from the validation simulations. Translocation of ssDNA was only observed on the timescale of these simulations by applying a 0.2 $\mathrm{V\,nm^{-1}}$ electric field, with test simulations done on pre-threaded ssDNA in the reference pore under a 0.1 $\mathrm{V\,nm^{-1}}$ electric field showing no instance of complete translocation.

Under an applied electric field strength of 0.2 $\mathrm{V\,nm^{-1}}$, DNA positioned outside the entrance to the reference pore did not enter the pore in 5 out of 6 repeat simulations

of 100 ns length, remaining associated with the pore mouth. In the remaining repeat, ssDNA progressed past the pore mouth but remained closely associated with the pore wall, and full translocation was not seen. In contrast to this, when the ssDNA strand was pre-threaded through the mouth of the pore (The phosphorous groups of the 5' and immediately following nucleotides positioned below the C-αs of the anchoring aromatic residues), complete and rapid translocation of the ssDNA strand was observed in under 10 ns, shown in Figure 3.5.

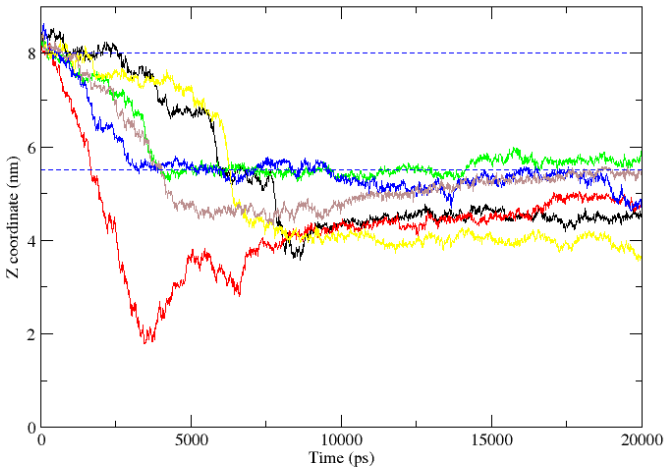


FIGURE 3.5: *The Z coordinate of the Cα of the leading nucleotide for all simulations in which ssDNA translocated through the reference pore. The pore entrance and exit are marked with dashed blue lines.All interruptions to translocation correspond to a new nucleotide coming into contact with the pore entrance.*

Overall, three distinct classes of DNA behaviour were observed:

1. ssDNA remained closely associated with the anchoring aromatic residues at the pore mouth.
2. ssDNA passed through the pore mouth but was unable to proceed through the hydrophobic constriction region.
3. ssDNA translocated fully through both the pore mouth and the constriction region.

In all simulations in which ssDNA was threaded through the constriction region, full translocation was observed. When the ssDNA was initially positioned at the pore mouth, all three DNA behaviours were seen in roughly equal proportion. Following full translocation, the ssDNA strand was seen to remain associated with the pore exit in a similar manner to the association seen with the aromatic anchoring residues at the pore mouth in instances where translocation was not seen, rather than moving away into the bulk solution. (Figure 3.7)

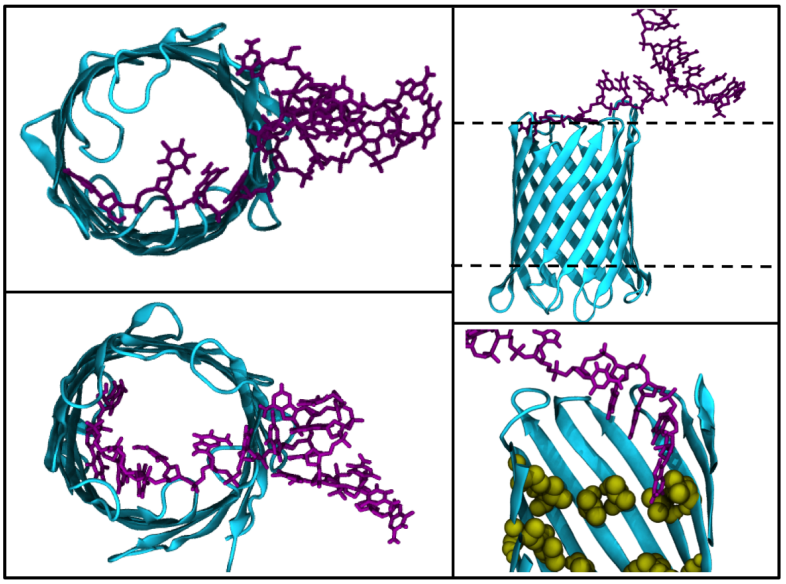


FIGURE 3.6: *ssDNA behaviour in two cases full translocation was not observed. (top) ssDNA does not enter the pore at all and remains closely associated with aromatic residues at the pore mouth and lipid head groups, and (bottom) where the leading nucleotide is unable to pass the constriction region, with valine residues shown in yellow.*

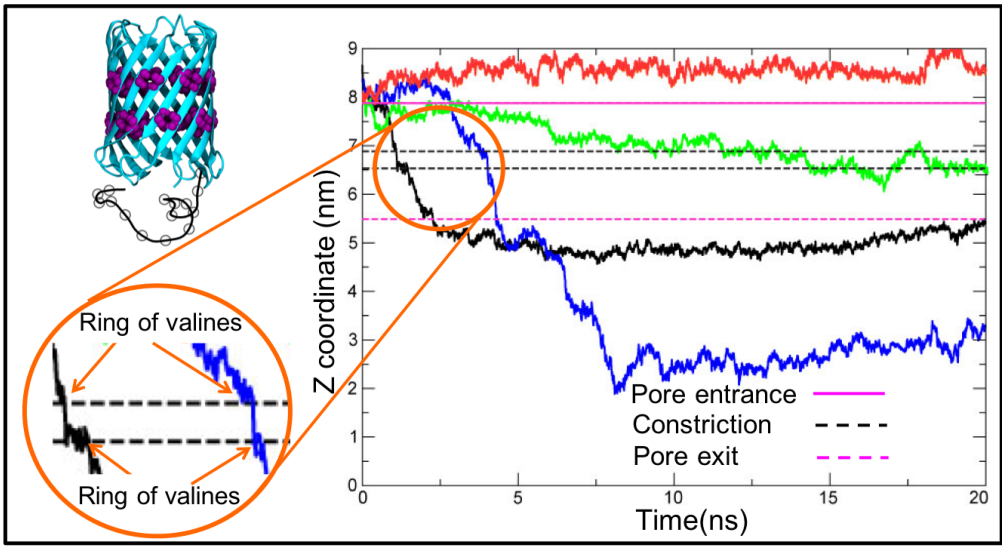


FIGURE 3.7: *Z position of the leading nucleotide as a function of time for four scenarios: (1) DNA remains at the pore entrance (red), (2) DNA is unable to pass the hydrophobic constriction (green), (3) DNA completes translocation and remains associated with the pore exit (black) and (4) DNA completes translocation and moves away from the pore into bulk solution. The inset highlights the DNA behaviour as it passes through the constriction region. The top-left image shows the tethering of the strand to the pore exit following translocation as described by line (4).*

In simulations that resulted in full ssDNA translocation, one of the aspects to be examined more closely was the conformational changes undergone by the ssDNA strand within the pore. Encouragingly, no significant secondary structure formation during translocation was observed, in contrast with previously reported work with α-hemolysin. The DNA end-to-end distances (Figure 3.8) show that in all translocating simulations the

DNA remains in an extended conformation during translocation, with a rapid reduction in end-to-end distance marking completion of translocation and subsequent strand coiling after exiting the pore.

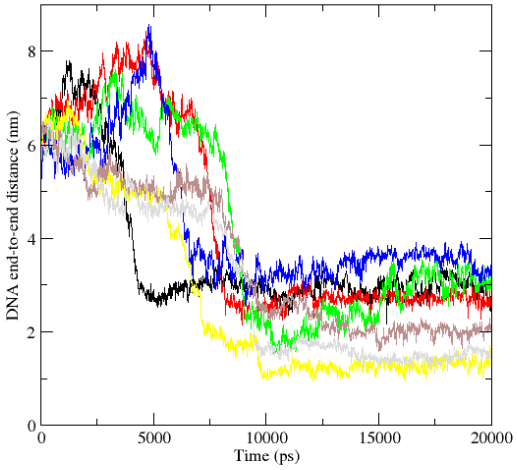


FIGURE 3.8: *DNA end-to-end distances as a function of time for all simulations in which ssDNA was positioned at the entrance to the HG pore and full translocation was observed. In all cases the initial (mostly linear) conformation was maintained or extended during translocation, with rapid coiling seen immediately following translocation.*

In the absence of the constriction, some interaction between the terminal nucleotide and pore wall residues is seen, but only results in partial bending of the ssDNA strand which does not proceed to true folding as shown in Figure 3.9.

3.5.2.3 Water flux during ssDNA translocation

During ssDNA translocation, the water flux was reduced by approximately 33% (to 45 $\pm$12 ns^{-1} and 43 $\pm$17 ns^{-1} in the reference and hourglass pores respectively) compared to that of the unobstructed pore but recovered to the expected rate once translocation was completed, shown in figure 3.10. This indicates both that the ssDNA strand does not induce any significant structural changes in the pore during translocation, and that ssDNA occludes species flow through the pore in a consistent and measurable manner. The rate of water flux during translocation was surprisingly similar between the reference and hourglass pores, which was somewhat unexpected.

In simulations where the ssDNA remained associated with the pore entrance, water flux was also reduced due to the increasing occlusion of the pore entrance by the coiling ssDNA strand.

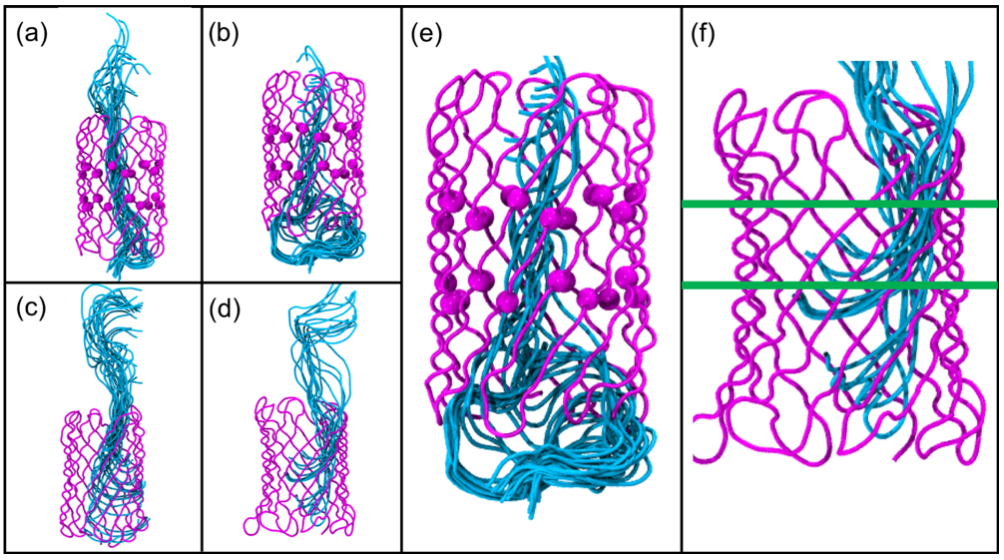


FIGURE 3.9: *Conformational comparison of ssDNA throughout simulation. ssDNA is shown in cyan at 200ps intervals whilst within the pore, with valine Cα shown as magenta spheres. (a) DNA enters and passes through the pore in an extended, linear conformation and only begins coiling when (b) the leading nucleotide becomes tethered to the pore exit. (c) and (d) show that in the reference pore, transient bending of the strand is seen throughout translocation but does not significantly hinder translocation. (e) and (f) show close up representations of (b) and (d) respectively. Green lines in (f) show the approximate position of the constriction region in an HG pore, demonstrating how the constriction encourages a linear conformation.*

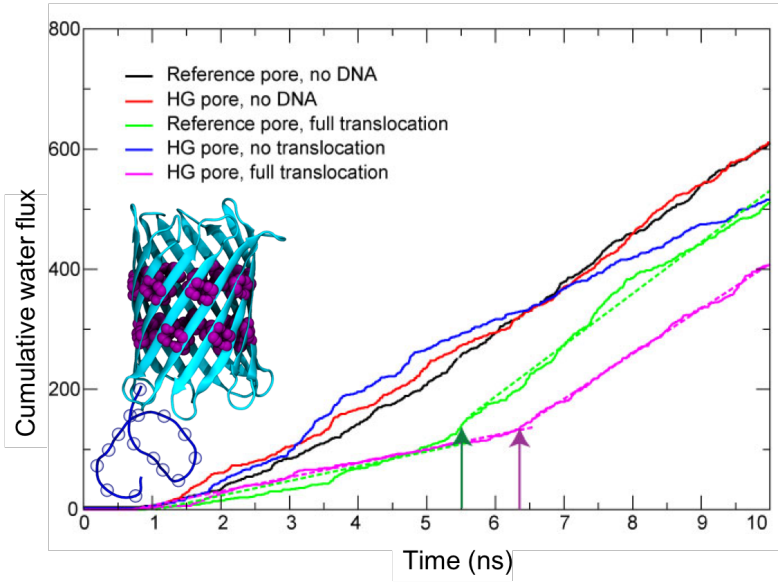


FIGURE 3.10: *Water flux (bidirectional) through the HG and reference pores, from a range of typical simulations. Linear regression for simulations with ssDNA translocation is included as dashed lines to emphasise the transition point from occluded to non-occluded water flux. Arrows are used to further indicate these points.*

3.6 Discussion

These results show that the artificial protein nanopores designed by Trick et al. can be successfully adapted for DNA translocation, with the modifiable constriction region providing a means of control for the translocation rate. The initial hypothesis for this work was that the hydrophobic nature of the artificial pore would be sufficient to maintain a linear conformation of a ssDNA strand during translocation, but in practice it was found that the constriction region is also required. Maintaining the linear conformation of ssDNA during translocation is of significant interest for sequencing applications, as it ensures the base read order corresponds to the actual sequence. Previous work by Bond et al.[52] showed that the presence of charged residues within a nanopore induce the formation of secondary structure in translocating ssDNA whereas aromatic side chains did not. The key interactions for inducing secondary structure formation within a nanopore were between DNA phosphate groups and cationic amino acid side chains, giving rise to an irregular and slow 'bind and slide' mechanism of translocation as well as the aformentioned secondary structure formation. The results reported here address both of those concerns, showing that a hydrophobic pore interior without the presence of any ionic side chains allows for a smooth mechanism of translocation that proceeds at a rapid pace. The hydrophobic pore in isolation does not prevent the ssDNA strand from kinking within the pore lumen, although full secondary structure formation was not observed. The inclusion of a constriction site within the pore removed even the small deformations in ssDNA conformation seen without, and thus demonstrate that both aspects are required for an extended linear conformation to be maintained during full translocation.

These improvements to ssDNA translocation profile and linear conformation maintenance have not created the 'ideal pore' however. Two impactful detriments to ssDNA translocation were identified , the more significant of which is the barrier to ssDNA entry seen at the pore mouth. Even under ideal conditions, with ssDNA approaching in a relatively linear conformation oriented parallel to the pore path, interactions at the pore mouth prevented ssDNA entry into the pore on these timescales in all but one repeat. This is clearly an area for improvement, but was outside the scope of this work.

The second barrier to translocation observed was at the constriction region itself, which was less disruptive than the interactions at the pore entrance but still proved to be a barrier, especially under lower electric field strengths. This is potentially a result of the somewhat severe profile demonstrated by the constriction site in these pores, where the pore lumen is reduced by 50% over a distance of 4nm. A modification to the pore models that may reduce the impact of this translocation barrier would be to increase the overall size of the nanopore. Moving from a 14-stranded β-barrel to a 16 stranded analogue would reduce the degree of constriction slightly, but also allow for greater variation in the modifications that could be made to it. Additionally, the wider pore lumen could

increase the area of approach in which capture by the anchoring residues at the pore mouth is reduced, lessening that barrier to translocation as well.

3.7 Future Direction

These results show that maintaining the ssDNA strand in a linear conformation is essential for smooth translocation to occur. To this end, further investigation into the anchoring residues present at the pore mouth would be of interest in order to either direct the ssDNA strand toward the hydrophobic constriction more effectively, or to reduce the need for a linear approach path altogether. The logical progression of this work is to investigate the sensitivity of these pores in distinguishing between the four DNA bases (A, C, T and G) to determine their suitability for DNA sequencing applications. The pores originally described by Trick et al. were not limited to those using a 14 stranded β-barrel scaffold, so it would be of interest to investigate the pores of other sizes for their respective suitability for DNA sequencing.

Chapter 4

Single-Stranded DNA Sequencing Using Hydrophobic Nanopores

4.1 Introduction

Having established that the hydrophobic nanopore models are stable under an applied persistent electric field, and that they enable smooth, linear translocation of ssDNA in a straighforward manner, the logical progression of this work was to examine the suitability of nanopores with a hydrophobic constriction region for DNA sequencing.

To this end, a series of simulations (summarised in Table 4.1)were carried out in which the suitability of the original hydrophobic HG pore previously described, as well as several variations on the constriction region, for distinguishing between the different DNA nucleotides was investigated.

The pore variants used in this study were based around two initial ideas; the previously studied pore with a *'VLV'* constriction region, and a pore with a constriction sequence of *'FNY'*. The inspiration for the latter pore came from the constriction found in the naturally occurring CsgG protein. The VLV pore was modified to examine the effect of the size of the constriction side chains on the pores sequencing ability, with alanine (small) and phenylalanine (large) mutations being introduced. The CsgG-inspired constriction is by nature asymmetrical, so I endeavoured to investigate the effect of each involved residue (F and Y) individually.

4.2 Simulation setup

The ssDNA strands used for these simulations are 12-nucleotide homogeneous strands, one for each DNA base (A, C, G and T). The 3DNA package[54, 55] was used to generate ssDNA strands in a helical conformation, which were then extended by restraining the 3' terminus and applying a moderate (0.5 $\mathrm{V\,nm^{-1}}$) electric field for a simulation of short duration in the NPT ensemble. These strands were positioned (in separate simulations)in the 5' direction in an extended conformation spanning the length of the pore, and harmonic restraints of 1000 $\mathrm{kJ\,nm^{-1}}$ were applied to the C-α atoms of the terminal residues

Two additional systems (Poly-G in the VLV constriction and poly-C in the FLF constriction) were set up with the DNA oriented in the 3' direction to investigate the effect of the strand directionality on the flow interruption, as has been previously postulated[68].

Under experimental conditions, distinguishing between ssDNA nucleotides during translocation is performed by measuring fluctuations in the magnitude of ionic current which are characteristic of each base. Measuring significant differences in ion conductance under simulation conditions would require substantially longer simulation times, as such water flux was used as a proxy measure for ion conductance to save on computational resource usage.

4.2.1 Simulation Protocols

The gromos53a6 force field [11, 63] with Berger lipid definitions [64] was used as described previously, with all systems being solvated using the SPC water model. Long range electrostatic interactions were treated using the PME method [65] with a short range cutoff of 1.4 nm. Van der Waals interactions were also truncated at 1.4 nm with long range dispersion corrections being added to the energy and pressure.

All simulation systems were equilibrated through 5 ns simulations in the NVT and NPT ensembles sequentially, using the Berendsen thermo- and barostats. Following this, a short (2 ns) simulation was performed using the full production parameters, velocity rescaling thermostat [39] and Parrinello-Rahman barostat [40] to ensure system stability prior to the longer simulations.

During simulation the system temperature was maintained at 310 K using the velocity rescaling thermostat with a coupling constant of 0.1 ps, and pressure was maintained semi-isotropically at 1 bar using the Parrinello-Rahman barostat with a time constant of 1 ps. A timestep of 2 fs was enabled through use of the LINCS algorithm for bond constraints [66]. Electric fields were applied as a constant voltage drop across the simulation box.

An electric field of 0.1 $\mathrm{V\,nm^{-1}}$ was applied for the full duration of all production simulations.

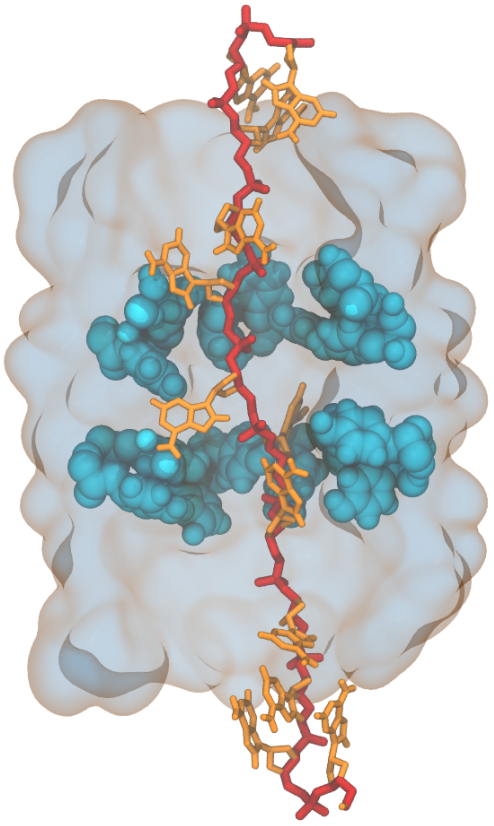


FIGURE 4.1: An example of system setup. Starting/Ending amino acid rings are shown in cyan, ssDNA side chains are shown in orange and the ssDNA backbone is highlighted in red to demonstrate its extended linear conformation.

TABLE 4.1: *A summary of all simulations run to investigate the DNA sequencing potential of nanopores with a hydrophobic constriction (Total Simulation time: 22.5* µs)

	DNA nucleotide			
Constriction Sequence	**A**	**C**	**G**	**T**
VLV	3x 600 ns	3x 300 ns	3x 300 ns	3x 300 ns
FLF	3x 300 ns	3x 300 ns	3x 300ns	3x 300 ns
ALA	3x 300 ns	3x 300 ns	3x 300ns	3x 300 ns
FNY	3x 300 ns	3x 300 ns	3x 300ns	3x 300 ns
FNF	3x 300 ns	3x 300 ns	3x 300ns	3x 300 ns
YNY	3x 300 ns	3x 300 ns	3x 300ns	3x 300 ns

TABLE 4.2: *Water flux rates (Given in bidirectional flux events per nanosecond) for simulations of 300* ns

Constriction Sequence	Repeat	DNA nucleotide A	C	G	T
VLV	1	20.24 ±5.56	23.24 ±6.32	17.44 ±5.74	22.67 ±5.23
	2	22.42 ±4.77	17.41 ±4.85	16.40 ±4.90	21.85 ±5.95
	3	25.77 ±6.42	17.23 ±4.42	17.00 ±4.76	23.79 ±5.95
FLF	1	14.98 ±4.83	10.24 ±3.69	7.89 ±3.13	19.28 ±6.08
	2	15.16 ±4.72	10.90 ±4.06	8.01 ±3.73	19.27 ±5.72
	3	12.98 ±4.29	11.17 ±4.42	7.67 ±3.21	16.97 ±4.83
ALA	1	27.73 ±6.89	22.50 ±5.79	24.11 ±6.80	26.49 ±6.18
	2	28.00 ±6.20	23.06 ±5.87	23.72 ±5.26	27.69 ±5.98
	3	27.00 ±6.13	24.71 ±6.23	25.71 ±5.98	28.33 ±6.95
FNY	1	8.53 ±3.36	4.52 ±2.23	6.47 ±3.06	6.98 ±3.05
	2	5.20 ±2.62	4.83 ±2.30	4.27 ±2.32	7.29 ±2.91
	3	8.66 ±3.49	4.24 ±2.10	8.49 ±3.45	7.84 ±3.50
FNF	1	6.82 ±2.94	7.86 ±2.79	6.50 ±3.04	10.68 ±4.09
	2	6.70 ±3.45	6.71 ±2.87	7.39 ±3.17	10.97 ±3.98
	3	5.15 ±2.64	7.13 ±2.77	8.46 ±3.61	12.29 ±4.51
ANA	1	6.06 ±3.05	2.92 ±1.79	3.46 ±1.88	5.63 ±3.05
	2	6.44 ±3.22	2.40 ±1.86	3.80 ±2.15	6.83 ±3.00
	3	5.97 ±3.12	3.93 ±2.27	2.58 ±1.68	6.21 ±3.07

4.3 Results

Table 4.2 contains the bidirectional water flux for all simulations run with restrained ssDNA within nanopores of varying constrictions. Some trends in the results are apparent, but not necessarily as expected. Generally speaking, the poly-T strand showed the highest flux rates in the wider pores (*VLV*, *FLF* and *ALA*), and the poly-G strand showed the lowest flux rates. However, the rates for the poly C, T and A strands were much more similar than expected, often giving results within error of one another. This is a bit counterintuitive, given that A and G are the bulkier pyrimidine bases and as such would be expected to show the most flux blockage, comparable with each other.

It is worth noting that the more constricted pores (*FNY, FNF and YNY*) all show quite similar flux rates, and the much lower numbers make the error much more significant. It is therefore difficult to draw meaningful conclusions from these results. One possible explanation for the similarity in these results is that the constriction aperture is narrow enough to be effectively completely occluded by more than one nucleotide. This would indicate that the nanopores based on the 14-stranded β-barrel scaffold with these constriction regions are too narrow to reliably distinguish between nucleotides of translocating ssDNA.

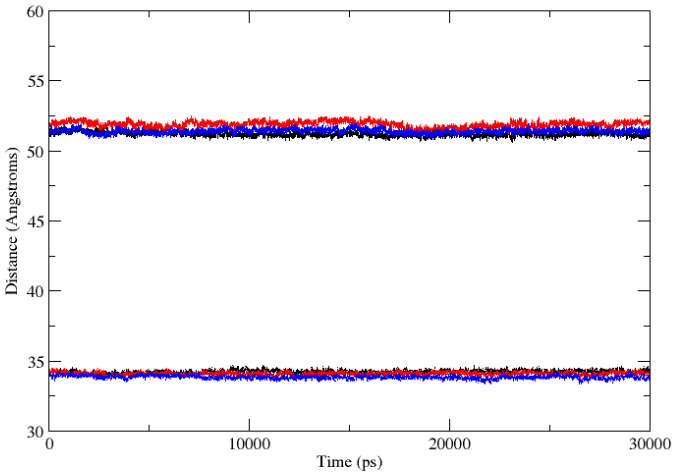


FIGURE 4.2: *ssDNA end-to-end distance for 3 repeats of the poly-C strand in the VLV constricted pore. The upper lines indicate the distance between terminal nucleotides, the lower indicate the distance spanned by the central 8 nucleotides*

An attempt was made to explain the water flux discrepancy by analyzing the conformation of the ssDNA strands in case any coiling was happening despite the restraints placed on the terminal nucleotides to maintain an extended conformation under the applied field. As shown in Figure 4.2, the end-to-end distance of the ssDNA throughout the simulation indicates that it maintained its extended conformation with little variation in all cases, and as such could not explain the unexpected results described previously.

4.4 Discussion

The aim of this work was to establish representative water flux values through model protein nanopores with an incorporated hydrophobic constriction for each of the four DNA bases, in order to lay a foundation to demonstrate the suitability of these pores for DNA sequencing purposes. Unfortunately convergence of the bidirectional water flux rates between repeats of each base was not seen, which could be due to a number of reasons, each of which would be a calid target for further investigation.

The short ssDNA strand length used, despite the application of restraints to keep it spanning the length of the pore, could be inciting unusual water behaviour at the entrance/exit of the pore, and as a related factor the restraints applied to the ssDNA to keep it spanning the pore could be forcing the bases into an unnatural arrangement, also interfering with water behaviour. The applied electric field used here (Potential difference of 300mV) across the membrane) is significantly stronger than has been used

in experimental studies of ssDNA in protein nanopores[69, 70], which could also be contributing to higher flux rates being seen than should be expected, particularly in cases of significantly occluded pores. The water model used here (SPC) may itself be responsible for the inconclusive results seen, which could be verified by application of alternative water models under the same simulation parameters.

4.5 Future Direction

In a further attempt to resolve the inconsistent water flux rates observed in the previously investigated systems, the next idea was to use a mobile ssDNA strand threaded through the pore, to investigate whether the rigidity of the restrained strand was interfering with the water translocation.

4.5.1 Periodic DNA

As shown previously, using a pre-threaded discrete ssDNA strand under an applied electric field results in rapid (sub-10ns) translocation through pores with a hydrophobic constriction. This timescale is too short to obtain meaningful flux data, so in order to extend the length of time that mobile DNA remained within the constriction region, a perdiodically bound ssDNA strand was used.

To create the periodic ssDNA models, a 62 base homonucleotide strand was created, following which the bases at both the 3' and 5' termini were removed. Next, a new bond definition was introduced between the exposed 'penultimate' bases at both ends of the strand, creating a 60 base long ssDNA strand periodically bound to itself in such a way that the cross-boundary linkage is indistinguishable from those present in the initial ssDNA strand.

These periodic ssDNA models were positioned such that they passed through the constriction region of the pore as shown in figure 4.3, with the simulation box dimensions adjusted such that the strand was allowed significant slack as well as free movement through the pore in order to attain an equilibrated mobile state. It was hoped that this approach would allow the water flux rates for each DNA base to converge to a more representative value, unlike those with large associated errors obtained previously. Unfortunately, time contraints led to these simulations being unable to be completed.

4.5.2 Multiple Constriction Sites

Further investigation for this work would be in the direction of investigating the implications of incorporating multiple constriction sites into hydrophobic pore models, in

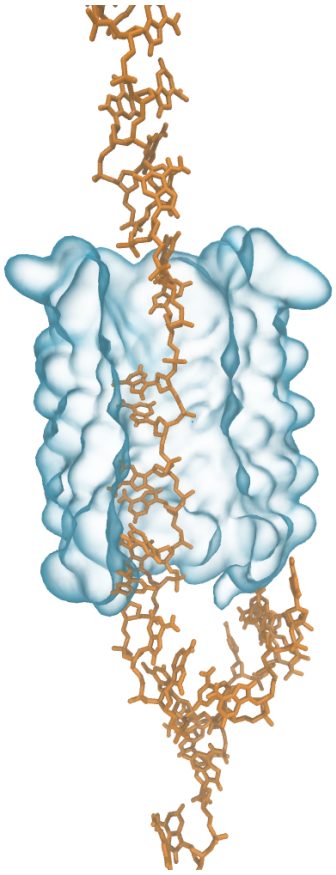


FIGURE 4.3: An example of a model pore with hydrophobic constriction (blue) set up with an ssDNA strand of infinite length (orange).

hopes of refining the sequencing ability of the pore by providing multiple read sites, as has been shown to improve sequencing accuracy in α-Hemolysin[71]. An example of this type of pore setup is shown in figure 4.4.

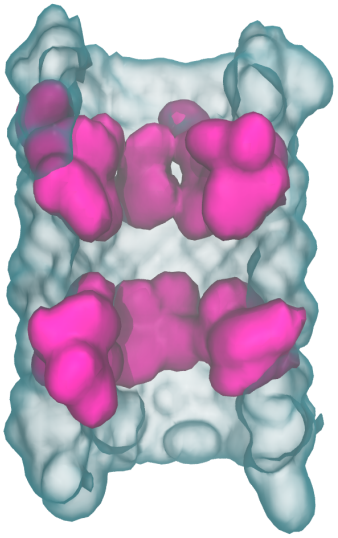


FIGURE 4.4: A model nanopore incorporating two constriction sites, highlighted in pink

Chapter 5

Conclusions

The stated aim of this work was to investigate the efficacy of incorporating hydrophobic gating mechanisms into transmembrane nanopores for the sequencing of ssDNA strands. The two major aspects of ssDNA sequencing investigated were the conformational behaviour of ssDNA during entry into, translocation through and exit from the nanopores of interest, and the ability of these hydrophobic gated pores to distinguish between individual bases using water flux as an analogue for ionic current.

I have shown that the hydrophobic gating structures incorporated here induce smooth, rapid translocation of an ssDNA strand in a linear conformation, without any formation of secondary structures within the pore that could interfere with an ionic current to cause inaccuracies in sequence reads. This is in contrast to the "bind and slide" behaviour seen in other ssDNA translocation simulation studies. A significant barrier to entry into hydrophobic pores was identified when the leading terminus of the ssDNA strand reaches the pore mouth. This barrier indicates that these pores are not viable for use in sequencing applications in isolation, as effective ssDNA capture from solution is an integral step in the sequencing process. A combination of proteins is a possible solution, incorporating a capping protein to the entrance of the hydrophobic transmembrane pore to enable more efficent capture of the ssDNA from solution and feeding into the pore to bypass the barrier to entry is one such example.

Investigation into the ability of nanopores with hydrophic constriction regions to distinguish between individual bases unfortunately did not provide conclusive results. This could have been due to a number of factors which were unable to be fully explored due to time constraints but which have been discussed in Chapter 4, section 4.4

Bibliography

(1) M. Chase, *J. Gen. Physiol.*, 1952, **36**, 39–56.

(2) J. Watson and F. Crick, *Nature*, 1953, **171**, 737–738.

(3) F. Sanger and S. Nicklen, *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 5463–5467.

(4) S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. Macdonald, A. Wing, C. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. Mcintosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y.-h. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg and J. C. Venter, *PLoS Biol.*, 2007, **5**, 2113–2144.

(5) L. D. Stein, *Genome Biol.*, 2010, **11**.

(6) C. J. Lo, T. Aref and A. Bezryadin, *Nanotech.*, 2006, **17**, 3264–3267.

(7) G. F. Schneider and C. Dekker, *Nature Biotech.*, 2012, **30**, 326–328.

(8) B. M. Venkatesan and R. Bashir, *Nat. Nanotechnol.*, 2011, **6**, 615–624.

(9) A. T. Kuan, J. A. Golovchenko, A. T. Kuan and J. A. Golovchenko, *Appl. Phys. Lett.*, 2012, **100**, 10–14.

(10) B. Bala, M. Venkatesan, A. B. Shah, J.-m. Zuo and R. Bashir, *Adv. Funct. Mater.*, 2010, **20**, 1266–1275.

(11) F. Schneider, S. W. Kowalczyk and V. E. Calado, *Nano. Lett.*, 2010, **10**, 3163–3167.

(12) Y. Tanaka, N. Hirano, J. Kaneko, Y. Kamio, M. Yao and I. Tanaka, *Protein Sci.*, 2011, **20**, 448–456.

(13) J. J. Kasianowicz, E. Brandin, D. Branton and D. W. Deamer, *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 13770–13773.

(14) M. Rincon-restrepo, E. Mikhailova, H. Bayley and G. Maglia, *Nano. Lett.*, 2011, **11**, 746–750.

(15) X.-f. Kang, S. Cheley, X. Guan and H. Bayley, *J. Am. Chem. Soc.*, 2006, **128**, 10684–10685.

(16) L. Song, M. R. Hobaugh, C. Shustak, S. Cheley, H. Bayley and J. E. Gouaux, *Science*, 1996, **274**, 1859–1866.

(17) M. Faller, M. Niederweis and G. E. Schulz, *Science*, 2004, **303**, 1189–1192.

(18) I. M. Derrington, T. Z. Butler, M. D. Collins, E. Manrao and M. Pavlenok, *Proc. Natl. Acad. Sci. USA*, 2010, **107**, 16060–16065.

(19) B. Cao, Y. Zhao, Y. Kou, D. Ni, X. C. Zhang and Y. Huang, *Proc. Natl. Acad. Sci. USA*, 2014, **111**, E5439–E5444.

(20) P. Goyal, P. V. Krasteva, N. V. Gerven, F. Gubellini, I. V. D. Broeck, J. S. Pinkner, M. R. Chapman, A. Troupiotis-tsa and H. Remaut, *Nature*, 2014, **516**, 250–256.

(21) J. Wang, N. E. Moore, Y. M. Deng, D. A. Eccles and R. J. Hall, *Frontiers in Microbiology*, 2015, **6**, 1–7.

(22) E. Karlsson, A. Lärkeryd, A. Sjödin, M. Forsman and P. Stenberg, *Sci. Rep.*, 2015, **5**, 1–8.

(23) P. M. Ashton, S. Nair, T. Dallman, S. Rubino, W. Rabsch, S. Mwaigwisya, J. Wain and J. O'Grady, *Nature Biotech.*, 2015, **33**, 296–299.

(24) A. L. Greninger, S. N. Naccache, S. Federman, G. Yu, P. Mbala, V. Bres, D. Stryke, J. Bouquet, S. Somasekar, J. M. Linnen, R. Dodd, P. Mulembakani, B. S. Schneider, J.-J. Muyembe-Tamfum, S. L. Stramer and C. Y. Chiu, *Genome Med.*, 2015, **7**.

(25) M. Misini Ignjatovic, O. Caldararu, G. Dong, C. Munoz-Gutierrez, F. Adasme-Carreno and U. Ryde, *J. Comput. Aided Mol. Des.*, 2016, 1–24.

(26) S. Lee, A. Tran, M. Allsopp and J. B. Lim, *J. Phys. Chem. B*, 2013, **118**, 547–556.

(27) A. H. D. Vries, *J. Phys. Chem. B*, 2007, **111**, 7812–7824.

(28) W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, 1995, **117**, 5179–5197.

(29) A. D. Mackerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wio, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, 3586–3616.

(30) W. R. P. Scott, P. H. Hu, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kru and W. F. V. Gunsteren, *J. Phys. Chem. A*, 1999, **103**, 3596–3607.

(31) R. Hockney, S. Goel and J. Eastwood, *Journal of Computational Physics*, 1974, **14**, 148–158.

(32) W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *The Journal of Chemical Physics*, 1982, **76**, 637–649.

(33) C. M. Welch, A. N. Camden, S. A. Barr, G. M. Leuty, G. S. Kedziora, R. J. Berry, C. M. Welch, A. N. Camden, S. A. Barr, G. M. Leuty, G. S. Kedziora and R. J. Berry, *J. Chem. Phys.*, 2015, **143**.

(34) D. Galimberti, C. Quarti and A. Milani, *Polymer*, 2015, **67**, 167–173.

(35) U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J. Chem. Phys.*, 1995, **103**, 8577–8593.

(36) H. J. C. Berendsen, J. P. M. Postma, W. F. V. Gunsteren, A. Dinola, J. R. Haak, H. J. C. Berendsen, J. P. M. Postma, W. F. V. Gunsteren, A. Dinola and J. R. Haak, *J Chem Phys*, 1984, **81**, 3684.

(37) S. Nosé, *J. Chem. Phys.*, 1984, **81**, 511–519.

(38) W. G. Hoover, *Physical Review A*, 1985, **31**, 1695–1697.

(39) G. Bussi, D. Donadio, M. Parrinello, G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, DOI: 10.1063/1.2408420.

(40) M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182–7190.

(41) R. Manara, Doctoral Thesis, University of Southampton, 2015.

(42) A. T. Guy, T. J. Piggot and S. Khalid, *Biophysical Journal*, 2012, **103**, 1028–1036.

(43) P. Blount and P. C. Moe, *Trends in Microbiology*, 1999, **7**, 420–424.

(44) K. Yoshimura, A. Batiza, M. Schroeder, P. Blount and C. Kung, *Biophysical Journal*, 1999, **77**, 1960–1972.

(45) O. Beckstein, O. Beckstein and M. S. P. Sansom, *Phys. Biol.*, 2006, **3**, 147–159.

(46) M. R. Powell, L. Cleary, M. Davenport, K. J. Shea and Z. S. Siwy, *Nature Nanotechnology*, 2011, **6**, 798–802.

(47) P. Aryal, F. Abd-Wahab, G. Bucci, M. S. P. Sansom and S. J. Tucker, *Nature Communications*, 2014, **5**, 4377.

(48) F. Zhu and G. Hummer, *Journal of Chemical Theory and Computation*, 2012, **8**, 3759–3768.

(49) J. L. Trick, E. J. Wallace, H. Bayley and M. S. P. Sansom, *ACS Nano*, 2014, **8**, 11268–11279.

(50) Q. Huang, D. Wang, Q. Zhao and X. Guan, 2015 International Conference on Manipulation, Manufacturing and Measurement on the Nanoscale (3M-NANO), 2015, pp. 267–271.

(51) R. S. S. de Zoysa, D. M. M. Krishantha, Q. Zhao, J. Gupta and X. Guan, *ELECTROPHORESIS*, 2011, **32**, 3034–3041.

(52) P. J. Bond, A. T. Guy, A. J. Heron, H. Bayley and S. Khalid, *Biochemistry*, 2011, **50**, 3777–3783.

(53) Schrödinger, LLC, "The {PyMOL} Molecular Graphics System, Version 1.8", 2015.

(54) X.-j. Lu and W. K. Olson, *Nucleic Acids Research*, 2003, **31**, 5108–5121.

(55) X.-J. Lu and W. K. Olson, *Nature Protocols*, 2008, **3**, 1213–1227.

(56) C. D. Berweger, W. F. van Gunsteren and F. Müller-Plathe, *Chemical Physics Letters*, 1995, **232**, 429–436.

(57) B. Hess, C. Kutzner, D. Van Der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.

(58) M. J. Abraham, T. Murtola, R. Schulz, S. Páall, J. C. Smith, B. Hess and E. Lindah, *SoftwareX*, 2015, **1-2**, 19–25.

(59) S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. Van Der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.

(60) H.J.C.Berendsen, D. der Spoel and R. Drunen, *GROMACS: A message-passing parallel molecular dynamics implementation*, 1995.

(61) D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.

(62) S. Páll, M. J. Abraham, C. Kutzner, B. Hess and E. Lindahl, *Lecture Notes in Computer Science*, 2015, **8759**, 3–27.

(63) C. Oostenbrink, A. Villa, A. E. Mark and W. F. Van Gunsteren, *J. Comput. Chem.*, 2004, **25**, 1656–1676.

(64) O. Berger, O. Edholm and F. Jahnig, *Biophys. J.*, 1997, **72**, 2002–2013.

(65) T. Darden, D. York and L. Pedersen, *The Journal of Chemical Physics*, 1993, **98**, 10089–10092.

(66) B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.

(67) O. S. Smart, J. G. Neduvelil, X. Wang, B. A. Wallace and M. S. P. Sansomt, *J. Mol. Graphics*, 1996, **7855**, 354–360.

(68) J. Mathé, A. Aksimentiev, D. R. Nelson, K. Schulten and A. Meller, *Proceedings of the National Academy of Sciences*, 2005, **102**, 12377–12382.

(69) G. Maglia, M. R. Restrepo, E. Mikhailova and H. Bayley, *Proceedings of the National Academy of Sciences*, 2008, **105**, 19720–19725.

(70) T. Vu, J. Borgesi, J. Soyring, M. D'Alia, S.-L. Davidson and J. Shim, *Nanoscale*, 2019, **11**, 10536–10545.

(71) D. Stoddart, G. Maglia, E. Mikhailova, A. J. Heron and H. Bayley, *Angewandte Chemie International Edition*, 2010, **49**, 556–559.