UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF CHEMISTRY

Optimising nanopores for DNA sequencing: A computational perspective

by

Punam Rattu

Thesis for the degree of Doctor of Philosophy

February 2023

University of Southampton

Abstract

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF CHEMISTRY

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

OPTIMISING NANOPORES FOR DNA SEQUENCING: A COMPUATIONAL PERSPECTIVE

by

Punam Rattu

Nanopore DNA sequencing is a well-established technology that has accelerated advancements in many fields, including medical research. Over the years, research has focussed on optimising protein nanopores for DNA sequencing. Optimisation strategies broadly focus on (1) slowing the translocation of DNA to increase the time available for base recognition and (2) improving the resolution of detection to attain single-base sensing. Molecular dynamics (MD) simulations have been invaluable in obtaining molecular-level insights to pave the way for informed nanopore optimisation. In this thesis, MD simulations were used to study DNA translocation through nanopores and elucidate the design principles for optimising nanopores for DNA sequencing.

In the first chapter, the translocation of short and longer single-stranded (ss)DNAs was studied through protein-inspired hydrophobic nanopores with dual-constrictions. It was found that DNA translocation is slowed down by aromatic residues, and when combined with a narrow geometry, DNA retains a largely linear conformation during translocation and without forming secondary structures that can impede DNA sequencing. Following this, the proteins CsgG and the CsgG-CsgF complex were characterised in terms of their conformational dynamics and ability to allow DNA translocation. Eyelet loops forming the CsgG constriction were found to exhibit large variations in their mobility, with at least one loop moving upwards into the vestibule under an applied electric field. CsgF was found to stabilise CsgG and the eyelet loop region. Subsequently, the translocation of short ssDNA through CsgG and the CsgG-CsgF complex was studied. The speed of DNA translocation was found to be primarily influenced by DNA interacting with key residues in the CsgG constriction region. DNA is retained in a more linear conformation during translocation through the dual-constriction hydrophobic channel formed by the CsgG-CsgF complex compared to CsgG. Next, the translocation of longer ssDNA in an applied electric field was studied through these proteins/protein complex. These simulations revealed that the eyelet loops of the CsgG constriction region are mobile during DNA translocation, and the stochastic nature of their mobility perturbs the pore geometry which may give rise to noise in the ionic current through the nanopores. In the last chapter, Markov State Model methodology was employed to characterise the kinetics of the mobility of the CsgG eyelet loops under an applied electric field. The model construction was limited by the duration of the MD simulations. The data and analyses presented in this, and previous, chapters emphasise the need for a model that describes the complex conformational dynamics of the CsgG eyelet loops.

Table of Contents

Table of Tables

Table of Tables

Table of Figures

Table of Figures

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Punam Rattu

Title of thesis: Optimising nanopores for DNA sequencing: A computational perspective

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while 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:
    - Rattu, P., Belzunces, B., Haynes, T., Skylaris, CK., and Khalid, S., *Translocation of flexible and tensioned ssDNA through in silico designed hydrophobic nanopores with two constrictions.* Nanoscale, 2021. **13**(3): p. 1673-1679.
    - Rattu, P., Glencross, F., Mader, SL., Skylaris, CK., Matthews, SJ., Rouse, SL., and Khalid, S., *Atomistic level characterisation of ssDNA translocation through the E. coli proteins CsgG and CsgF for nanopore sequencing.* Computational and Structural Biotechnology Journal, 2021. **19**: p. 6417-6430.

Signature: ..... Date:

Acknowledgements

I would like to thank my supervisor, Professor Syma Khalid, for her guidance and endless support during the course of my PhD. I would also like to thank Jayne Wallace and Lakmal Jayasinghe from Oxford Nanopore Technologies for their insightful discussions.

I would also like to thank all the members of the Khalid group for all their help and support. In particular, a special thank you to Kamolrat Somboon for inspiring and motivating me; in her, I have found a friend for life.

On a personal note, I am eternally grateful for my parents, sister, and extended family, without whom this would have been impossible. In addition, I would like to thank Sophie Lewis for making me laugh whenever I needed it the most. Lastly, a special thanks to the whole team of BBC Asian Network for soundtracking the last four years with laughs, banter, and fantastic music.

***

ਮੈਂ ਆਪਣੀ ਸੁਪਰਵਾਈਜ਼ਰ ਪ੍ਰੋਫ਼ੈਸਰ ਸਾਇਮਾ ਖਾਲਿਦ ਦਾ ਧੰਨਵਾਦ ਕਰਨਾ ਚਾਹਾਂਗੀ, ਮੇਰੀ PhD ਦੇ ਦੌਰਾਨ ਉਸ ਦੇ ਮਾਰਗਦਰਸ਼ਨ ਅਤੇ ਬੇਅੰਤ ਸਮਰਥਨ ਵਾਸਤੇ। ਔਕਸਫੋਰਡ ਨੈਨੋਪੋਰ ਟੈਕਨੋਲੋਜੀਜ਼ ਤੋਂ ਜੇਨ ਵੈਲੇਸ ਅਤੇ ਲਕਮਲ ਜੈਸਿੰਘੇ ਦਾ ਧੰਨਵਾਦ ਕਰਨਾ ਚਾਹਾਂਗੀ ਉਹਨਾਂ ਦੀ ਸੂਝ ਭਰਪੂਰ ਚਰਚਾ ਵਾਸਤੇ।

ਮੈਂ ਖਾਲਿਦ ਸਮੂਹ ਦੇ ਸਾਰੇ ਮੈਂਬਰਾਂ ਦਾ ਧੰਨਵਾਦ ਕਰਨਾ ਚਾਹਾਂਗੀ, ਉਹਨਾਂ ਦੀ ਹਰ ਮਦਦ ਅਤੇ ਸਮਰਥਨ ਵਾਸਤੇ। ਖਾਸ ਤੌਰ ਤੇ, ਮੈਨੂੰ ਪ੍ਰੇਰਿਤ ਕਰਨ ਲਈ ਕਮੋਲਰੈਟ ਸੋਮਬੂਨ ਦਾ ਵਿਸ਼ੇਸ਼ ਧੰਨਵਾਦ। ਉਸ ਵਿੱਚ, ਮੈਨੂੰ ਜੀਵਨ ਲਈ ਇੱਕ ਦੋਸਤ ਮਿਲੀ ਹੈ।

ਇੱਕ ਨਿੱਜੀ ਨੋਟ ਤੇ, ਮੈਂ ਆਪਣੇ ਮਾਤਾ-ਪਿਤਾ, ਭੈਣ ਅਤੇ ਪਰਿਵਾਰ ਦਾ ਸਦਾ ਲਈ ਧੰਨਵਾਦੀ ਹਾਂ, ਜਿਨ੍ਹਾਂ ਤੋਂ ਬਿਨਾਂ ਇਹ ਅਸੰਭਵ ਸੀ। ਇਸ ਤੋਂ ਇਲਾਵਾ, ਮੈਂ ਸੋਫੀ ਲੁਈਸ ਦਾ ਧੰਨਵਾਦ ਕਰਨਾ ਚਾਹਾਂਗੀ, ਕਿ ਜਦੋਂ ਵੀ ਮੈਨੂੰ ਸਭ ਤੋਂ ਵੱਧ ਲੋੜ ਹੋਵੇ ਤਾਂ ਮੈਨੂੰ ਹਸਾਉਣ ਵਾਸਤੇ। ਅੰਤ ਵਿੱਚ, ਬੀ.ਬੀ.ਸੀ. ਏਸ਼ੀਅਨ ਨੈੱਟਵਰਕ ਦੀ ਪੂਰੀ ਟੀਮ ਦਾ ਪਿਛਲੇ ਚਾਰ ਸਾਲਾਂ ਵਿੱਚ ਹਾਸੇ, ਮਜ਼ਾਕ ਅਤੇ ਸ਼ਾਨਦਾਰ ਸੰਗੀਤ ਲਈ ਵਿਸ਼ੇਸ਼ ਧੰਨਵਾਦ।

***

मैं अपने पर्यवेक्षक प्रोफेसर सायमा खालिद को तैह दिल से धन्यवाद देना चाहती हूं, मेरी PhD के दौरान उनके मार्गदर्शन और समर्थन के लिए । मैं ऑक्सफोर्ड नैनोपोर टेक्नोलॉजीज के जेने वालेस और लकमल जयसिंघे को भी उनकी अंतर्दृष्टिपूर्ण चर्चा के लिए धन्यवाद देना चाहती हूं।

मैं खालिद समूह के सभी सदस्यों को उनकी मदद और समर्थन के लिए धन्यवाद देना चाहती हूं। विशेष रूप से, मुझे प्रेरित करने के लिए कौम्लराट सोम्बून को विशेष धन्यवाद , उसमें मुझे जीवन भर के लिए एक दोस्त मिला है।

व्यक्तिगत रूप से, मैं अपने माता-पिता, बहन और परिवार के लिए सदा आभारी हूं, जिनके बिना यह असंभव होता। इसके अलावा, मैं सोफी लुईस को धन्यवाद देना चाहता हूं कि जब भी मुझे इसकी सबसे ज्यादा जरूरत थी, मुझे हंसाने के लिए। अंत में, पिछले चार वर्षों में हंसी, मजाक और शानदार संगीत के लिए बीबीसी एशियन नेटवर्क की पूरी टीम को विशेष धन्यवाद।

Abbreviations and Definitions

ATP ..... Adenosine triphosphate

CF ..... Constant force

cryo-EM ..... Cryo-electron microscopy

CV ..... Constant velocity

DFT ..... Density Functional Theory

DNA ..... Deoxyribonucleic acid

dNTP ..... Deoxynucleoside triphosphate

DPPC 1,2-dipalmitoyl- sn-glycero-3-phosphocholine

dsDNA ..... Double-stranded deoxyribonucleic acid

E. coli ..... *Escherichia coli*

ITS ..... Implied timescale

LJ ..... Lennard-Jones

MD ..... Molecular dynamics

MM ..... Molecular mechanics

MSM ..... Markov state model

NMR ..... Nuclear magnetic resonance

NTP ..... Nucleotide triphosphate

OM ..... Outer membrane

ONT ..... Oxford Nanopore Technologies

PBC ..... Periodic boundary conditions

PCA ..... Principal component analysis

PDB ..... Protein data bank

PME ..... Particle Mesh Ewald

POPC 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine

RMSD ..... Root mean square deviation

RMSF ..... Root mean square fluctuation

Abbreviations and Definitions

RNA ..... Ribonucleic acid

SPC ..... Simple point charge

ssDNA..... Single-stranded deoxyribonucleic acid

TCM..... Transition count matrix

TICA..... Time-lagged independent component analysis

TIP3P ..... Transferable intermolecular potential with 3 points

TPM..... Transition probability matrix

USB..... Universal Serial Bus

VAMP ..... Variational Approach to Markov Processes

vdW..... van der Waals

VMD ..... Visual Molecular Dynamics

WT..... Wild type

Chapter 1 Introduction

1.1 Nanopore-based sensing

Nanopore sensors are powerful single-molecule sensors that have been successfully employed to detect biomolecules. Nanopore-based sensing technology has been used for sequencing DNA [1-4], RNA, and proteins [5-8] and also for rapidly detecting a wide variety of analytes [9] with organic chemistries such as polymers [10, 11], and inorganic chemistries, such as ions [12-14] and metallic nanoparticles [15-17], at a relatively low cost [18]. In medical diagnostics, nanopores have been shown to detect molecules that are markers of disease at low concentrations in small sample volumes [19], enabling diagnosis and progression monitoring of diseases and various types of cancers [20], including breast cancer [21].

A nanopore sensor consists of a nanometer-sized aperture in an insulating membrane that segregates two electrolyte-filled compartments (*cis* and *trans*). When an electrical potential difference is applied between the two compartments, ions traverse through the nanopore, which results in an ionic current. The ionic current is temporarily disrupted when analytes enter the nanopore. The capture and transport of analytes by the nanopores depends on the analyte's properties; whilst charged analytes are pulled through the nanopore *via* electrophoresis [22], neutral analytes are transported by the electroosmotic flow [23, 24]. Once inside the nanopore, analytes block the ionic current *via* the volume exclusion model of the nanopore conductance [25] and by forming non-bonded interactions with the nanopore lumen [26]. As the magnitude and duration of the disruption in ionic current is dependent on the analyte's properties, such as its size, charge distribution, and frequency and lifetime of interactions with the nanopore [27], the change in ionic current be used to gain insights into the analyte properties (Figure 1.1). Additionally, the concentration of the analyte can be discerned from the frequency at which the ionic current is disrupted [28].

Nanopores can be classified into two types: biological nanopores, formed by transmembrane proteins embedded in lipid bilayers, and solid-state nanopores, fabricated into synthetic materials such as graphene [29-31]. Although solid-state nanopores are highly stable and durable, it is difficult to manufacture nanopores accurately with a precise pore geometry [32]. Biological nanopores offer many advantages over solid-state nanopores; as transmembrane proteins consistently form pores with atomically precise dimensions, it is easy to mass-produce identical nanopores [9]. Proteins can also be precisely tailored by introducing mutations to alter nanopore properties, such as pore diameter and surface charges, to increase the sensitivity and specificity

of the nanopore [9, 33]. In addition, protein nanopores are less noisy than solid-state nanopores, as they are typically operated at low voltages [34].

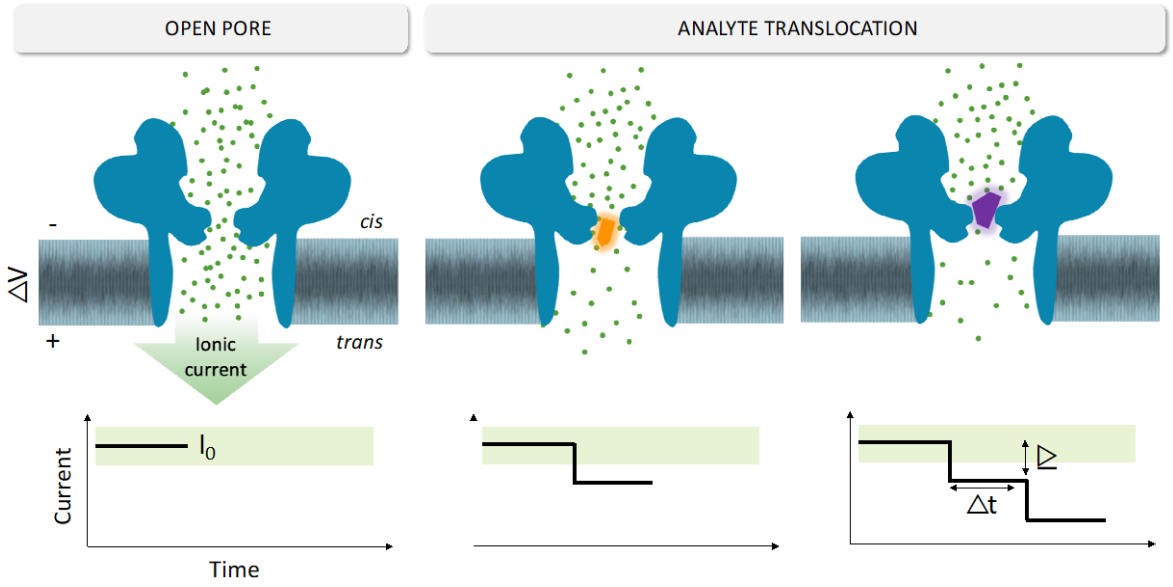

Figure 1.1: The nanopore sequencing device. A nanopore device consists of a nanometer-sized hole separating two electrolyte-filled compartments, called *cis* and *trans*. An applied voltage across the membrane drives molecules in the *cis* compartment to translocate to the *trans* compartment through the nanopore. The fluctuations in ionic current during analyte passage are the primary signal for nanopore sensing. I_0=open pore current, △I=current blockage, △t=dwell time of the analyte.

1.2 Nanopore DNA sequencing

Deoxyribonucleic acid (DNA) forms the genetic material of almost all known living organisms. DNA is composed of nucleotides, which contain one of the four bases, adenine (A), guanine (G), cytosine (C), and thymine (T). The sequence of these nucleotides within the DNA molecule encodes the hereditary instructions required for building and maintaining organisms. DNA sequencing, therefore, enables the investigation of the impact of genetic variations on physiological and pathological levels and the identification of the genetic risk factors associated with complex human diseases [35, 36]. In addition, it continues to have an emerging role in therapeutics and personalised medicine [37, 38].

DNA sequencing using nanopores is now a well-established 'next-generation sequencing' approach that can be used to sequence DNA with high specificity [39], sensitivity [40], and speed [41]. Nanopore sensors can sequence long fragments and eliminate the DNA sample amplification or chemical labelling steps, reducing errors that arise when reconstructing the original sequence [42]. Nanopores have been successfully commercialised for DNA and RNA sequencing by Oxford Nanopore Technologies (ONT) [43]. The devices developed by ONT contain multiple nanopore sensors that simultaneously sequence ultra-long DNA fragments that can exceed a megabase (1,000,000 bases) in length [42]. The smallest nanopore device offered by ONT is the MinION, which measures only 3.3 × 10.5 × 2.3 cm in size and can fit in the palm of one's hand. MinION can be powered through laptop computers' Universal Serial Bus (USB) port, making the device extremely portable and sequencing to be conducted anywhere [44]. MinION has been shown to sequence DNA with very high accuracy exceeding 99.8 % [42]. The cost of the device is also much lower compared to other DNA sequencing technologies, the initial cost being only around $1,000, including the device and initial reagents [45].

1.2.1 DNA structure and properties

Deoxyribonucleic acid (DNA) is a biological macromolecule consisting of two strands coiling around each other to form a double helix. Each strand is a polynucleotide composed of subunits referred to as nucleotides. A nucleotide consists of a 5-carbon sugar called deoxyribose, a phosphate group, and a nucleobase (Figure 1.2). In a polynucleotide, nucleotides are linked *via* a covalent bond between the deoxyribose of one nucleotide and the phosphate group of the consecutive nucleotide, forming the DNA sugar-phosphate backbone. The two polynucleotide strands that form DNA are also referred to as single-stranded (ss)DNA. The two ends of ssDNA are termed 5' terminus and 3' terminus, named according to the number of the carbon atom in deoxyribose to which the phosphate group is bonded (Figure 1.2).

Nucleotides differ due to the presence of one of the four nitrogenous bases, adenine (A), guanine (G), cytosine (C), and thymine (T) (Figure 1.3). Each nucleobase differs in its chemical structure, but the four bases are classified into two groups according to the number of ring groups they contain. Nucleobases cytosine and thymine are referred to as pyrimidines, as they contain pyrimidine, a heterocyclic aromatic organic compound containing two nitrogen atoms. The larger bases adenine and guanine are purines and contain an additional imidazole group fused to the pyrimidine group.

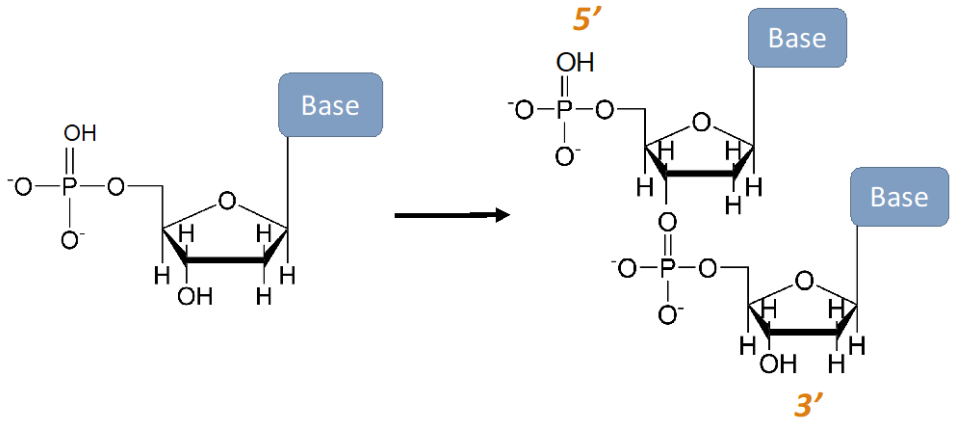

Figure 1.2: Structure of a DNA nucleotide (left), and 2 nucleotides linked *via* a covalent bond between the deoxyribose and phosphate groups in a polynucleotide (rigt). The 5' terminus and 3' terminus of the strand are labelled.

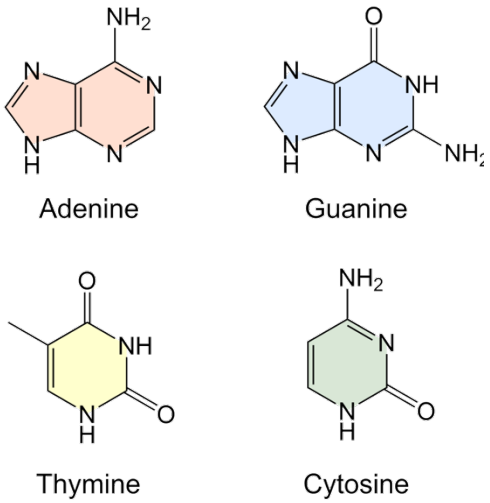

Figure 1.3: Structure of the four DNA nucleobases.

In the DNA double helix, the two ssDNA strands are held together by hydrogen bonds between their bases. Each nucleobase interacts with its complementary partner; adenine pairs with thymine *via* two hydrogen bonds, and guanine pairs with cytosine *via* three hydrogen bonds (Figure 1.4). This complementary pairing is caused by the size of the bases and the hydrogen bonds that can be formed between the nucleobase pairs.

DNA is anionic due to a negatively charged oxygen atom in the phosphate group of each nucleotide in the polynucleotide strands. This property is exploited during nanopore DNA sequencing, where ssDNA is pulled through the nanopore *via* electrophoresis.

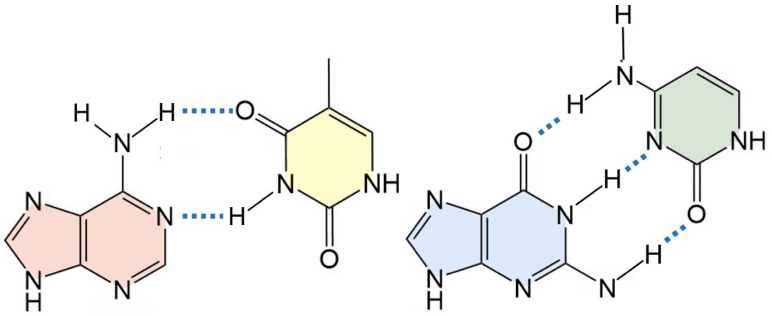

Figure 1.4: Complementary nucleobases interact by forming hydrogen bonds (marked by dashed lines), which hold the two strands together in DNA.

1.2.2 Biological nanopores for DNA sequencing

Proteins that naturally form nanometer-sized pores and act as ion channels, porins, aquaporins, pore-forming toxins, and viral pores can be repurposed as nanopores for DNA sequencing. For a protein to be considered for use as a nanopore, it should ideally have several fundamental properties. The protein must remain stable under conditions implemented for DNA sequencing and following engineering *via* mutagenesis, insertions, or deletions. It must also form a channel of stable geometry that is not altered by, e.g., gating activity, to ensure there is little noise in baseline ionic conductance. The protein must also be easy to insert into membranes used in nanopore sequencing devices. β-barrel proteins are considered suitable candidates for nanopores as they are resistant to whole-protein changes due to the engineering of specific regions and insert into lipid membranes more readily than those with an α-helical transmembrane region.

Several protein pores have been investigated for their potential use for DNA sequencing, including α-hemolysin, MspA, aerolysin, FhuA, OmpF, OmpG, and ClyA [9]. Of these, α-hemolysin, MspA, aerolysin, and CsgG have been extensively studied and optimised for DNA sequencing. Therefore, this chapter will focus on these proteins from here on.

1.2.2.1 α-hemolysin

α-hemolysin was the first protein used to demonstrate the possibility of DNA sequencing using nanopores [1]. α-hemolysin is a pore-forming toxin secreted by the gram-positive bacterium *Staphylococcus aureus.* Secreted as water-soluble monomers, the protein self-assembles in target cell membranes to form an aqueous channel that allows the unregulated exit of water, ions, and vital molecules, ultimately leading to target cell death *via* permeabilization and lysis.

The high-resolution X-ray structure (PDB 7AHL, 1.9 Å) [46] shows that α-hemolysin is a homo-heptamer protein characterised by an extracellular cap domain connected to a transmembrane 14-stranded β-barrel stem, and seven rim domains lining the cap that lies close to the cell membrane. The extracellular cap domain conceals a large solvent-accessible vestibule that tapers down to a ~ 1.5 nm-wide constriction where it connects to the transmembrane stem. The stem is ~ 5.2 nm in height and varies in diameter (~ 1.5-2.4 nm) due to the volume of amino acid side chains protruding into the primarily hydrophobic interior. Both ends of the stem lumen are capped by rings of acidic and basic residues, unlike the rest of the stem, which is lined with neutral amino acids (Figure 1.5).

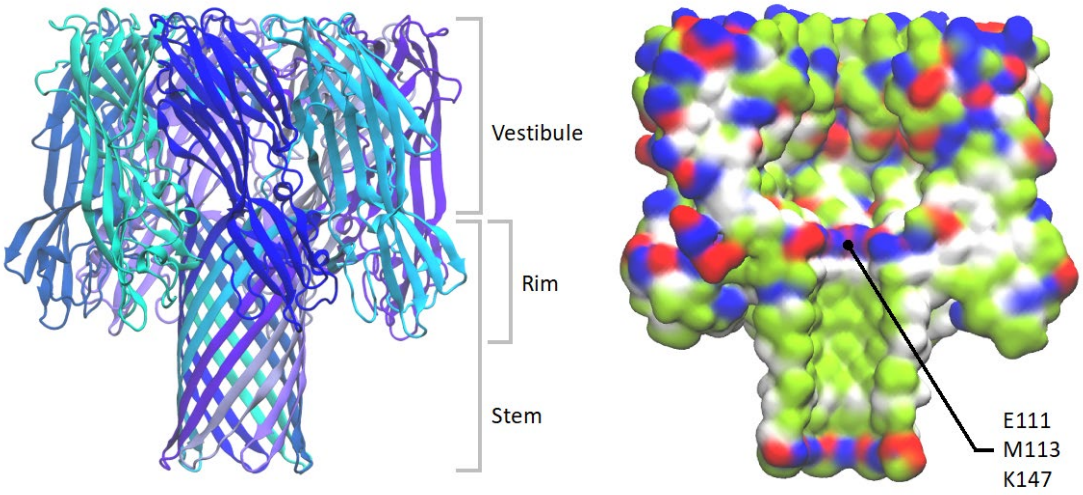

Figure 1.5: α-hemolysin (PDB 7AHL, 1.9 Å) shown from the side, approximately parallel to the membrane, and depicted in ribbon (left) and surface representations (right). The vestibule, rim, and stem domains are labelled. The inner surface of α-hemolysin is shown with the three residues in the constriction region labelled (right). Red: anionic side chains, blue: cationic side chains, green: polar side chains, and white: non-polar side chains.

The constriction in the α-hemolysin channel is composed of alternating rings of glutamic acid, methionine, and lysine residues (Glu-111, Met-113, and Lys-147) (Figure 1.5). It is wide enough to accommodate linear macromolecules such as ssDNA, which is ~ 1.2 nm in diameter; hence, in 1996, Kasianowicz et al. demonstrated that when a voltage bias of ~ 120-180 mV is applied across the membrane with the embedded α-hemolysin channel, ssDNA and RNA molecules placed in *cis* chamber electrophoresed through the nanopore to the *trans* chamber. Their passage caused an 80-90% reduction in the ionic current of the α-hemolysin pore, of which the duration correlated

with polymer length and thus the time it took for the molecule to fully translocate through the nanopore. As the diameter of α-hemolysin constriction is slightly wider than the ssDNA, polynucleotides translocate through the aperture in an extended conformation and with nucleotides moving in a single file [1]. In addition to ssDNA [4, 47-50] and RNA [51-53], α-hemolysin has been used to detect DNA hairpins [54], peptides [55, 56], and proteins [57]. α-hemolysin is stable and does not undergo critical structural transitions in a wide range of temperatures and conditions required for nanopore sequencing (in KCl solution and an applied voltage bias of ~ 120-180 mV) [58].

1.2.2.2 MspA

MspA (PDB 1UUN, 2.5 Å) [59] is a porin found in the outer membranes of *Mycobacterium smegmatis*, where it facilitates the entry of small hydrophilic nutrients. It is a goblet-shaped homo-octameric protein with a periplasmic rim domain, a transmembrane stem domain, and an extracellular base domain. The stem domain is a ~ 4 nm-wide 16-stranded β-barrel connected to the base domain, which is another shorter 16-stranded β-barrel (Figure 1.6). The base domain sharply narrows to a ~ 1 nm wide and ~ 1 nm long channel constriction formed by acidic side chains of aspartate residues (Asp-90 and Asp-91).

MspA was proposed as a promising candidate for nanopore sequencing because the ionic current blockades during DNA translocation arise as the polynucleotide traverses the short and narrow constriction region. Wild-type MspA is unsuitable for DNA sequencing because the open pore current is unsteady due to spontaneous and frequent current blockades. The negatively charged environment of the constriction region also prevents the anionic ssDNA from entering the pore. Using site-directed mutagenesis, MspA was engineered for DNA sequencing by replacing anionic residues with neutral asparagine residues in pore constriction and cationic residues such as arginine and lysine at the vestibule entrance, which resulted in an increased rate of DNA capture and translocation through the pore [60]. Experiments with immobilised ssDNA have shown that the current blockage in MspA is induced by three nucleotides residing within or near the constriction region [61]. MspA is one of the most stable porins characterised to date; its structure remains stable over pH values of 2-14 and temperatures up to 92 °C [59].

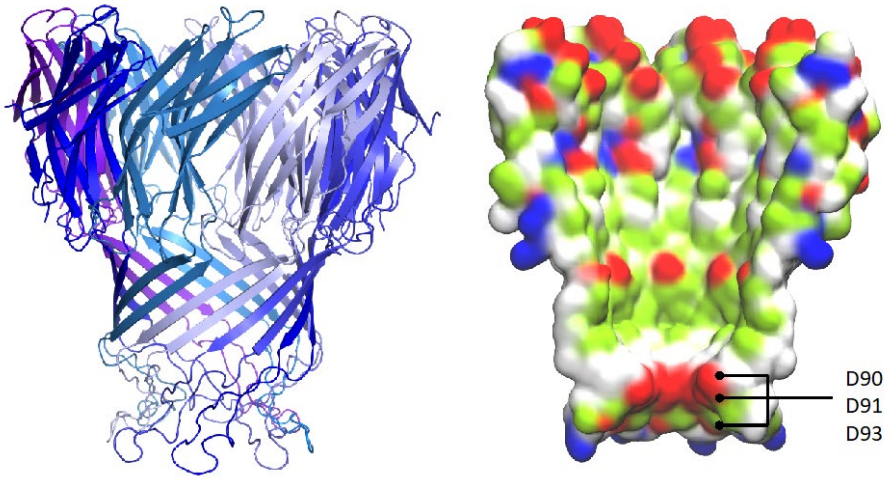

Figure 1.6: MspA (PDB 1UUN, 2.5 Å) shown from the side, approximately parallel to the membrane, depicted in ribbon (left) and surface representation (right). The inner surface of MspA is shown with the residues in the constriction region labelled (right). Red: anionic side chains, blue: cationic side chains, green: polar side chains, and white: non-polar side chains.

1.2.2.3 Aerolysin

Aerolysin (PDB 5JZT, 7.4 Å) [62] is a pore-forming toxin secreted by *Aeromonas hydrophila.* It is secreted as water-soluble monomers that self-assemble in target cell membranes to induce cell death. Like α-haemolysin, aerolysin is a homo-heptameric protein with a mushroom-liked structure and a transmembrane β-barrel stem. However, aerolysin contains a small ~ 1.8 nm wide extracellular cap instead of a large vestibule region. The cap is connected to a ~10 nm long transmembrane 14-stranded β-barrel that forms a channel with a diameter ranging from ~ 1.0-1.7 nm [63]. Two ~ 1.0 nm-wide constriction regions flank the β-barrel, referred to as R1 and R2, at the exocellular entrance and adjacent to the intracellular side, respectively [26]. Basic residues form both constriction regions; R1 contains two arginine residues, Arg-282 and Arg-220, and R2 includes two lysine residues, Lys-238 and Lys-242 [64]. The pore is lined with many positively charged residue sidechains pointing into its lumen [63]. Aerolysin exhibits a unique fold constituted by two concentric β-barrels at the top of the pore, which are stabilised by hydrophobic interactions [65].

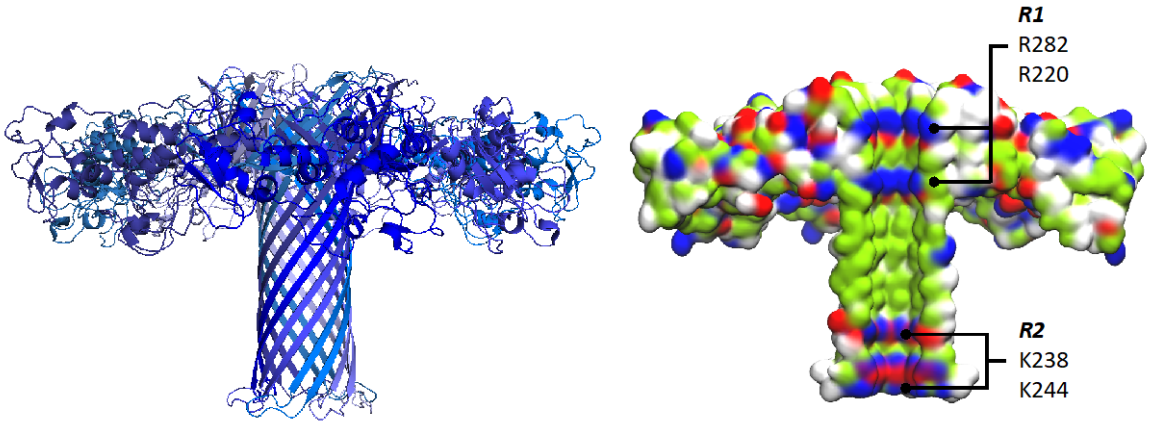

Figure 1.7: Aerolysin (PDB 5JZT, 7.4 Å) shown from the side, approximately parallel to the membrane, and depicted in ribbon (left) and surface representations (right). The inner surface of aerolysin is shown with the two constriction regions labelled (right). Red: anionic side chains, blue: cationic side chains, green: polar side chains, and white: non-polar side chains.

Aerolysin was first demonstrated as a nanopore sensor in 2006 and applied in the analysis of the alpha-helix collagens [56]. Since then, it has been used for detecting DNA [66, 67] and for studying peptides [8, 56, 68] and the dynamics of protein unfolding [69]. More recently, WT aerolysin was shown to successfully recognise and discriminate amongst all 20 amino acids during the protein sequencing [8]. For DNA sequencing, WT aerolysin has demonstrated optimal single-base resolution and translocation rate that is~ 3 orders of magnitude slower [70] than the average translocation rate of nucleotides in WT and mutants of α-haemolysin [2, 71], without requiring any additional engineering of the pore [71] nor DNA immobilisation [72]. Furthermore, aerolysin can discriminate differences as minor as a single methyl group; the current blockage and the translocation duration of an oligonucleotide containing methylcytosine were 0.82 pA greater and 1.36-fold longer, respectively, compared to cytosine, due to the difference in size and the non-covalent interactions formed by each with the charged residues in the aerolysin pore lumen [26]. Despite the nanopore containing two constriction regions, the R1 constriction is responsible for base recognition [64]. Aerolysin is a highly stable protein thanks to its double barrel fold, which is ideal for use as a nanopore sensor.

1.2.2.4 CsgG

CsgG is an outer membrane lipoprotein found in *Escherichia coli.* It is a component of the curli biogenesis system, which is multi-protein machinery that facilitates the secretion of curli subunits and their assembly into highly aggregative amyloid fibres associated with biofilm formation in

Gram-negative bacteria. It is a specialised class of transporters that do not rely on ATP or chemical gradients for secretion. CsgG is transported to the outer membrane by the lipoprotein localisation transport pathway, during which the cysteine residues at the amino terminus (C16) are lipidated before being inserted into the outer membrane [73].

The crystal structure of full-length lipidated CsgG extracted from the outer membrane (PDB 4UV3, 3.59 Å) [74], along with negative-stain electron micrographs, revealed the protein to be a crown-shaped symmetric nonameric complex that is 12 nm in width and 8.5 nm in height, with a 36-stranded β-barrel traversing the outer membrane connected to a large solvent-accessible vestibule that opens up into the periplasmic space (Figure 1.8). Two of the long β-strands forming the transmembrane β-barrel (β3 and β4) extend and, along with an additional β1 strand, partake in forming the vestibule. The vestibule also contains α1 and α3 helices lining the outer surface and α2, the inner surface of the channel between which the three-stranded β-sheet is sandwiched.

The protein lumen is predominantly lined with acidic residues in the β-barrel apart from the top and bottom regions of the pore that are highly hydrophobic. Like the transmembrane domain, the interior of the vestibule is negatively charged near the apex but positively charged near the mouth opening to the periplasm. The structure also showed two aromatic belts formed from tyrosine residues anchoring the protein in the membrane.

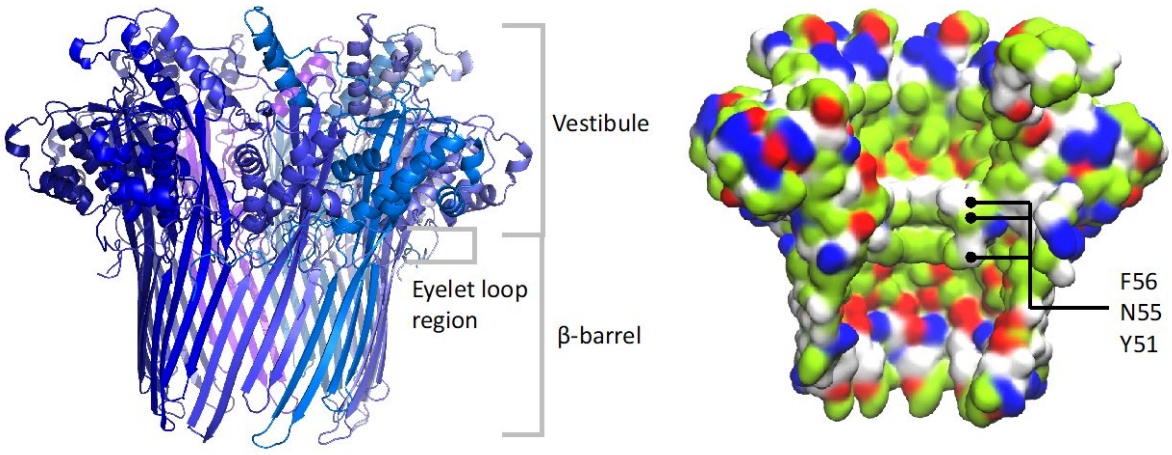

Figure 1.8: CsgG (PDB 4UV3, 3.59 Å) shown from the side, approximately parallel to the membrane, and depicted in ribbon (left) and surface representations (right). The vestibule, eyelet loop region, and the β-barrel domains are labelled. The inner surface of CsgG is shown with key residues forming the constriction region labelled (right). Red: anionic side chains, blue: cationic side chains, green: polar side chains, and white: non-polar side chains.

CsgG forms a channel of ~ 4 nm diameter that widens to ~ 5 nm in the vestibule mouth formed by α2 helices and connecting loops. The β-barrel and the vestibule regions inside the channel are partitioned by a 12-residue N terminal eyelet loop linking the β1 to α1, which rises from the bottom of the channel (periplasmic domain) and folds back at the outer membrane-periplasm interface. The eyelet loops of CsgG protomers form a solvent-excluded ~ 0.9 nm diameter constriction at the level of lipid-aqueous interface, characterised by stacked rings of tyrosine, asparagine, and phenylalanine residues (Figure 1.9). Mutagenesis experiments and multiple sequence alignments of CsgG to similar translocator proteins have shown that the presence of the stacked phenylalanine ring and hydrogen bond donor/acceptor like asparagine and tyrosine is required for translocation of curli subunits [74].

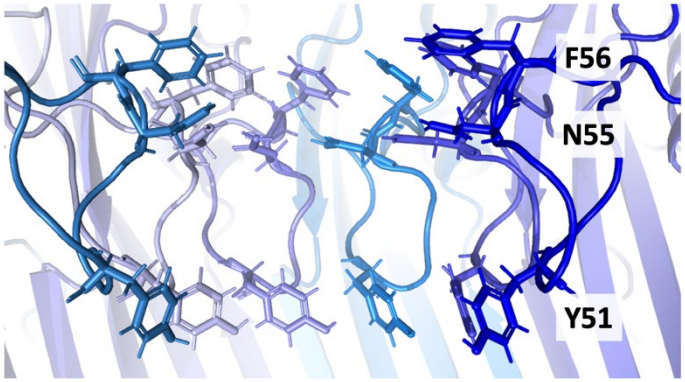

Figure 1.9: CsgG eyelet loops that form the constriction region are shown, with key residues shown in stick representation.

The X-ray structure and the electron density maps could not resolve the diacylglycerol- and amide-linked acyl chain added to the N terminal cysteine residue. Still, it is expected to lie outside the transmembrane β-barrel domain anchoring the protein in the outer membrane. The membrane-bound CsgG has also been resolved to a higher resolution of 2.9 Å (PDB 3X2R) [75]. However, it lacks the first 34 and last 12 residues from the N terminus and C terminus, respectively, due to proteolytic digestion before crystallography, and residues 116-128 and 251-265 due to crystallographic disorder. An accompanying electron density map of the whole CsgG protein used for phase determination is also available (PDB 4Q79, 3.1 Å).

Modifications to the CsgG protein were first reported to be potentially suitable for DNA sequencing in 2014 [43]. CsgG was later adopted by Oxford Nanopore Technologies (ONT) and is now being used commercially in nanopore devices after further engineering [76].

1.2.3 Comparison of biological nanopores

α-hemolysin, MspA, aerolysin, and CsgG share features in common that are promising for DNA sequencing. The sensing region of the four nanopores narrows to ~ 1-1.5 nm diameter, similar to the ~ 1.2 nm diameter of ssDNA. In this region, the confined DNA strand is forced into an extended conformation, which ensures that nucleotides traverse the sensing region in a single file so that the blockade in the ionic current directly correlates to the passage of a single nucleotide. Additionally, the slight difference in diameter results in the differences in current blockages to be enhanced.

Although similar in diameter, the length of the sensing region varies amongst the three proteins. Early experiments involving the translocation of polynucleotides of varying lengths through α-hemolysin showed that the ~ 90-92 % reduction in the open pore current was caused by a minimum of 12 nucleotides. As the β-barrel can accommodate up to ~ 12 nucleotides at a time, the entirety of the stem region acts as the sensing region in α-hemolysin, and the nucleotide-induced blockade in the ~ 1 nm wide constriction region is diluted [4]. In contrast, the length of sensing regions in MspA and CsgG nanopores are shorter and can only accommodate ~ 4-5 nucleotides at a given time. Therefore, the signal resolution is higher than α-hemolysin as fewer nucleotides contribute to the reduction in ionic current measured through the nanopore. The sensing region in aerolysin is even shorter; it can detect oligonucleotides as short as 2 nucleotides long and discriminate between oligonucleotides that range from 2 to 10 nucleotides in length with high sensitivity [67].

α-hemolysin, MspA, and CsgG also contain large vestibule regions through which the DNA strand enters the narrow sensing region of the nanopore. However, the vestibule of α-hemolysin is narrower (~ 1.5 nm) than in MspA (~ 4 nm) and CsgG (~ 5 nm), therefore the current blockage during sequencing is affected to a greater extent by DNA occupying the vestibule region in the α-hemolysin pore. Aerolysin lacks a large vestibule region and instead contains a smaller cap region that is slightly wider than α-hemolysin (~ 1.8 nm).
Examination of open pore current and ssDNA-induced current blockage through the four pores showed that the current in both scenarios is higher through MspA and CsgG, as they form wider channels (barring the constriction regions) than α-hemolysin and aerolysin. The current blockage induced by ssDNA is almost the same in all pores, indicating that the four pores represent promising candidates for sensing [77].

1.3 Nanopore optimisation for DNA sequencing

Ideally, the current measured through the nanopore during DNA sequencing can be used to identify the bases and read the DNA sequence with confidence. To achieve this, firstly, each base must reside in the sensing region for a duration that is long enough to measure the associated ionic current, and secondly, the differences in the ionic current measured must be distinguishable for bases. Therefore, nanopore optimisation strategies focus on controlling the rate of DNA translocation and the base-specific blockade of the ionic current through the nanopore.

1.3.1 Molecular dynamics simulations

DNA translocation through nanopores is influenced by a multitude of factors, such as DNA conformations during translocation, the interactions between residues in the pore lumen with the four bases, channel dynamics, and the behaviour of the ions and water molecules inside the channel during DNA translocation. An immensely complex interplay between these factors makes the rational optimisation of nanopores a challenging endeavour. Molecular dynamics (MD) simulations have proven to be highly useful, as they have provided atomistic insights into the aforementioned factors to pave the way for informed nanopore optimisation. The MD method is described in detail in Chapter 2.

MD simulation studies of protein nanopores and the translocation of DNA through them typically employ methods such as applying an electric field [78-80] and using steered MD [81, 82]. Electric field simulations replicate experimental conditions as the transmembrane bias enables the movement of charged species, such as ions and DNA, through the nanopore [83]. These simulations have been demonstrated to quantitively predict the open pore conductance through protein nanopores α-hemolysin [80], MspA, CsgG, and aerolysin [77], and also through other proteins [84-87], yielding values that are in rough agreement with experiments. Additionally, the conductance of α-hemolysin has been faithfully reproduced in simulations of simplified models of the protein which omit the vestibule and contain the transmembrane β-barrel region only [88]. Electric field simulations are also used to characterise the molecular details of DNA translocation and the modulations in the nanopore's ionic current caused by DNA due to its conformation [79, 89-93]. Steered MD simulations involve the application of an external force to slowly pull the analyte through the nanopore. This method has been used to determine the free energies of translocation of polynucleotides through α-hemolysin [94], and can also aid in identifying the pathways that DNA may take through the nanopore in a computationally efficient manner [26, 93]. A detailed description of this method is provided in section 2.7.

DNA translocation through nanopores has been extensively studied using MD simulations varying levels of detail, ranging from atomic level in all-atom simulations to modelling multiple atoms as a single bead in coarse-grained simulations. Whilst coarse-grained simulations can simulate DNA translocation on millisecond timescales [95-97], all-atom simulations are more accurate and provide atomistic detail and hence are typically used to study nanopores and DNA translocation.

1.3.2 Controlling the rate of DNA translocation

In experiments of WT α-hemolysin, DNA transits through the pore at the rate of ~ 1 μs/nucleotide (in 100 mV). Hence, the ionic current for each nucleotide is measured for the short duration of 1 μs only. As the ionic current during ssDNA passage is typically ~ 10 pA, equivalent to ~ 60 ions/μs flowing through the nanopore, DNA translocation at the rate of ~ 1 μs/nucleotide results in ~ 60 ions being used for detecting each base [98]. Consequently, the precision of the current measurement (equivalent to the standard deviation divided by the mean) is at most ± 13 %, which is too low to distinguish the subtle > ~ 1 pA differences in the ionic currents between the four bases. The thermodynamic fluctuations associated with DNA translocation further convolute the differences in ionic current blockage caused by the bases. To circumvent this issue, the DNA translocation rate can be reduced to increase the nucleotide dwell time in the sensing region, so that more ions are used for measuring the associated ionic current precisely.

To control the high translocation speeds, some of the approaches investigated include increasing the viscosity of the solvent [99], decreasing the temperature [58], adding organic salts [100], attaching complementary DNA sequences [101], and decreasing the voltage bias [1]. However, these approaches reduce the translocation speeds by less than one order of magnitude whilst also causing a substantial decrease in the ionic current [102]. One of the strategies proven to be successful is the use of processive motor enzymes, which ratchet DNA nucleotides through the pore in a controlled and stepwise manner. Motor enzymes harness the free energy released during the catalytic hydrolysis of high-energy phosphate bonds in nucleotide triphosphates (NTPs) or deoxynucleoside triphosphates (dNTPs) to step from one nucleotide to the next along the DNA strand. As the enzyme moves along the strand, it ratchets the nucleotides so that each pauses inside the pore for a few milliseconds before proceeding translocation.

The motor enzyme used for nanopore sequencing must be selected carefully, as almost half of the errors in sequencing reads are due to its malfunction [103, 104]. The enzyme must generate sufficient force to translocate the DNA strand against a competing electrophoretic force of the applied voltage. Additionally, the enzyme's association with DNA must be stable enough for the strand to undergo multiple catalytic cycles such that very long DNA strands can be translocated

through the nanopore. Finally, the motor enzyme must be large enough so that it is not pulled through the nanopore.

Some of the enzymes demonstrated as motors include DNA polymerases from *Escherichia coli* (Klenow fragment) and bacteriophage *phi29* DNA polymerase [3, 105-107], and helicases such as Hel308 [108-110]. DNA polymerase associates with the DNA strand by binding to the primer sequence ligated to the 3' terminus of the strand. Once the polymerase-bound ssDNA is captured and drawn into the nanopore, the enzyme sits atop the nanopore and adds new nucleotides to the 3' terminus of the primer as it steps in 3' to 5' direction along the DNA strand that is gradually pulled out of the nanopore. The *phi29* DNA polymerase is a promising motor enzyme as it can remain tightly associated with the DNA against the pulling voltage bias during nanopore sequencing [107]. It can also synthesise long DNA strands under a wide range of forces, which is necessary to efficiently oppose the electrophoretic force required to move the DNA strand through the pore [111]. Furthermore, it can operate at salt concentrations typically used during nanopore DNA sequencing [112].

Helicases motor enzymes separate double-stranded DNA (dsDNA) strands into single-stranded DNA (ssDNA). During nanopore sequencing, helicases associate with dsDNA strands, to which a synthetic anionic polymer is attached to the 5' terminus. The helicase-bound dsDNA is captured by the nanopore due to the anionic polymer being pulled through the nanopore. Following this, the helicase separates dsDNA into two ssDNA strands and steps one of the ssDNA through the nanopore at a rate of ~ 500 nucleotides/second, whilst the complementary strand is displaced in the surrounding solution [110]. In contrast to polymerases that pull ssDNA through nanopores against the force of the electric field, helicases are used to limit the rate at which the electric field drives a DNA strand through the nanopore. Helicases offer an advantage over polymerases as they bind single-stranded nucleic acid molecules at and initiate movement from random positions along the lattice, while polymerases require a partial duplex where the new nucleotides are added to the 3' terminus of the primer [113].

Overall, motor enzymes significantly reduce the translocation rate through nanopores from more than 1,000,000 nucleotides/second to 10-1,000 nucleotides/second [114].

1.3.3 Understanding DNA-nanopore interactions for optimising DNA base identification

Along with slowing the rate of DNA translocation, another challenge is to achieve single-base resolution, i.e., the current measured through the nanopore corresponds to the blockade caused by a single base only. Currently, there are no known protein nanopores with a constriction region

that hosts a single base; the shortest sensing region is in aerolysin, which can host ~ 2 bases simultaneously [67].

The sensing capabilities of nanopores are influenced by pore geometry (the length of the channel and its sensing region, and the geometry of the vestibular region if present) and its chemistry as determined by the nature of amino acids that line the pore lumen [115]. Both affect the mobility of ions inside the nanopore and modulate the pore current [116-119]. The geometry of the constriction region determines the sensitivity of the nanopore. The ionic current is dominated by the residues occupying the space in the constriction region because the electric field is most intense at the narrowest section of the pore [120]. Detection is most sensitive when the pore diameter is similar to the size of ssDNA, as it results in a large ratio between open-pore and blocked-pore ionic currents. For this reason, engineering the sensing region is one of the strategies employed to improve the nanopore's capability for sequencing DNA.

During DNA sequencing, along with the size of the bases, the pore current is also influenced by non-bonded interactions that DNA forms with the pore lumen. In aerolysin, the DNA-pore interactions were shown to prolong the duration for which the strand resides in the pore, which led to a higher current blockage. The non-bonded interactions between DNA and the pore also affected the pathway of ions inside aerolysin; electrostatic interactions affected ion mobility, along with van der Waals interactions which introduced the dipole-induced electrostatic potential to influence ion mobility. Additionally, hydrogen bonds greatly influence the dipole moment and further lead to the change in the dielectric constant inside the nanopore [26]. Therefore, optimising the chemistry of the pore lining is another strategy employed for optimising nanopores.

Protein engineering of α-hemolysin followed by electrophysiology measurements revealed that basic residues introduced into the lumen of the pore of α-hemolysin slowed down the translocation of ssDNA [71, 121]. Subsequent MD simulation studies of simplified models of the α-hemolysin pore, consisting of the transmembrane β-barrel region only, showed that DNA translocation is slowed when the DNA backbone tethers to key lysine or arginine residues of transmembrane β-barrel in wild-type and mutant pores respectively. The side chains of these basic amino acids interact electrostatically with the negatively charged phosphate groups in the DNA backbone, affecting its conformation; in the presence of arginine rings in mutant pores, ssDNA formed hairpins and other nonlinear structures, which explained the slower speeds of translocations observed experimentally [88]. Similarly, the translocation rate through WT aerolysin was observed to be ~ 3 orders of magnitude slower than the average translocation rate of nucleotides in α-haemolysin due to basic residues predominantly lining the pore lumen [67]. A

subsequent study involving model α-hemolysin pore showed that although basic residues slowed down DNA translocation rate, the introduction of charged lysine or arginine residues (G119K or G119R mutants) resulted in DNA coiling and non-sequential DNA translocation [81]. Thus, along with slower translocation speeds, it is essential to ensure that the ssDNA remains in a linear conformation during translocation for accurate sequencing.

The approach of using model pores was later employed to study nanopores with a hydrophobic interior to prevent the tethering of charged DNA strands inside the pore. Nanopores mimicking the β-barrel architecture of proteins, with a central hydrophobic constriction region (a ring of leucine residues sandwiched between two rings of valine residues) giving it an 'hourglass' shape, maintained ssDNA in an extended conformation during translocation under applied electric field. Removing this constriction region not only resulted in DNA coiling but also a faster translocation rate; thus, the hydrophobic constriction region also influenced the rate of DNA translocation [122].

1.3.4 Sequencing DNA homonucleotides: Dual constriction nanopores

DNA typically contain homonucleotides, which are stretches composed of multiple consecutive nucleotides containing the same base. It is challenging to sequence homonucleotides accurately as the ionic current through the nanopore is uniform during their passage, which makes it difficult to determine the number of nucleotides in the segment directly. Instead, the length of the homonucleotides is estimated using their dwell time. However, the accuracy of this relies on the precision of the algorithms used.

A promising solution is to use nanopores with two constriction regions, which output more sequence information as the translocating DNA sequence is read twice. This concept was first demonstrated using α-hemolysin [50] and more recently shown for sequencing homonucleotides with improved accuracy using the CsgG-CsgF complex [123].

1.3.4.1 CsgG-CsgF complex

CsgG forms an ungated peptide diffusion channel that becomes substrate-specific *via* the binding of soluble accessory proteins during the curli biogenesis [75, 124, 125]. One accessory protein is CsgF, which CsgG secretes into the extracellular milieu where the growing amyloid fibre attaches to the CsgF C terminus during its secretion [75, 124, 125]. However, the CsgF N terminus remains associated with the CsgG β-barrel to form a stable complex. The cryo-EM structure of the CsgG-CsgF complex (PDB 6SI7, 3.4 Å) [123] shows that it creates a channel containing two constriction

regions, one formed by CsgG eyelet loops and another formed by a ring of asparagine (Asn-17) residues in CsgF that is ~ 1.5 nm wide [123].

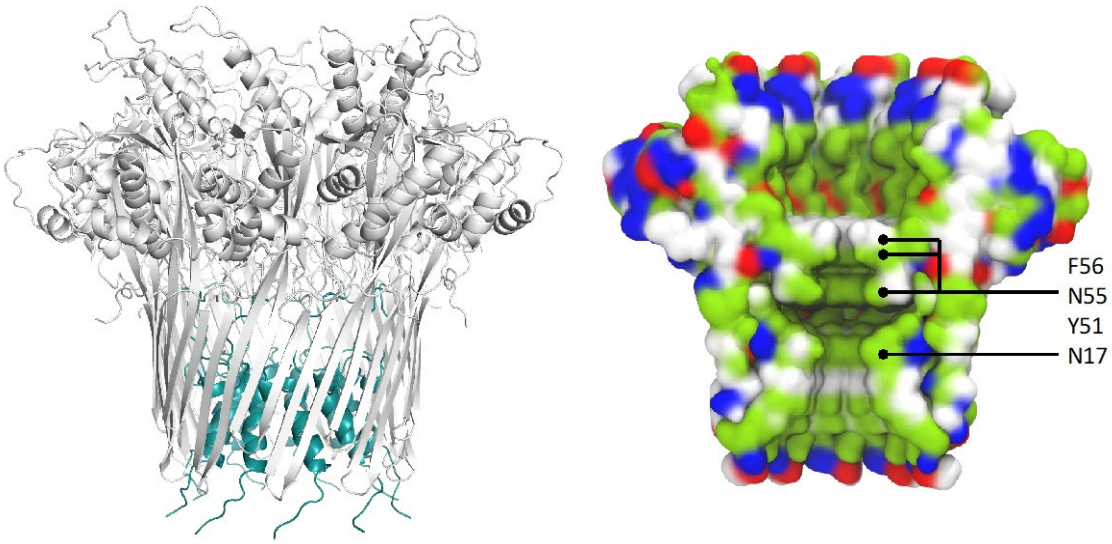

Figure 1.10: The CsgG-CsgF complex (PDB 6SI7, 3.4 Å) is shown from the side, approximately parallel to the membrane, and depicted in ribbon (left) and surface representations (right). CsgG and CsgF are coloured white and teal, respectively (left). The inner surface of the CsgG-CsgF complex is shown, with key residues forming the constriction regions labelled (right). Red: anionic side chains, blue: cationic side chains, green: polar side chains, and white: non-polar side chains.

1.4 Project aims

This thesis aims to elucidate the design principles for optimising nanopores for DNA sequencing using computational methods. To achieve this, Molecular dynamics (MD) simulations have been used to study the translocation of ssDNA through simplistic designed model nanopores and protein nanopores.

Inspired by experimental studies which showed the potential advantages of using nanopores with two constriction regions for DNA sequencing [50, 123], and following on from simulations of model pores with hourglass geometry containing hydrophobic constrictions which maintained DNA in an extended conformation during translocation [122], model nanopores with two hydrophobic constriction regions are designed to investigate the translocation of short and longer ssDNA. The short ssDNA models DNA that is conformationally labile during nanopore sequencing and the longer ssDNA is retained under tension to mimic DNA entering a narrow pore from a less confined vestibule region of proteins during DNA sequencing. The aim is to study the impact of nanopore geometry and its chemical nature on DNA conformation and translocation rate, using

MD simulations. Secondly, CsgG and the CsgG-CsgF complex have been extensively characterised for use as nanopores for DNA sequencing. The translocation of short and longer ssDNA is investigated to explore the impact of single- or dual-constriction geometry on the conformational dynamics of DNA during nanopore sequencing.

Chapter 2 Methods

2.1 Molecular dynamics

The study of biological systems is limited by the temporal and spatial resolution accessible to experimental techniques. Molecular simulations can confirm and expand upon phenomena observed *via* experimental methods as they enable the study of processes occurring at short timescales (< ns) and changes taking place at a molecular level.

Molecular dynamics (MD) is a simulation technique that can investigate changes in structural and thermodynamic properties of molecular systems over time [126, 127]. To perform MD simulations, initial structures of molecules, such as proteins, DNA, and lipids, are obtained from either experimental data or by modelling techniques. Protein structures can be obtained from structural biology techniques such as X-ray crystallography, electron microscopy, or nuclear magnetic resonance (NMR). More recently, protein structures not yet resolved can be obtained *via* AlphaFold 2, an AI system that can predict structures from amino acid sequences using evolutionary relationships between proteins [128, 129]. The protein structure can then be used to create a system, for example, a protein embedded in a lipid membrane and immersed in water, in which molecules are described as collections of point particles.

In MD simulations, the motion of particles in the system is modelled using classical Newtonian mechanics. The evolution of particle positions with respect to time can be obtained as the positions are related to the force exerted on the particle and the system's potential energy. The force acting on each particle can be expressed as the negative gradient of the potential energy of the system,

$$F = -\frac{dU(r^N)}{dr_i} \tag{2.1}$$

where r_i is the position of particle i, and $U(r^N)$ is the potential energy of the system containing N particles. The system's potential energy is obtained from the force field (see section 2.4). The force can be used to calculate the acceleration experienced by the particle. As acceleration is the second derivative of the displacement, the evolution of particle positions with respect to time t can be calculated,

$$\vec{F} = m_i\vec{a}_i = m_i\frac{d^2\vec{r}_i(t)}{dt^2} \tag{2.2}$$

where m_i is the mass of particle i. The two equations, therefore, allow the calculation of the evolution of particle positions concerning time using the potential energy of the system:

$$\vec{F} = m_i \frac{d^2\vec{r}_i(t)}{dt^2} = -\frac{dU(r^N)}{dr_i} \tag{2.3}$$

The workflow of an MD simulation is shown in Figure 2.1. In summary, the positions and velocities of particles at any given time are calculated from their previous positions. The velocities at the start of the simulation are obtained from Maxwell-Boltzmann distribution at a given temperature [130].

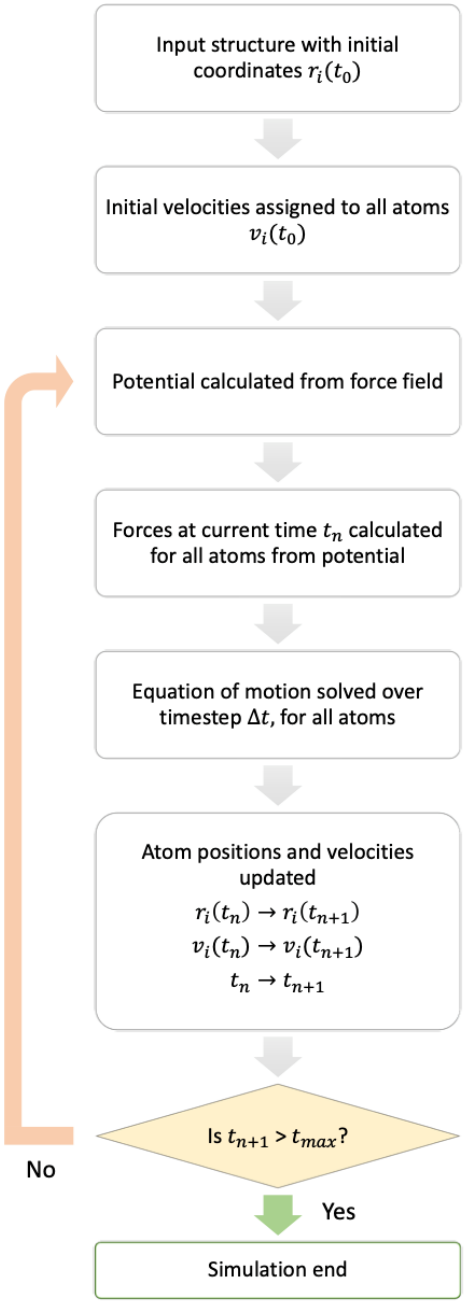

Figure 2.1: Flowchart of MD simulation workflow [130].

2.2 Integrators

MD simulations systems contain a large number of particles and, as the force experienced by any particle is dependent on its position relative to other particles, Newton's equations of motion must be solved simultaneously for all particles at a given time. However, the many-body problem makes it impossible to analytically describe the motion of particles in systems larger than two particles. Therefore, finite difference integration algorithms are used to integrate the equations of motions numerically. Finite difference integration algorithms break down the integration into multiple steps, each separated by a fixed time. The algorithms assume that the positions, velocities, and accelerations can be approximated using a Taylor expansion of Newton's equations of motion for discrete timesteps δt,

$$r_i(t+\delta t) = r_i(t) + v_i(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots \tag{2.4}$$

$$v_i(t+\delta t) = v_i(t) + a_i(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \cdots \tag{2.5}$$

$$a_i(t+\delta t) = a_i(t) + b_i(t)\delta t + \cdots \tag{2.6}$$

where r_i refers to the position, v_i is velocity, and a_i is acceleration, both of which are first and second derivatives of position r_i with respect to time t.

The integrator used in this thesis is the leapfrog integrator [131]. It calculates initial velocities at half-timestep $t + \frac{1}{2}\delta t$ using the velocities at the previous half-timestep $t - \frac{1}{2}\delta t$, and acceleration at the current timestep t,

$$v_i\left(t + \frac{1}{2}\delta t\right) = v_i\left(t - \frac{1}{2}\delta t\right) + a_i(t)\delta t \tag{2.7}$$

where δt is the size of the timestep. The calculated velocities and current position coordinates r_i are then used to obtain the positions at the next timestep $t + \delta t$:

$$r_i(t+\delta t) = r_i(t) + v_i\left(t + \frac{1}{2}\delta t\right)\delta t \tag{2.8}$$

The positions and velocities are not generated simultaneously; they 'leap' over each other during this calculation cycle until the simulation is complete (Figure 2.2).

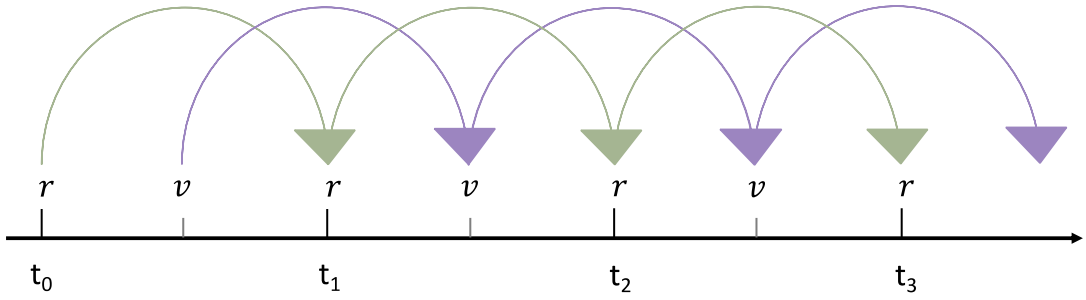

Figure 2.2: Schematic representation of the leap-frog algorithm. The positions and velocities are calculated alternatively and over half-timesteps.

2.3 Constraints

The value of the timestep delta t must be chosen carefully. If the timestep is too small, the MD simulation will cover a limited region of the phase space. Contrariwise, a too large timestep will result in energetically unfavourable collisions between particles and subsequently, the system will become unstable. The timestep, therefore, should be set to be an order of magnitude shorter than the time of the fastest motion in the system, which corresponds to the C-H bond vibration that occurs on the timescale of ~ 10 fs [132] in atomistic simulations.

The rapid bond vibrations are not essential for the systems simulated in this thesis; therefore, they can be discounted using constraint algorithms. Constraint algorithms restrain the bond lengths during the simulation, which enables larger integration timesteps to be used and systems to be simulated efficiently for longer timescales. The SHAKE [133] and LINCS [134] methods are typically used for atomistic simulations to fix bond lengths or angles. In this thesis, the faster LINCS method is used.

2.4 Force Fields

MD simulations can use methods that model molecular systems at various resolutions. The method depends on the system size, which can be as small as a few atoms to whole proteins, and the phenomena to be studied. Modelling atoms with their nuclei and electrons included is chemically accurate but computationally costly, and systems cannot exceed 20, 000 atoms and be simulated for more than a few picoseconds. When investigating long-time biological processes such as conformational changes in proteins, details of the electronic structure are irrelevant; therefore, classical molecular mechanics (MM) is used [132].

In classical MM, atoms are modelled as 'soft balls' with no explicit representation of electrons. A classical MM force field describes a set of potential energy functions and associated parameters derived from experimental data, and quantum mechanical calculations used to determine the potential energy of particles in the MD simulation system. The total potential energy of a particle is the sum of all bonded and non-bonded potential energy terms, described further in sections 2.4.1 and 2.4.2.

The resolution of force fields can vary; for example, each atom may be represented explicitly or united-atom, in which methyl groups containing carbon and non-polar hydrogen atoms are represented as one interaction centre, but all other atoms are represented explicitly, or coarse-grained, in which groups of heavy atoms are combined into particles [135]. The force fields implemented in this thesis are the GROMOS 53A6 united-atom force field [136] and the CHARMM36m all-atom force field [137]. AMBER is another popular force field for biological systems, but it was not used because it has been previously shown to be potentially unsuitable for representing the ssDNA structure [81]. More recently, the DES-Amber force field was developed with greatly improved descriptions of proteins and nucleic acids such as RNA and DNA, especially the ssDNA structures [138]. Although relevant for the DNA-protein systems, DES-Amber is not used in this work as it was unavailable during the time in which the research presented in this thesis was conducted.

2.4.1 Bonded Potential Energy Terms

The bonded potential energy (U_{bonded}) includes terms for bond stretching, bond bending, and rotation about dihedral angles,

$$U_{bonded} = \sum_{stretching} \frac{k_r}{2}(r - r_0)^2 + \sum_{bending} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \sum_{improper} \frac{k_\phi}{2}(\phi - \phi_0)^2 + \sum_{proper} k_\chi[1 + \cos(n\chi - \sigma)] \quad (2.9)$$

where k is the force constant, r is the distance between the two bonded atoms, θ is the angle formed between two bond vectors, ϕ is the improper torsion angle; n and σ are the periodicity and the phase of proper torsion angle χ.

The stretching and bending terms are modelled as harmonic oscillators centred on equilibrium values r_0 and θ_0, with displacement r and θ, respectively, and spring constants k_r and k_θ, respectively, which determine the strength of the terms. The improper term enforces the planarity of functional groups such as aromatic rings by applying a harmonic potential to control

the torsion angle ϕ between two planes. The proper term controls the rotation of two atom groups around a covalent bond. It is modelled as a sum of cosine functions, as multiple minima are associated with the proper torsion angle χ. The bonded potential energy terms are illustrated in Figure 2.3.

2.4.2 Non-bonded Potential Energy Terms

Non-bonded terms describe the pair-wise sum of the energies of all possible intermolecular interactions between atoms not connected *via* bonds or are separated by three or more bonds in the same molecule. It comprises electrostatic and van der Waals (vdW) terms, which are represented by the Coulomb and Lennard-Jones (LJ) potentials [139], respectively, summed over N atoms,

$$U_{i,nonbonded} = U_{i,elec} + U_{i,vdw} = \sum_{j}^{N} \frac{q_i q_j}{4\pi \in_0 r_{ij}} + 4\varepsilon \sum_{j}^{N} \left(\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right) \quad (2.10)$$

where i and j are two atoms separated by a distance r, q is the atom's partial charge, $\in_0$ is the permittivity of free space, ε is the well depth of the LJ potential energy function, and σ is the distance between two atoms at which the LJ potential energy is zero.

The electrostatic term calculates the potential energy between the two atoms, i and j, with partial charges q_i and q_j. The vdW term approximates the potential energy between the two neutral atoms. It includes the strong repulsive term (r^{-12}), which describes the repulsive force arising when the electron orbitals of the two atoms overlap (Pauli repulsion), and the attractive term (r^{-6}), which describes the weak attractive force between the two atoms arising from short-ranged instantaneous dipoles (London dispersion). The well depth of the potential energy function, $\mathcal{E}$, sets the strength of the interaction [132] (Figure 2.4).

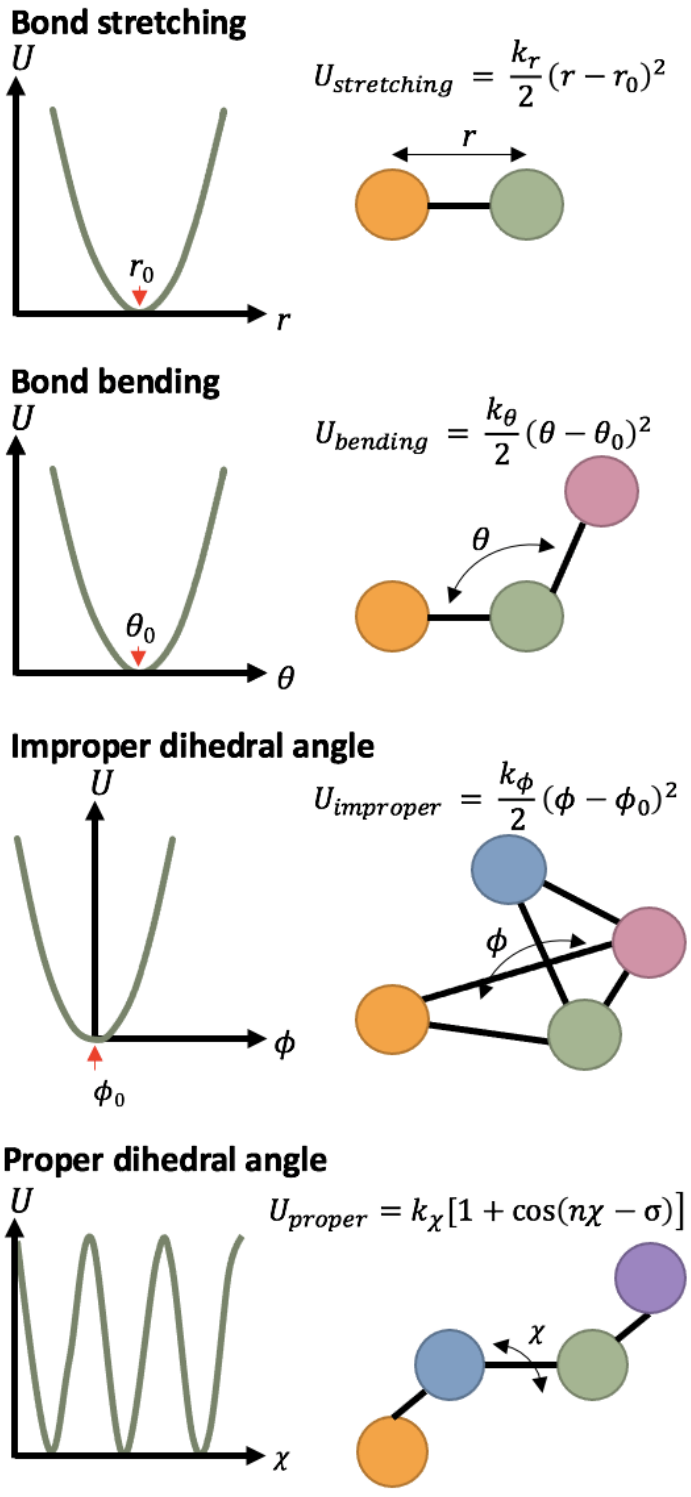

Figure 2.3: Bonded potential energy terms.

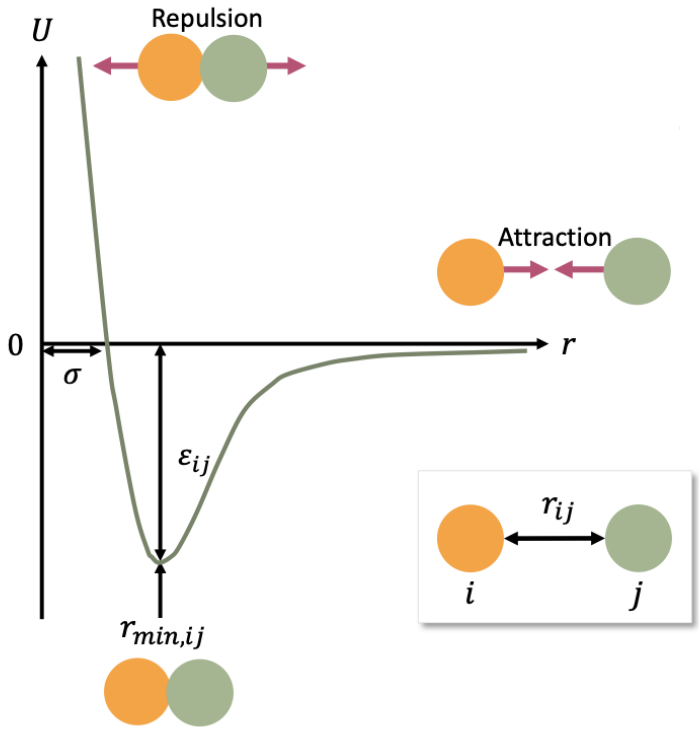

Figure 2.4: Form of the Lennard-Jones Potential describing van der Waals interactions between two atoms.

2.5 Periodic boundary conditions

MD simulations are limited to a range of hindered to millions of atoms in a simulation cell, a fraction of the macroscopic experimental system. Applying periodic boundary conditions (PBC) allows the small simulation system to mimic a macroscopic system. PBC takes the simulation cell and forms a lattice of infinite cell copies, known as periodic images. Any particles drifting outside the simulation cell into a periodic image re-enter the cell from a different periodic image. This ensures that the number of particles remain constant during the simulation (Figure 2.5).

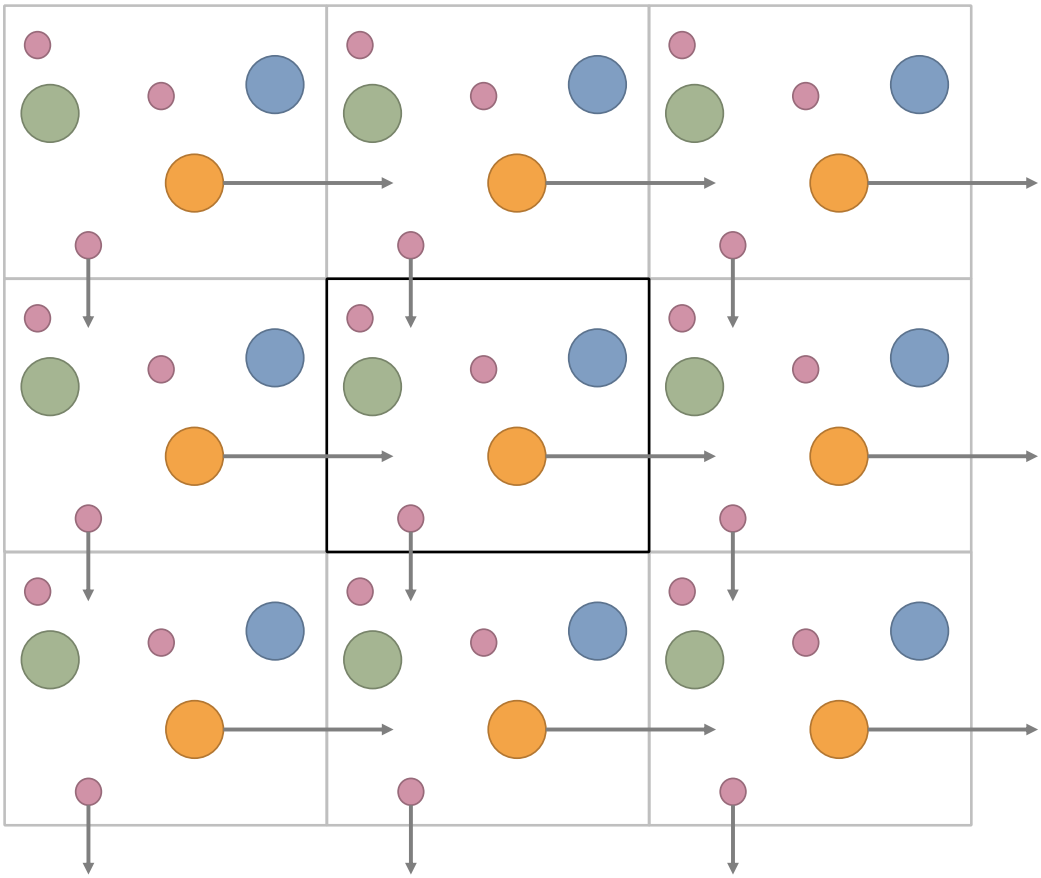

Figure 2.5: Schematic modelling a simulation system when periodic boundary conditions are applied. The initial simulation cell is highlighted in grey, surrounded by periodic images of itself. Particles can leave the simulation cell and re-enter from the opposite side of the cell.

2.5.1 Cut-offs

A consequence of applying periodic boundary conditions is that a particle in a periodic cell forms non-bonded interactions with other particles in the same cell and the surrounding periodic images. Additionally, as mentioned earlier, the non-bonded terms for a particle are calculated for all pairs in the system. Together, this requires an infinite number of calculations to be performed for the non-bonded potential energy term, which is both computationally expensive and unfeasible. Therefore, a cut-off is applied so that longer inter-particle distances are not regarded as non-bonded interactions. This also prevents the calculation of the potential between a particle and its periodic image unless the cut-off distance is greater than half the length of the periodic cell. This method works well for vdW interactions; however, a cut-off excludes electrostatic interactions at longer inter-particle distances.

One of the models developed for the efficient calculation of long-range interactions is the Ewald summation method, which splits the interaction into a short-range contribution (cut-off range) and a long-range contribution (beyond the cut-off range). The short-range contribution is calculated using a convergence function, and the long-range contribution is approximated using a Fourier transform, resulting in two rapidly convergent sums that are more accurate than a single calculation. The Ewald summation is computationally expensive for large systems as the calculation is of order N^2. In the Particle Mesh Ewald method (PME), the Fourier transform in the Ewald summation method is replaced with the particle-mesh method, which has a faster algorithm $N \log N$, for approximating the long-range contributions [140, 141]. PME is comparatively more efficient and accurately calculates long-range interactions in large system sizes. Hence, the PME method is used in most atomistic force fields.

2.6 Statistical ensembles

The microscopic motion of atoms in an MD simulation is related to the properties of a macroscopic systems *via* statistical mechanics. In statistical mechanics, an ensemble is a collection of many virtual microscopic states of a system that are macroscopically identical. Although the microscopic states differ in properties such as temperature, pressure, and volume, the average properties are constant. An ensemble can be considered an experiment performed multiple times under the same macroscopic conditions, where the system has different microscopic properties in each repetition.

In an MD simulation, the system adopts multiple microscopic states over time. Over a sufficiently long time, the simulation's properties' averages correspond to the ensemble averages, as stated by the ergodic hypothesis [142]. Therefore, the thermodynamic properties of a system can be obtained by averaging the properties across its thermodynamic ensemble.

Macroscopic constraints corresponding to ensembles with distinct statistical characteristics can be implemented in MD simulations. In simulations of biological systems containing lipid membranes, the Isobaric-Isothermal ensemble (NPT) is implemented, in which the number of particles (N), the system pressure (P), and the temperature (T) are maintained at constant values. The canonical ensemble (NVT) is implemented to prevent fluctuations in the system volume, in which the number of atoms, the system volume (V), and temperature remain constant.

2.6.1 Pressure and temperature control

The conditions described by the implemented ensemble are regulated using a thermostat and barostat for temperature and pressure, respectively. Thermostats regulate the temperature of the system by scaling the velocities, as the temperature is related to the velocity of the particles,

$$T = \frac{1}{3k_B N}\sum_{i}^{N} m_i {v_i}^2 \tag{2.11}$$

where T is the temperature, k_B is the Boltzmann constant, N is the number of particles, and m_i and v_i are the mass and the velocity of the particle i respectively. A commonly used thermostat is the Berendsen thermostat, which scales the velocity of each particle proportionately to the change in the temperature from the reference value [143]. A drawback of this thermostat is that it does not maintain the ensemble conditions. Based on the Berendsen thermostat, the velocity-rescale thermostat maintains the ensemble due to an additional stochastic term which scales the velocities using kinetic energy taken from the Maxwell-Boltzmann distribution [144].

Barostats regulate the pressure of the system by adjusting the system volume, as pressure is related to volume,

$$P = \frac{Nk_B T}{V} + \frac{1}{3V}\sum_{i} r_i \cdot f_i \tag{2.12}$$

where P is the pressure, V is the volume of the simulation cell, r_i is the position of particle i, and is the force acting on particle i. As with thermostats, the system volume is scaled proportionately to the change in pressure from the reference value. The volume is adjusted by altering the dimensions of the simulation cell. Barostats can be implemented as isotropic, which scales all the cell dimensions equally, and semi-isotropic, which scales the cell in x and y dimensions only. Semi-isotropic barostats are typically used for systems containing lipid membranes, as the stress parallel to the membrane surface greatly differs from the stress perpendicularly. The Parinello-Rahman barostat is a commonly used barostat that can scale each dimension of the simulation cell separately [145].

2.7 Steered molecular dynamics

Steered molecular dynamics is an MD derivative inspired by single-molecule pulling experiments. Systems are driven to evolve towards non-equilibrium states in timescales shorter than in MD simulations [146]. In a steered MD simulation, a virtual harmonic spring is added to two groups of the system, which can be the centre of mass of an atom, molecule, or group of molecules. A force

is applied to the harmonic spring along a defined vector to overcome energy barriers and progress along processes such as protein unfolding, binding or unbinding of molecules from their protein targets, and the transportation of molecules through membrane channels. The force applied can be either a constant value or fluctuate, resulting in the harmonic spring moving with a constant velocity in a defined direction, in constant velocity (CV) or constant force (CF) steered MD, respectively. The force experienced by the pulled group is:

$$F = -k(x - vt) \tag{2.13}$$

where k is the spring constant, x is the displacement of the pulled group from its original position, v is the velocity, and t is time.

In CF simulations, the duration of the conformational changes varies according to their associated energy barrier, as the force cannot be increased to overcome high energy barriers. CF simulations are more accurate for computational elasticity measurements than CV simulations [147, 148]. However, multiple CF simulations with different force strengths are required to generate a stress-strain curve compared to CV, making CF steered MD computationally expensive [148].

In CV simulations, the fluctuating pulling force can be measured over time and thus can be correlated with the observed conformational changes of the system. For example, a higher force is applied to overcome energy barriers associated with pulling the ligand away from a region in the protein where it forms several interactions with the protein residues. It is often computationally unfeasible to match pulling velocities to experiments; hence faster velocities are typically used in steered MD simulations. A side effect of this is that the required force is overestimated compared to experiments. Although the forces obtained cannot be compared directly with experimental data, CV simulations can be useful for analysing details of interactions at an atomic level and for relative comparisons e.g., comparing the binding of different ligands to a given protein [149].

The method implemented in this thesis is CV steered MD.

Chapter 3 DNA translocation through hydrophobic nanopores with two constriction regions

3.1 Introduction

The impact of the chemical nature of nanopores on the rate of DNA translocation and the conformations adopted by the translocating DNA has been extensively studied [29, 67, 71, 81, 88, 122, 123, 150-152]. During DNA sequencing, slow translocation of DNA retained in a largely linear conformation is ideal, because it enables a distinct current reading to be obtained for each base and ensures that the bases pass through the nanopore sequentially [81].

The introduction of basic residues in the lumen of α-hemolysin has been experimentally shown to result in a slower rate of ssDNA translocation compared to the unmodified protein [71, 150]. Similarly, the translocation rate through WT aerolysin was observed to be ~ 3 orders of magnitude slower than the average translocation rate of nucleotides in α-haemolysin due to basic residues predominantly lining the pore lumen [67]. Investigation *via* MD simulations of simplified models of the α-hemolysin nanopore region showed that DNA translocation is slowed down due to the strand transiently tethering to the pore *via* strong electrostatic interactions formed between the acidic phosphate groups of the DNA backbone and the basic residues of the pore lumen [88]. However, these interactions resulted in the short DNA strand forming undesirable conformations during translocation, which could impact the accuracy of DNA sequencing [81]. Following this, model nanopores mimicking 14-stranded β-barrel architecture were used to investigate DNA translocation through hourglass-shaped pores with a central hydrophobic constriction region [122]. DNA was slow to enter these pores; however, once inside, the strand was maintained in a linear conformation during translocation, compared to the absence of the hydrophobic constriction (Figure 3.1).

Taking this forward, and inspired by experimental work showing the potential advantages of nanopores with two constriction regions for DNA sequencing [123, 151], four model nanopores with β-barrel architecture were designed with two hydrophobic constriction regions. The model nanopores differ from protein nanopores as they lack the large vestibule region through which DNA enters. However, they resemble solid-state nanopores such as those made from graphene, as their outer surface is cylindrical [29-31]. In this study, DNA translocation through the model nanopores is investigated to ascertain the impact of the chemical nature and geometry of

nanopores on DNA translocation rate and conformation, with the aim of achieving slow translocation of DNA retained in a largely linear conformation.

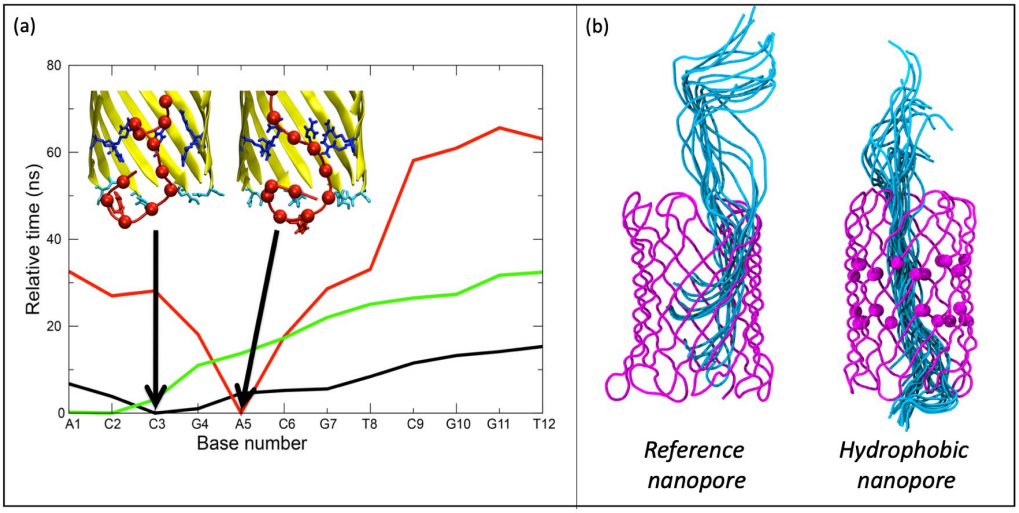

Figure 3.1: (a) The presence of basic residues in model α-hemolysin nanopore (arginine and asparagine are shown in blue and cyan respectively) resulted in the DNA nucleotides (red) exiting the nanopore non-sequentially. (b) DNA (cyan) is maintained in a linear conformation through model nanopore with a central hydrophobic constriction. The absence of the constriction (in reference nanopore) resulted in DNA coiling. Image in panel (a) is reproduced from Guy et al. [81], and in panel (b) is from Haynes et al. [122].

Notes

This chapter is based on the publication titled "Translocation of flexible and tensioned ssDNA through *in silico* designed hydrophobic nanopores with two constrictions", published in the RSC Nanoscale Journal [153]. The figures used in this chapter are reproduced from ref. [153] with permission from the Royal Society of Chemistry.

The parameters for generating continuous tensioned ssDNA were obtained from Dr Richard Manara. The Density Functional Theory (DFT) calculations in this chapter were performed by Dr Sophie Mader.

3.2 Methods

3.2.1 Generation of model nanopores

The model nanopores are based on the peptide backbone of pores first reported by Sansom et al. [154], which consist of antiparallel β-sheets formed by alanine residues connected with glycine or serine loops (for 14- or 16-stranded pores, respectively) and with a band of tryptophan residues anchoring the pore in lipid bilayer. The model nanopores were constructed using Modeller [155] and PyMOL [156].

3.2.2 Generation of ssDNA

DNA translocation through the model nanopores was investigated for two scenarios: short and flexible ssDNA (12 nucleotides in length), and a longer tensioned ssDNA retained in a linear conformation. The models of ssDNA with polyA sequence were generated using the 3DNA package [157]. To generate tensioned ssDNA, a 42-nucleotide long strand was stretched by applying an electric field of 0.1 V nm^{-1} for 10 ns, with positional restraints with force constant 1000 kJ mol^{-1} nm^2 applied to the 5' terminal nucleotide. Following this, the 5' and 3' terminal nucleotides were removed, and bond definitions were introduced between the new terminal nucleotides of the resultant 40-nucleotide long strand across periodic boundaries to generate a continuous strand of ssDNA under tension (Figure 3.2).

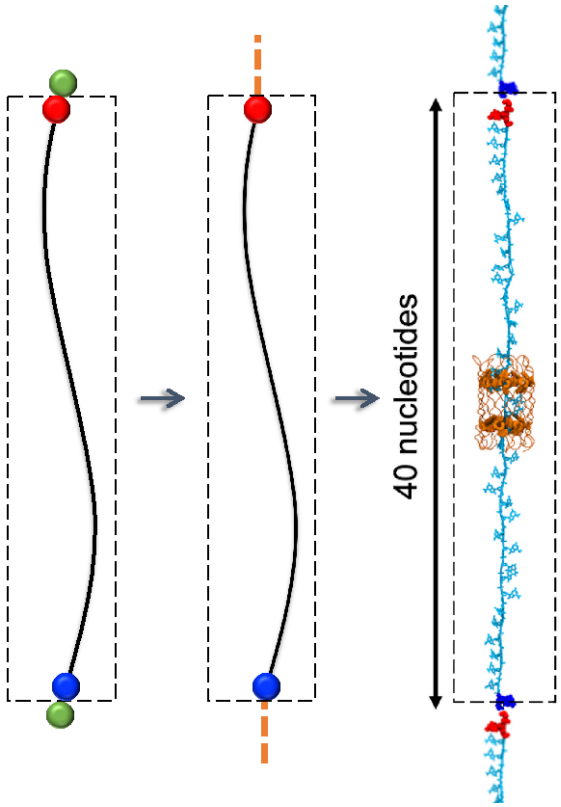

Figure 3.2: Schematic overview of generating continuous tensioned ssDNA. The 5' and 3' terminal nucleotides of long ssDNA (green circles) were removed, and bond definitions were introduced between the two ends of the 40-nucleotide ssDNA across periodic boundaries (orange dashed lines). The final system with continuous tensioned ssDNA threaded through the model nanopore is shown.

3.2.3 Simulation protocol and analyses

The model nanopores were embedded in a membrane composed of 505 1,2-dipalmitoyl- sn-glycero-3-phosphocholine (DPPC) lipids, and the systems were immersed in 1.0 M NaCl electrolyte solution using the simple point charge (SPC) water model [158]. Additional ions were added to neutralise the system. Simulations were performed using GROMACS 2018.3 and the GROMOS 53A6 force field with added Berger lipid definitions [159-161]. To prevent the movement of DPPC lipids in the Z dimension, positional restraints were applied to the phosphate moieties with a force constant of 1000 kJ mol^{-1} nm^2. Positional restraints were also applied to the Cα atoms of the model nanopores in systems with continuous tensioned ssDNA, with a force constant of 1000 kJ mol^{-1} nm^2. The Particle Mesh Ewald (PME) method was used to treat long-range electrostatic interactions with a short-range cut-off of 1.4 nm [140]. The van der Waals interactions were curtailed at 1.4 nm, with long-range dispersion corrections applied to the pressure and energy.

The temperature was sustained at 310 K using the velocity-rescale thermostat [144] and a coupling constant of 0.1 ps.

Systems with continuous tensioned ssDNA were simulated in the NVT ensemble, whereas systems with short ssDNA were simulated in the NPT ensemble. The pressure was maintained semi-isotropically using the Parrinello−Rahman barostat at 1 bar and a time constant of 1 ps [145]. The lengths of all bonds were constrained using the LINCS algorithm enabling a timestep of 2 fs [134]. An electric field was imposed by a constant voltage drop across the simulation cell in the Z dimension. When a 'reversed' electric field was imposed, charged particles experienced a force in the direction opposite to the original electric field. The periodic boundary conditions were applied to all systems in three dimensions, as in previous studies [81, 88, 89, 162]. Replicate simulations were initiated using coordinates extracted at random time points from the last 100 ps of the equilibration run. The initial coordinates and velocities differ for each replicate simulation for the systems. Systems with short ssDNA contained ~ 205,600 atoms, while larger systems with long ssDNA contained ~ 355,000 atoms. The short ssDNA systems were simulated with the performance of ~ 24 ns/day, and the long ssDNA were simulated with the performance of ~ 15 ns/day.

The 14-stranded nanopore systems without ssDNA were also simulated using the CHARMM36m force field [137] and the TIP3P water model [163]. These simulations are discussed in detail in section 3.3.5.

Analyses were performed using GROMACS utilities and locally written code. Pore radius profiles of model nanopores were calculated as an average across the simulations for a given pore using HOLE [164]. The molecular graphics images were generated using the Visual Molecular Dynamics (VMD) package [165]. The interaction energies for the short ssDNA systems were calculated using GROMACS utilities. The interaction energies for continuous tensioned ssDNA systems were computed using density functional theory (DFT) implemented within the ONETEP code [166].

3.2.4 Density Functional Theory Calculations

The interaction energies were calculated for one of the simulations of systems with continuous tensioned ssDNA, during which a DNA nucleotide is captured by a 'gate' formed by two residues in the constriction region. Three snapshots were taken, each for the following scenarios: the nucleotide is trapped in the gate; the conformation of the residues has changed, resulting in the gate opening; the nucleotide has moved out of the gate. The coordinates retained were of the DNA nucleotide and the sidechains of the two residues. All the carbonyl and amine groups were replaced with hydrogen atoms, and the missing hydrogen atoms (in the united-atom systems) were added

using the CHARMM-GUI web server PDB Reader tool [167, 168]. The resultant systems used for calculations were of net charge of –1 *e* due to the phosphate moiety of the nucleotide.

Linear-scaling Density Functional Theory (DFT) was used for geometry optimisation of the three snapshots using the ONETEP program [166, 169]. The Perdew-Burke-Ernzerhof exchange-correlation functional [170] were used with the D2 Grimme dispersion correction (PBE-D2) [171]. Open boundary conditions were used for a 125 nm^3 cubic simulation cell *via* the multigrid Poisson-Boltzmann solver. Norm-conserving pseudopotentials were used, and the psinc basis set, equivalent to a plane wave basis set, was employed with a kinetic energy cut-off of 800 eV. The radii of 8.0 Bohr were used for the nonorthogonal generalised Wannier functions (NGWFs) [172]. No cut-off was applied to the density kernel.

The interaction energies were calculated as the difference between the total DFT energy of the complex (the DNA nucleotide containing 35 atoms and the two residue sidechains containing 36 atoms) and the energy of the DNA nucleotide and the residue sidechains,

$$E_{int} = E_{(nucleotide-residue)} - E_{(nucleotide)} - E_{(residue)} \quad (3.1)$$

where E_{int} is the interaction energy, $E_{(nucleotide)}$ is the DFT energy of the DNA nucleotide, and $E_{(residue)}$ is the DFT energy of the residue sidechains.

3.3 Results and Discussion

3.3.1 Model nanopores in an applied electric field

The model nanopores mimicking protein β-barrel are composed of either 14 or 16 antiparallel strands and contain 2 hydrophobic constriction regions. The 14-stranded β-barrel model nanopores are **14LLx2**, in which each constriction region is composed of two rings of leucine (LEU) residues, and **14Fx2**, in which each constriction region is composed of one ring of phenylalanine (PHE) residues. The 16-stranded β-barrel model nanopores are **16FFx2**, in which each constriction region is composed of two rings of PHE residues, and **16WWx2**, in which each constriction region is composed of two rings of tryptophan (TRP) residues (Figure 3.3).

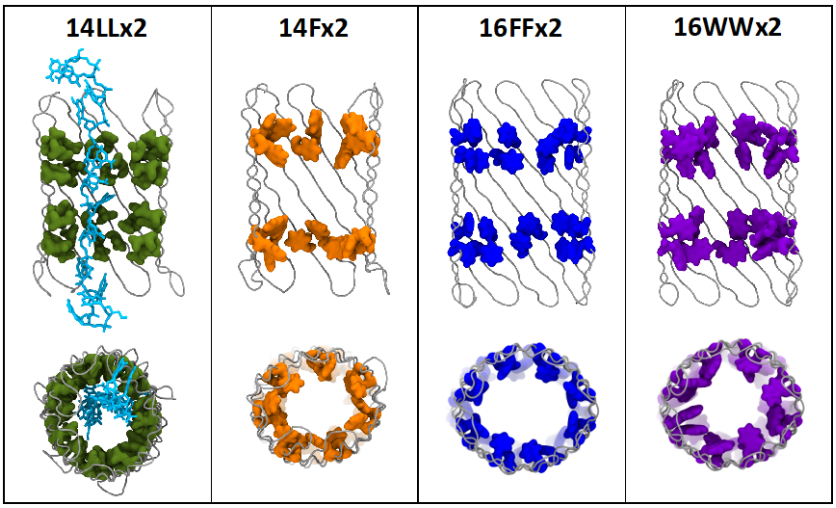

Figure 3.3: The model nanopores studied are shown as cross-sectional and birds-eye views, with the residues forming the constriction regions shown in surface representation. 14LLx2 is shown with short ssDNA (cyan) threaded.

The model nanopores were simulated under an applied electric field of 0.15 V nm^{-1}. The dimensions of the model nanopores fluctuated to a small degree during the 20 ns simulations due to the flexibility of the sidechains (Figure 3.4). The constriction regions of the 14-stranded pores are narrower than the 16-stranded pores, which is expected as the 14-stranded β-barrel is narrower. The radii of the constriction regions are: ~ 0.50-0.55 nm in 14Fx2, ~ 0.58-0.60 nm in 14LLx2, ~ 0.62-0.65 nm in 16WWx2, and ~ 0.75-0.80 nm in 16FFx2. The flux of water and ions through the model nanopores correlated with the pore radius, with lower flux observed through the 14-stranded pores and higher flux through the 16-stranded pores (Table 3.1). Interesting, there was no water and ion flux observed through 14LLx2; although the 14LLx2 constriction regions are wider than 14Fx2, the separation between the constriction regions in 14Fx2 is wider and longer which may be the cause of the higher flux through the pore than 14LLx2.

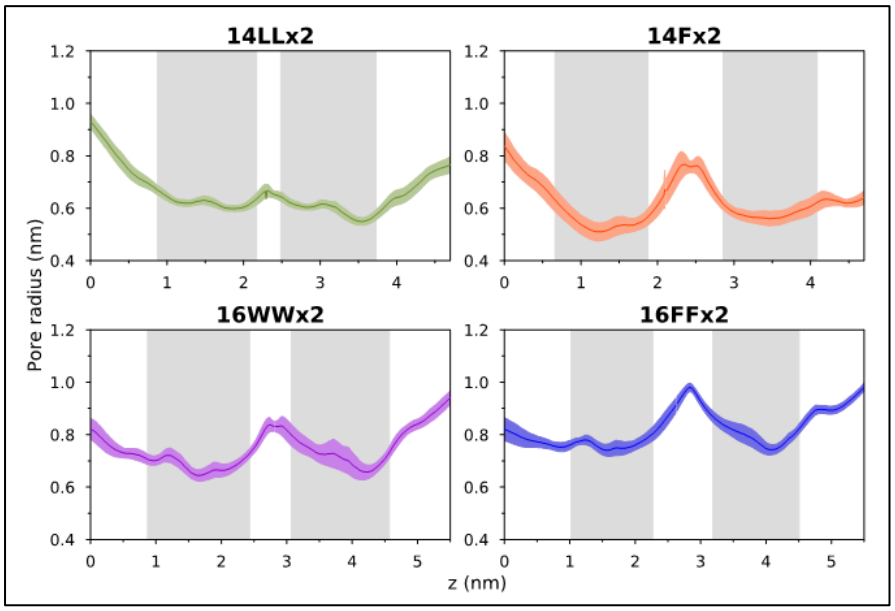

Figure 3.4: Average pore radius profiles of the model nanopores under an applied electric field of 0.15 V nm^{-1}, with standard deviations shown. The constriction regions for each pore are shaded in grey.

Table 3.1. The average mean flux of water and ions through the model nanopores under an electric field of 0.15 V nm^{-1}.

Model nanopore	**Average mean flux (ns^{-1}) ± SD**		
	Water	Na^+	Cl^-
14LLx2	0	0	0
14Fx2	24 ± 3.0	1.2 ± 0.4	1.9 ± 0.2
16FFx2	55 ± 0.8	3.0 ± 0.8	5.2 ± 0.4
16WWx2	45 ± 4.7	1.3 ± 0.5	4.6 ± 0.1

Next, the translocation of DNA through the model nanopores was investigated for two scenarios: short flexible ssDNA of finite length, and a longer continuous tensioned ssDNA of 'infinite' length due to the strand being bonded to itself across periodic boundaries. Both DNA strands have a

polyA sequence. For all model nanopores, constriction 1 refers to the constriction nearest to the pore entrance, and constriction 2 refers to the constriction nearest to the pore exit (Figure 3.5).

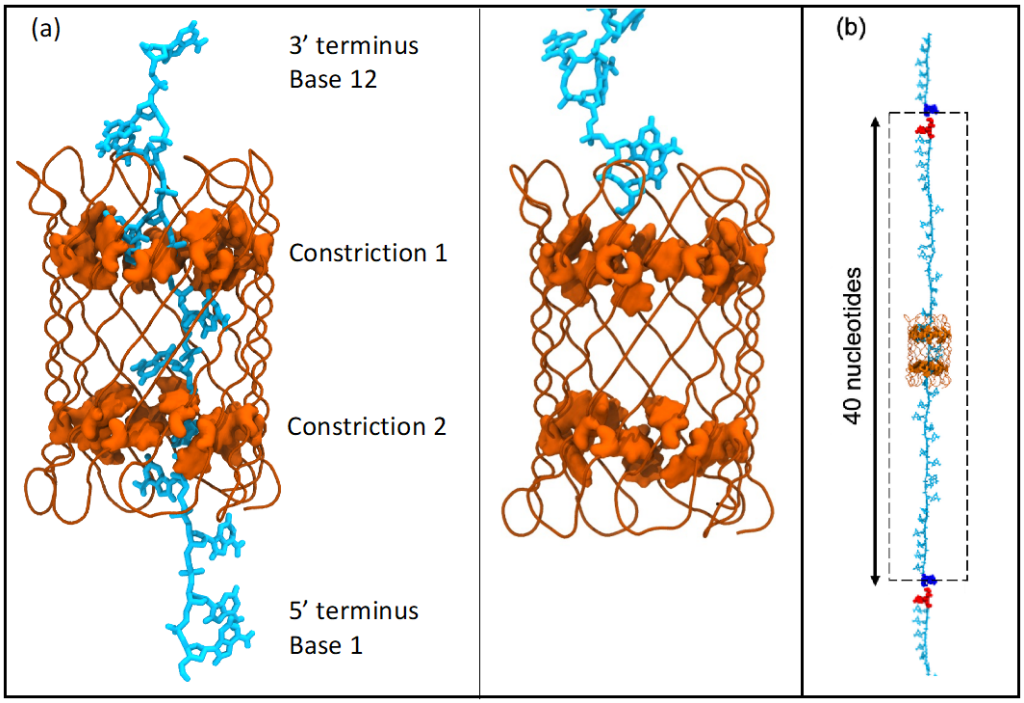

Figure 3.5: The initial position of ssDNA (cyan) in model nanopores (14Fx2 shown in orange) for the translocation studies. (a) The translocation of short flexible ssDNA was investigated for two scenarios: DNA strand pre-threaded through the pore (left), or the DNA 5' terminal nucleotide located at the pore entrance (right). The naming conventions adopted throughout this chapter are labelled. (b) Long continuous tensioned ssDNA, with the terminal nucleotides (red and blue) bonded across the periodic boundaries (dashed lines).

3.3.2 Entry of short ssDNA into model nanopores under an applied electric field

The model nanopores were simulated with the short ssDNA 5' terminal nucleotide initially placed at the pore entrance (Figure 3.5). DNA did not enter the 14-stranded pores (4 x 14Fx2 and 4 x 14LLx2) in all eight simulations under an applied electric field of 0.15 V nm^{-1}. This concurs with a previous study, which showed that DNA does not enter the 14-stranded hydrophobic pores with a single constriction region in absence of an electric field, and rarely under an applied electric field [122]. In contrast, DNA entered the 16-stranded pores (4 x 16FFx2 and 4 x 16WWx2) in all eight simulations (Figure 3.6). These simulations indicate a barrier to DNA entry into 14-stranded pores

and not the wider 16-stranded pores. This could be due to the entropic penalty associated with DNA moving into a more restricted geometry. DNA entry into the 14-stranded pores could be facilitated by incorporating a cap or vestibule region, like present in α-hemolysin and aerolysin proteins in which DNA is known to enter the 14-stranded β-barrel region.

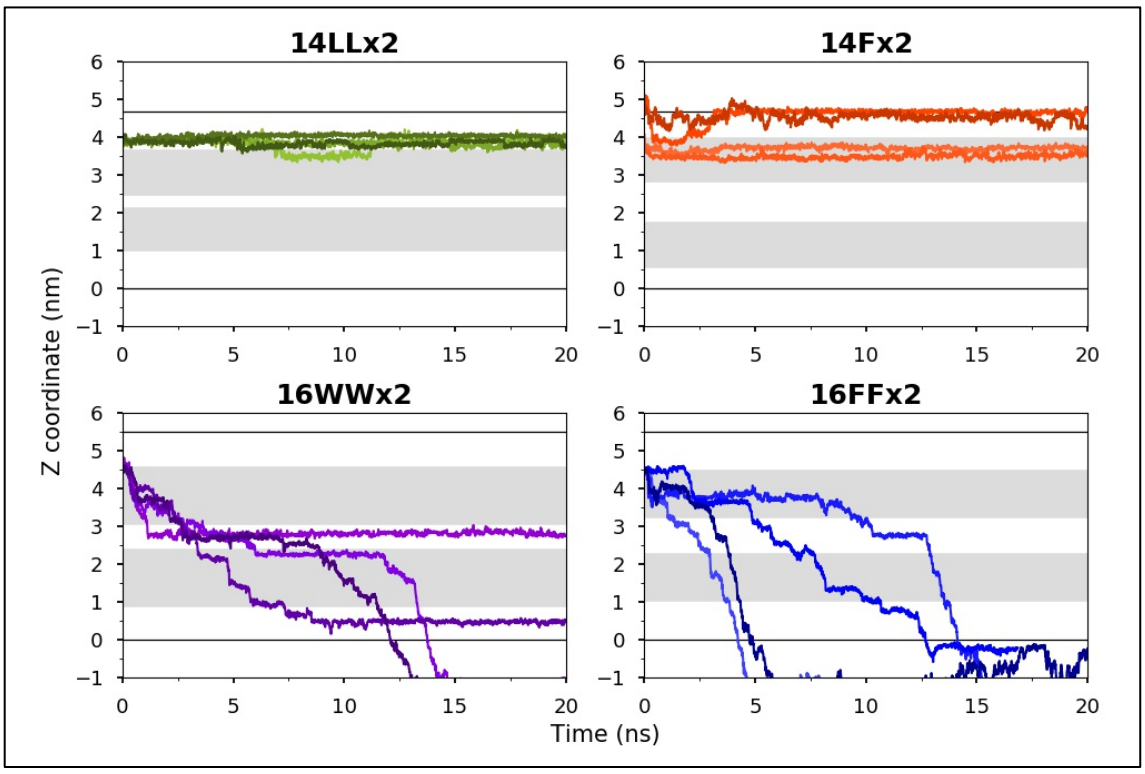

Figure 3.6: DNA translocation through the model nanopores, with short ssDNA initially at the pore entrance, is measured as the Z coordinate of the center of mass of the 5' terminal nucleotide over time, in four simulations for each pore. The constriction regions for each pore are shaded in grey, and the mouths of the pores are represented by solid lines.

3.3.3 Translocation of short ssDNA through model nanopores under an applied electric field

Next, the translocation of short ssDNA was investigated when the DNA strand is initially pre-threaded through the model nanopores (Figure 3.5) under an applied electric field of 0.15 V nm^{-1}. The centre of mass movement of the 3' terminal nucleotide as a function of time was calculated to characterise the rate of DNA translocation through the pores (Figure 3.7). In simulations of 14LLx2, DNA exited constriction 2 of the pore by 4-17 ns in all eight simulations and exited the

pore by 8 ns and 15 ns in two simulations. DNA was unable to exit 14LLx2 in six simulations due to the nucleotides in the pore exit interacting with TRP residues (Figure 3.8). Translocation was comparatively slower through 14Fx2; DNA remained threaded through both constriction regions in two simulations and remained threaded through constriction 2 only in six simulations. For the 16-stranded pores, DNA remained threaded through constriction 2 of 16WWx2 and 16FFx2, in eight and five simulations, respectively (Figure 3.7). DNA was unable to exit constriction 2 of 16WWx2 even when the simulations were extended to 40 ns (Figure 3.9). In contrast, DNA translocation was more variable through 16FFx2, with DNA exiting constriction 2 within 10 and 15 ns in two simulations. In summary, DNA was retained in constriction region(s) of model nanopores containing aromatic residues in 22 of 24 independent simulations (and exited the pore in two simulations of 16FFx2).

The translocation rate of DNA was calculated as the number of nucleotides exiting constriction 2 of the model nanopore over time in eight independent simulations (Figure 3.10). The rate is affected by DNA translocation being halted due to either one end of the strand physically occluding the pore exit or nucleotides at the pore entrance interacting with TRP residues. It is important to note that this is not influenced by the constriction regions of the pores; it is a function of the DNA strand being short and flexible. For example, the translocation rate through 14LLx2 is ~ 0.40 nucleotides ns^{-1} at 20 ns, but the rate is faster during ~ 5-15 ns. It is clear to see that overall, the DNA translocation rate is slower through 14Fx2 and 16WWx2 compared to through 14LLx2 and 16FFx2. The translocation rate at 20 ns is ~ 0.22 nucleotides ns^{-1} through 14Fx2, ~ 0.33 nucleotides ns^{-1} through 16FFx2, and ~ 0.23 nucleotides ns^{-1} through 16WWx2.

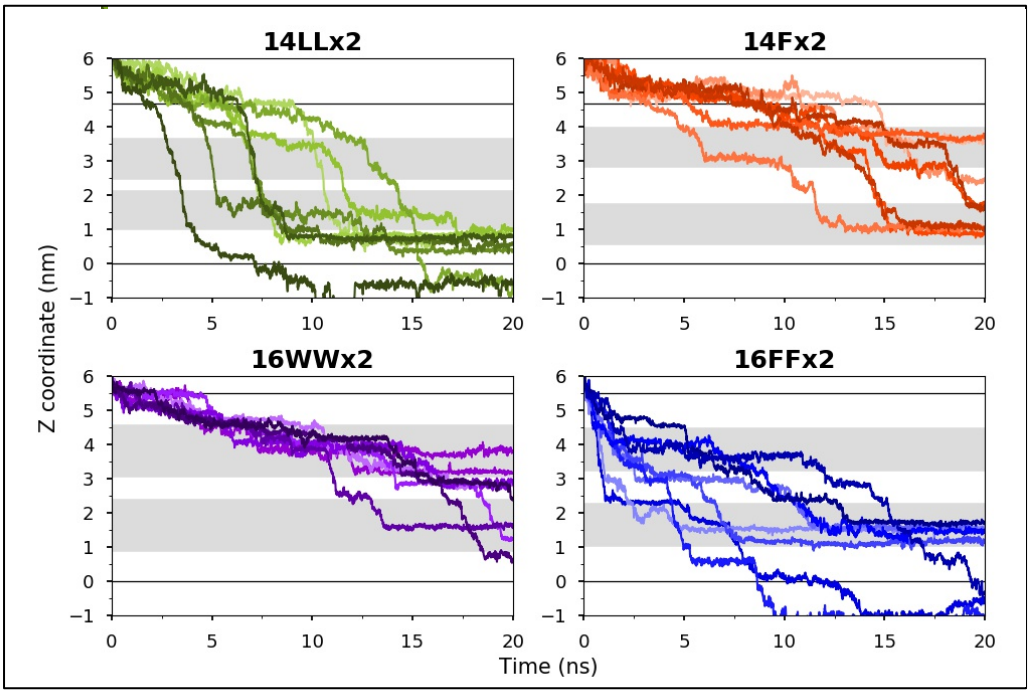

Figure 3.7: DNA translocation through the model nanopores, with short ssDNA pre-threaded through the pore, is measured as the Z coordinate of the centre of mass of the 3' terminal nucleotide over time in eight simulations for each pore. The constriction regions for each pore are shaded in grey, and solid lines represent the mouths of the pores.

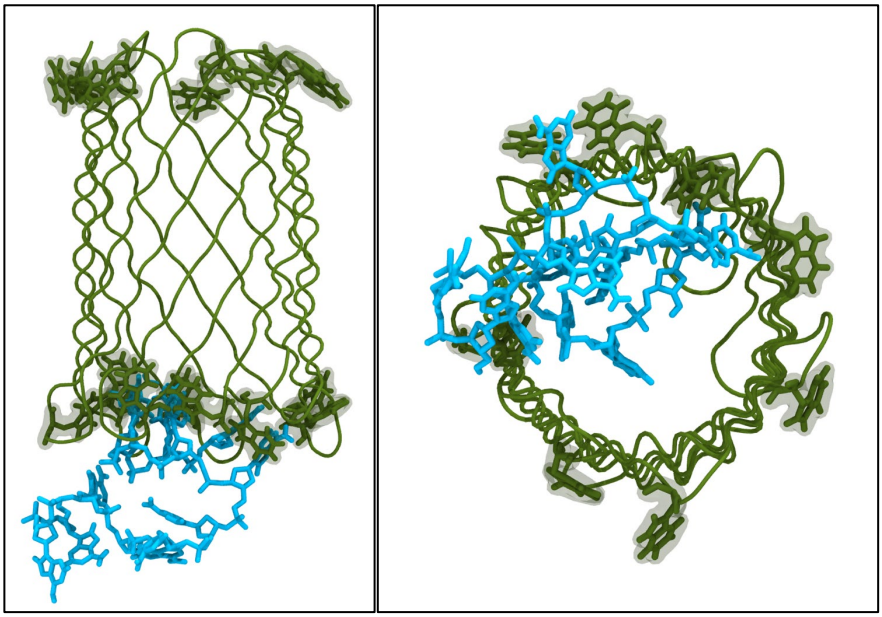

Figure 3.8: DNA (cyan) remained coiled in the 14LLx2 pore exit and associated with anchoring TRP residues (green) in six simulations.

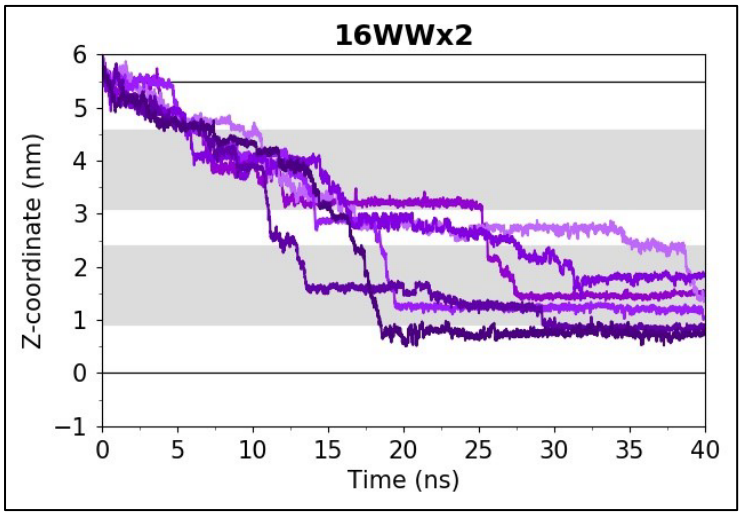

Figure 3.9: DNA translocation through 16WWx2, with short ssDNA pre-threaded through the pore, is measured as the Z coordinate of the centre of mass of the 3' terminal nucleotide over time in eight simulations extended to 40 ns. The constriction regions for each pore are shaded in grey, and solid lines represent the mouths of the pores.

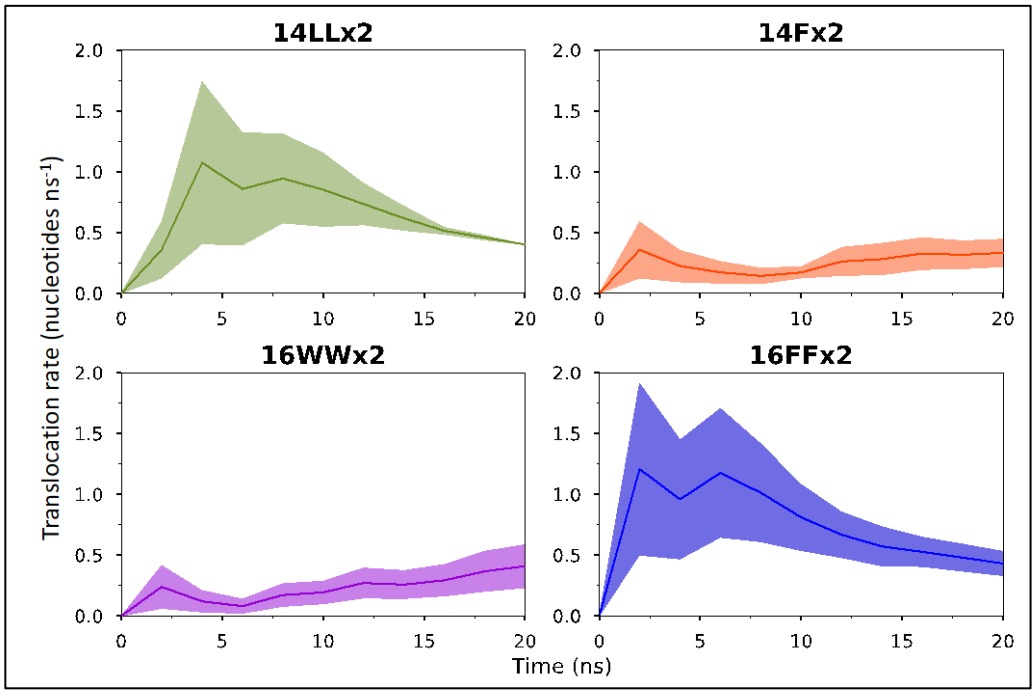

Figure 3.10: DNA translocation rate through the model nanopores, with short ssDNA pre-threaded through the pore, is calculated as an average rate (eight simulations) at which nucleotides exited constriction 2 as a function of time. Standard deviations are shown.

To understand the origins of the differences in DNA translocation between the model nanopores, the conformational behaviour of the DNA strand and the DNA-pore interactions during translocation were examined. The DNA end-to-end distance was measured to identify conformations predominantly adopted by DNA in each pore. Only the conformations inside the pores were considered by using portions of the trajectories during which a 6-nucleotide DNA segment occupied the pore. DNA nucleotides outside the pores were excluded to omit conformations induced by the interactions between DNA and the pore mouth or the membrane. The relative frequency of the end-to-end distances of the DNA segment and the representative conformations of the most frequent end-to-end distance in each pore are shown in Figure 3.11.

In the 16-stranded pores, DNA was observed to coil to varying degrees, with the DNA end-to-end distances ranging between 1.8-3.6 nm. The most frequent DNA end-to-end distance ranged between 2.2-2.6 nm in 16FFx2 and 2.2-2.4 nm in 16WWx2 (note that the DNA end-to-end distance was 3.4-3.6 nm at the beginning of the simulations). A closer inspection revealed that DNA interacted with the aromatic residues in the constriction regions, which led to the strand deviating from an extended conformation. These interactions arose as the aromatic sidechains formed 'pocket' structures within which translocating nucleotides could fit. Up to four nucleotides were observed to occupy a constriction region at a given time, leading to DNA coiling inside both pores. In 16WWx2, the 3' terminal nucleotides were observed to break free from constriction 1 and, as they are unhindered, rapidly translocate downwards into the already occupied constriction 2. As nucleotides could not exit constriction 2 due to DNA-protein interactions, occasionally up to six nucleotides occupied this constriction region, leading to DNA being retained inside the pore by 20 ns. The stepwise movement of DNA from constriction 1 to constriction 2 can be seen in Figure 3.6. This effect was diminished and inconsistent in 16FFx2; hence DNA coiled to a lesser degree, and DNA translocation was faster overall through 16FFx2 than 16WWx2.

In the 14-stranded pores, DNA was maintained in a largely linear conformation in 14LLx2 (Figure 3.11). Although the DNA end-to-end distance ranged between 2.4-3.4 nm, ~ 58% of DNA conformations in eight simulations corresponded to the end-to-end distance between 3.0-3.4 nm. This is comparable to 3.6 nm at the beginning of the simulations when DNA is in an extended conformation. There was little interaction between the translocating DNA and 14LLx2 residues despite the pore being narrow, which resulted in DNA maintaining a largely linear conformation and translocating rapidly through the pore. The DNA strand was observed to coil only after exiting the second constriction region, indicating that the hydrophobic constriction regions maintain the extended conformation adopted by DNA during translocation.

In 14Fx2, the DNA end-to-end distance ranged between 2.3-3.4 nm. This is similar to 14LLx2; however, the end-to-end distances are evenly distributed (compared to 14LLx2, which is skewed towards more extended conformations), indicating that DNA adopted more coiled conformations in 14Fx2. Like in the 16-stranded pores, DNA coiling was facilitated by the strand interacting with the aromatic PHE residues of the constriction regions during translocation. However, DNA coiled to a lesser degree in 14Fx2 than in the 16-stranded pores, as the pore is narrower. Concurrently, the DNA translocation rate through 14Fx2 was slower than through 14LLx2 but faster than through the 16-stranded pores. Together, the results indicate that DNA interactions with aromatic residues led to DNA deviating from a linear conformation and slower DNA translocation rates.

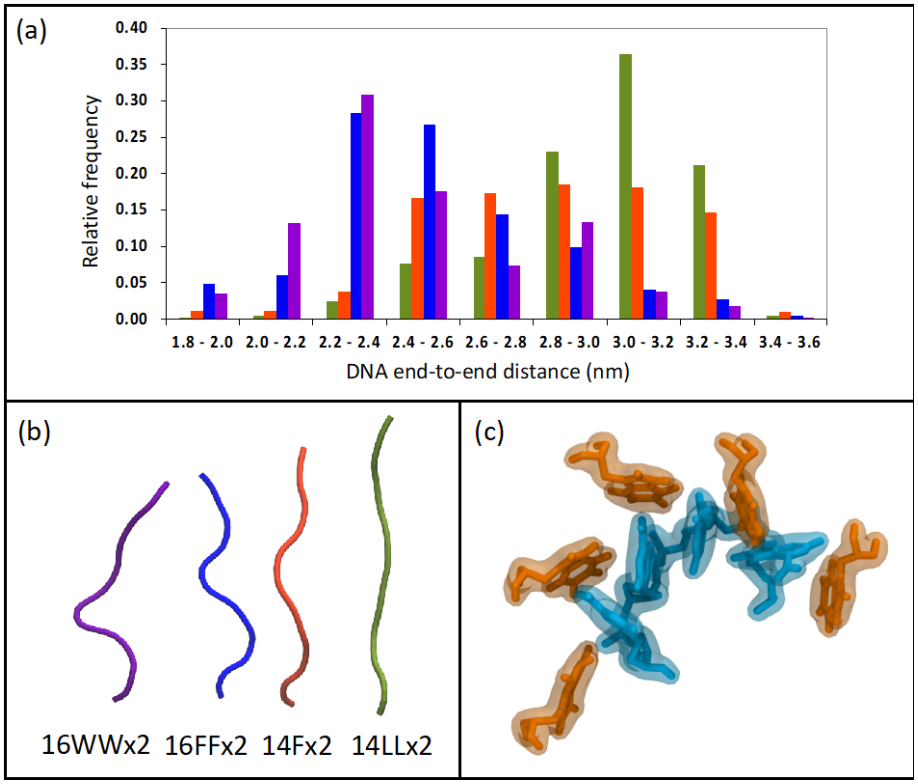

Figure 3.11: DNA conformation of short ssDNA during translocation through model nanopores. (a) The relative frequency distribution of DNA end-to-end distances during translocation through each pore. (b) Representative conformation of the DNA backbone in the model nanopores. (c) Four DNA nucleotides (cyan) interact with TRP residues of 16WWx2 (orange); the nucleotides slot into gaps between the TRP residues.

Next, the interactions between DNA and the pore during translocation were examined. For each pore, the interaction between a DNA nucleotide and two residues of the constriction region was calculated. Portions of two trajectories were post-processed for each pore, in which DNA was

retained within the pore due to the nucleotides interacting with aromatic residues in the constriction regions, or, in the case of 14LLx2, in which DNA simply occupied the constriction regions (given that DNA was not retained due to interactions with LEU residues) (Figure 3.12). The interaction energies show that DNA interaction is greater for the aromatic residues compared to LEU in 14LLx2 (Table 3.2). This once again concurs with the slower DNA translocation rates observed in the presence of aromatic residues in the constriction regions of the pores.

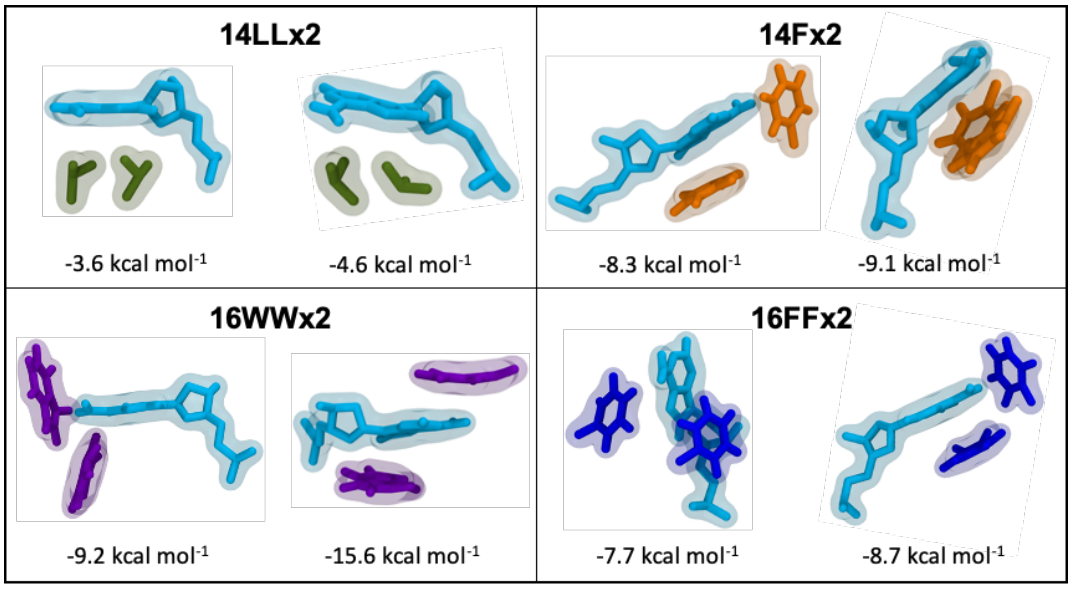

Figure 3.12: The conformation of a DNA nucleotide (cyan) interacting with the sidechains of two residues of the constriction regions in two simulations for each pore, for which the interaction energies were calculated.

Table 3.2. DNA-pore interactions energies, calculated for sidechains of two residues of the constriction region interacting with a DNA nucleotide.

Model nanopore	**Interaction energy (kcal mol^{-1}) ± SD**	
	Simulation 1	Simulation 2
14LLx2	-3.6 ± 0.3	-4.6 ± 0.1
14Fx2	-8.3 ± 0.3	-9.1 ± 0.2
16FFx2	-7.7 ± 0.2	-8.7 ± 0.1
16WWx2	-9.2 ± 0.3	-15.6 ± 0.5

3.3.3.1 Translocation of short ssDNA through model nanopores in a reversed 5' to 3' orientation

The translocation of short ssDNA was simulated in reversed 5' to 3' direction to compare the effect of DNA orientation on its translocation through the model pores [91]. Similar to DNA 3' to 5' translocation simulations, DNA translocation was the fastest through 14LLx2 and the slowest through 16WWx2 (Figures 3.13 and 3.14). In simulations of 14LLx2, DNA exited constriction 2 of the pore by 5-15 ns in seven simulations and exited the pore by 10 ns, 13 ns, and 15 ns in three simulations. Translocation was comparatively slower through 14Fx2; DNA remained threaded through both constriction regions in six simulations and remained threaded through constriction 2 in two simulations. For the 16-stranded pores, DNA remained threaded through both constriction regions of 16WWx2 and 16FFx2 in eight and three simulations, respectively. Like in DNA 3' to 5' translocation simulations, DNA 5' to 3' translocation was more variable through 16FFx2 compared to 16WWx2; the strand remained threaded through constriction 2 of 16FFx2 in four simulations and exited the pore after ~ 17 ns in one simulation (Figure 3.13). In summary, DNA was retained in the constriction region(s) of model pores containing aromatic residues in 23 of 24 independent simulations (and exited the pore in one simulation of 16FFx2), which is very similar to what was observed for DNA 3' to 5' translocation simulations.

The translocation rate at 20 ns through 14LLx2 is ~ 0.42 nucleotides ns^{-1} with faster rates of ~ 0.50 - 0.80 nucleotides ns^{-1} during 5-15 ns as observed in DNA 3' to 5' translocation simulations. The translocation rates are slower through pores with aromatic constriction regions; the translocation rate by 20 ns is ~ 0.20 nucleotides ns^{-1} through 14Fx2, ~ 0.26 nucleotides ns^{-1} through 16FFx2, and ~ 0.01 nucleotides ns^{-1} through 16WWx2. Overall, the trends in DNA translocation through the model pores are similar for both translocation orientations.

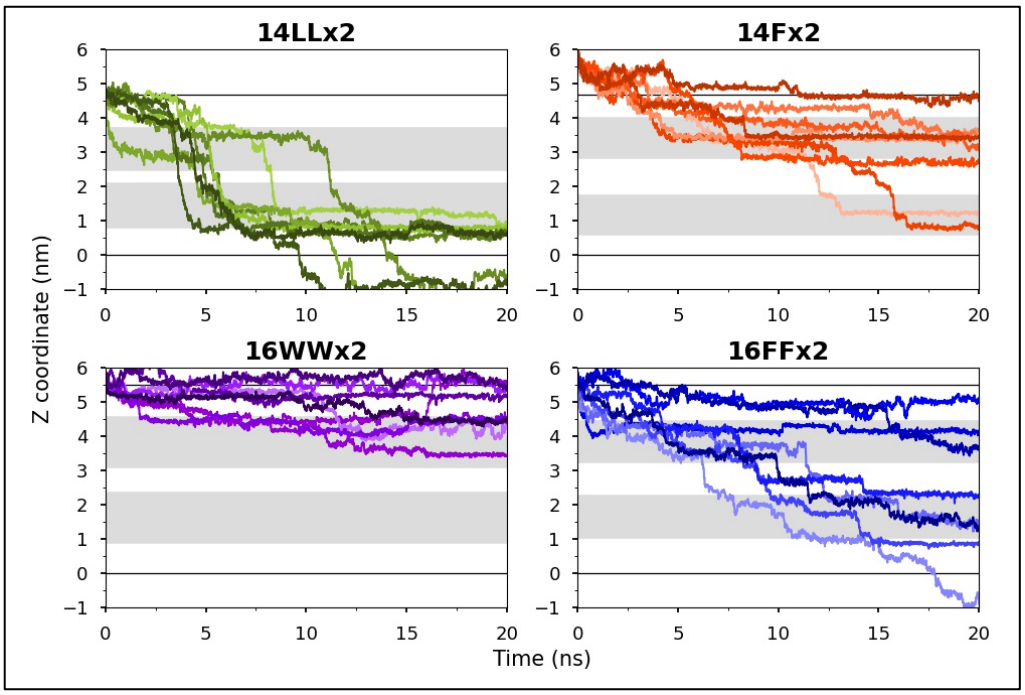

Figure 3.13: DNA translocation through the model pores, with short ssDNA pre-threaded through the pore and moving in 5' to 3' direction, is measured as the Z coordinate of the centre of mass of the 5' terminal nucleotide over time, in eight simulations for each pore. The constriction regions for each pore are shaded in grey, and solid lines represent the mouths of the pores.

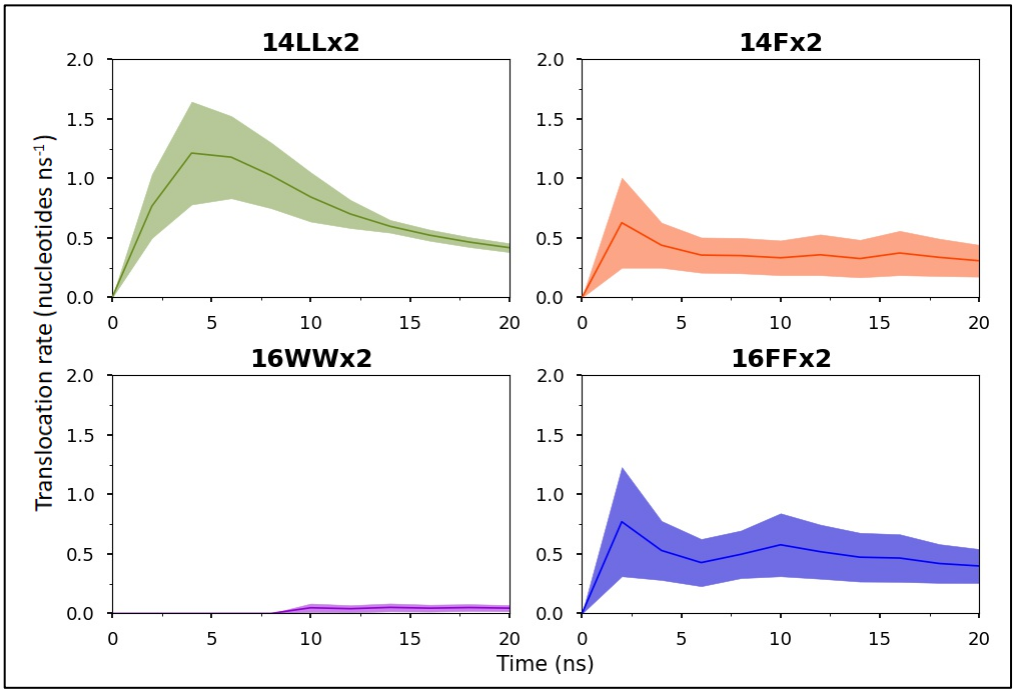

Figure 3.14: DNA translocation rate through the model pores, with short ssDNA pre-threaded through the pore and moving in 5' to 3' direction, is calculated as an average rate

(eight simulations) at which nucleotides exited constriction 2 as a function of time. Standard deviations are shown.

3.3.4 Translocation of long tensioned ssDNA through model nanopores under an applied electric field

It has been previously shown that DNA adopts an extended conformation and is under tension as it enters the solid-state nanopores from a less confined geometry in bulk solution [173]. Proteins used for DNA sequencing often contain large vestibule regions through which DNA enters the narrower pore where the sequence is read [61, 150, 162]. Thus, to explore the behaviour of DNA if it were entering the model nanopores from a vestibule region, the translocation of a longer tensioned ssDNA maintained in a largely linear conformation was studied. The 40-nucleotide polyA ssDNA was bonded to itself across the periodic boundaries in the Z dimension to generate a continuous tensioned ssDNA threaded through the pores. These systems were simulated in NVT ensemble, and as the box dimensions are kept constant, DNA is retained under tension and therefore cannot coil or kink. A single and long 200 ns simulation was performed for each pore under an applied electric field of 0.09 V nm^{-1}, during which multiple translocation events could be observed (as the DNA strand is continuous).

The cumulative number of DNA nucleotides exiting constriction 2 of the model nanopores as a function of time shows that DNA translocation is uniform and rapid through 14LLx2 compared to the rest of the pores (Figure 3.15). The centre of mass movement of the 3' terminal nucleotide as a function of time shows that it took ~ 40 ns for 40 nucleotides to translocate through 14LLx2 (Figure 3.16). When DNA was halted during ~ 180-200 ns, this was due to a nucleotide interacting with a TRP residue in the pore exit and not due to interactions between DNA and residues of the constriction regions. DNA translocation was restored after 208 ns once this interaction was disrupted (Figure 3.17). Hence, DNA is not retained by LEU residues in 14LLx2.

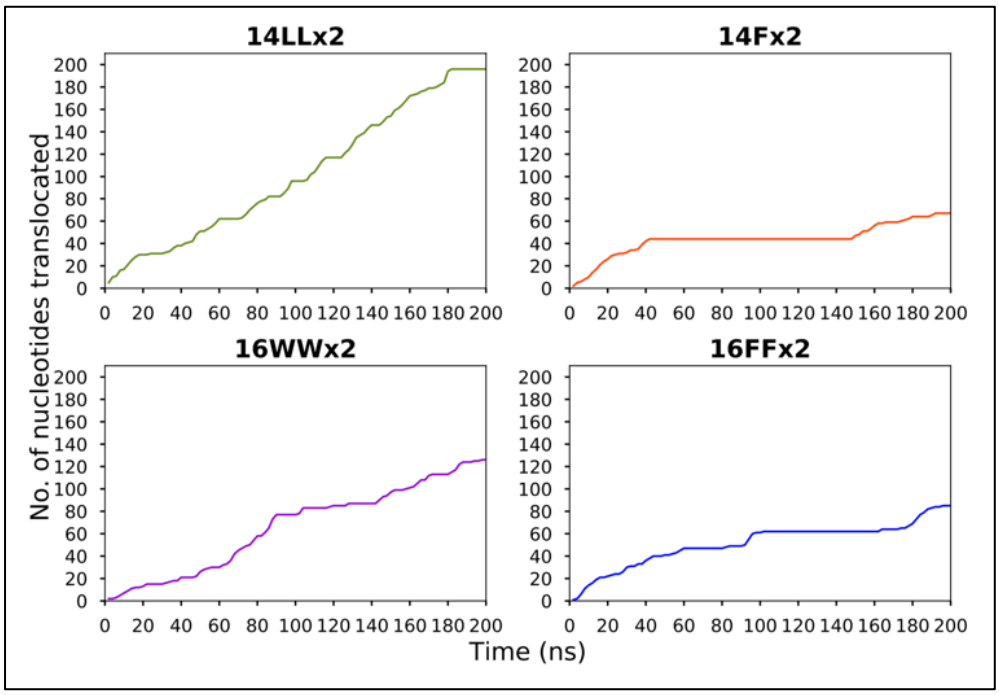

Figure 3.15: DNA translocation through the model nanopores, with continuous tensioned ssDNA pre-threaded through the pore, is shown as the cumulative number of DNA nucleotides exiting constriction 2 as a function of time.

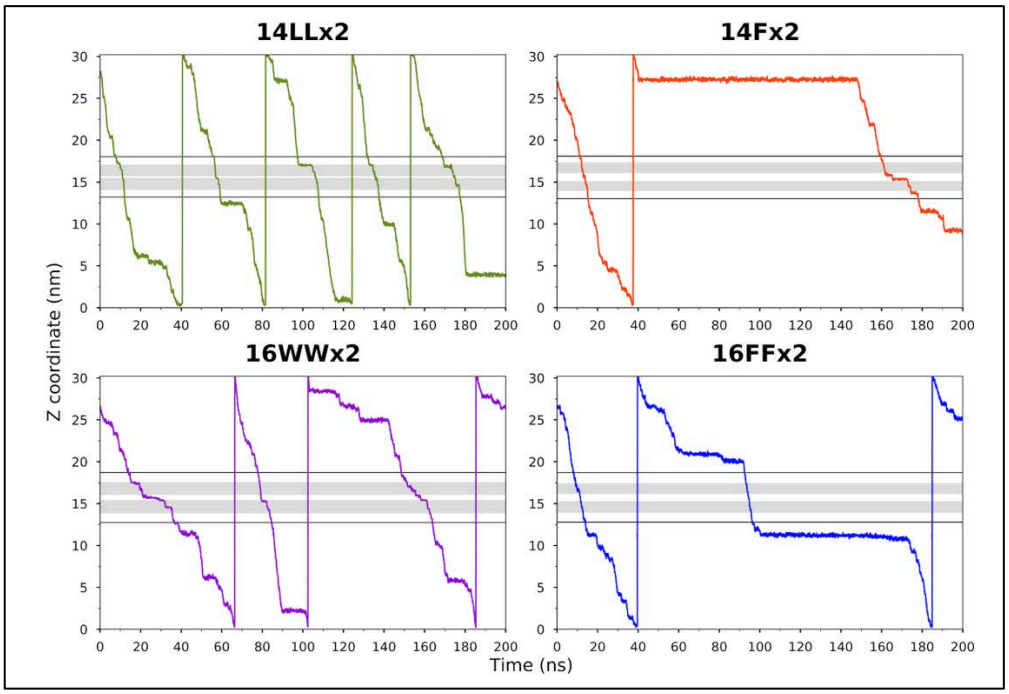

Figure 3.16: DNA translocation through the model nanopores, with continuous tensioned ssDNA pre-threaded through the pore, is measured as the Z coordinate over time of the centre of mass of the 3' terminal nucleotide starting furthest away from constriction

1. The constriction regions for each pore are shaded in grey, and solid lines represent the mouths of the pores.

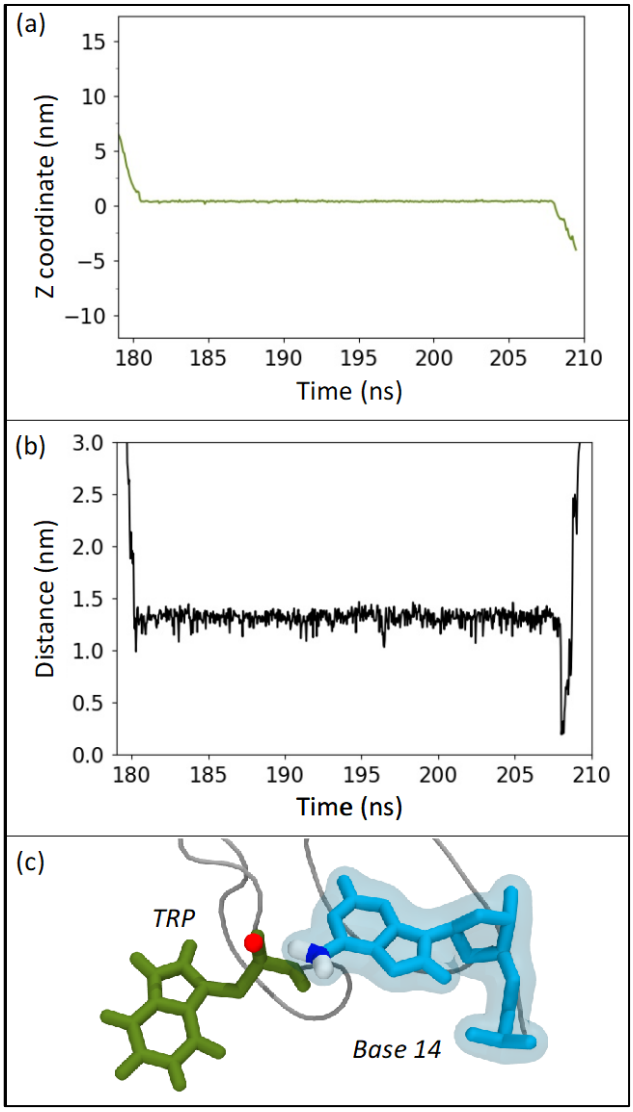

Figure 3.17: The translocation of continuous tensioned ssDNA through 14LLx2 is halted during ~ 180-208 ns. (a) Z coordinate of the centre of mass of nucleotide 14 over time. (b) The distance of the interaction between nucleotide 14 and TRP residue in the pore exit. (c) A molecular view of the nucleotide 14 and TRP residue.

DNA translocation was slower and less uniform through the pores with aromatic constriction regions compared to 14LLx2 (Figure 3.15). For 14Fx2, although the first translocation event of 40 nucleotides occurred within ~40 ns, like through 14LLx2, the second translocation event did not complete during the rest of the simulation (Figure 3.16). During ~ 42-145 ns, DNA translocation was halted as two neighbouring PHE residues in constriction 1 formed a ‘gate’ in which one of the

DNA nucleotides was 'trapped' *via* non-specific steric interactions (Figure 3.18). The rearrangement of PHE sidechains, such that they were ~ > 0.5 nm apart, led to the loss of the 'gate', and the DNA nucleotide subsequently moved out of constriction 1 under the influence of the electric field.

The interaction between the captured DNA nucleotide and the two PHE residues forming the 'gate' was calculated using DFT. The interaction energy is -11.4 kcal mol^{-1} when the nucleotide is trapped and the gate is closed, and this is reduced to -7.5 kcal mol^{-1} upon opening of the gate and further reduced to -2.2 kcal mol^{-1} once the nucleotide moved out of the constriction region (Figure 3.19).

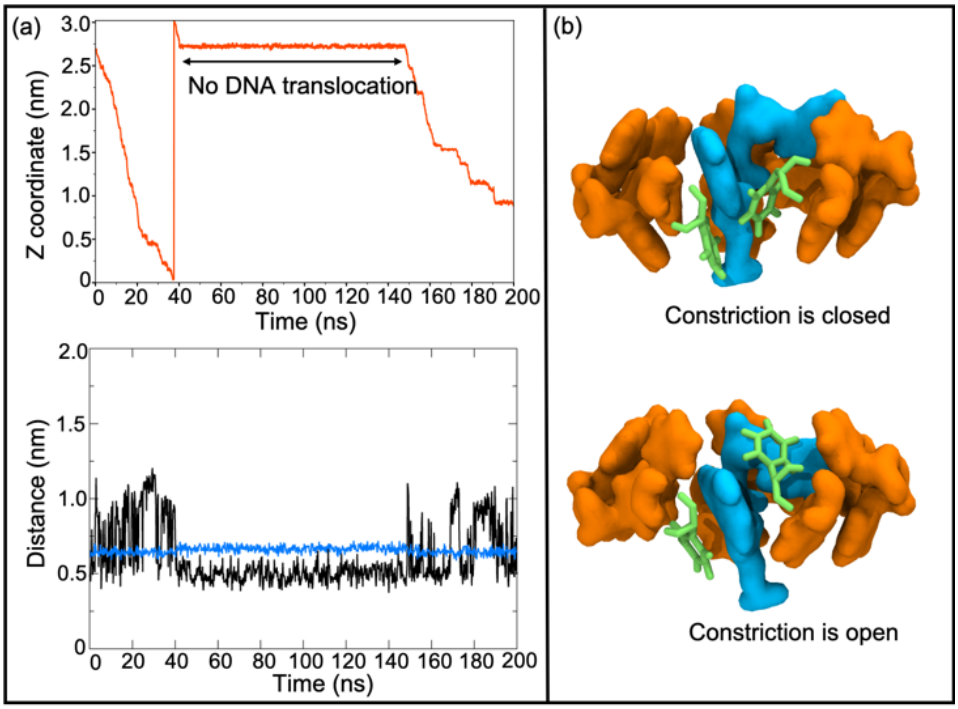

Figure 3.18: The translocation of continuous tensioned ssDNA through 14Fx2 is halted during ~ 42-145 ns. (a, top) Z coordinate over time of the centre of mass of the 3' terminal nucleotide starting furthest away from constriction 1. (a, bottom) The distance over time between two PHE residue sidechains (black) and their backbone Cα atoms (blue) in constriction 1. (b) Molecular view of the PHE residues (green) interacting with two DNA nucleotides (cyan) that are halted in the constriction during 42-145 ns. The other PHE residues forming constriction 1 are also shown (orange).

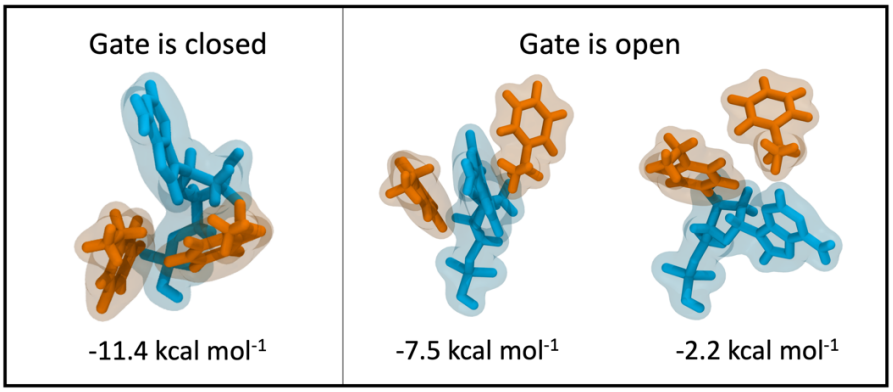

Figure 3.19: The conformation of a DNA nucleotide of continuous tensioned ssDNA (cyan) and two PHE residues forming the 'gate' in the constriction region when the gate is closed and after it opens. The interaction energies for each conformation calculated using DFT are shown.

To further investigate the dynamics of the gate in 14Fx2, two simulations were run with the gate initially in a closed conformation under applied electric fields of 0.08, 0.09, or 0.10 V nm^{-1} (Figures 3.20, 3.21, and 3.22). The gate opened in one of the two simulations for each electric field strength. In a simulation under an applied electric field of 0.10 V nm^{-1}, a third PHE residue also participated in the gate, and DNA translocation was subsequently halted for 100 ns (Figure 3.23). Overall, these simulations demonstrate that the DNA-protein interactions are stochastic in nature when DNA is retained under tension.

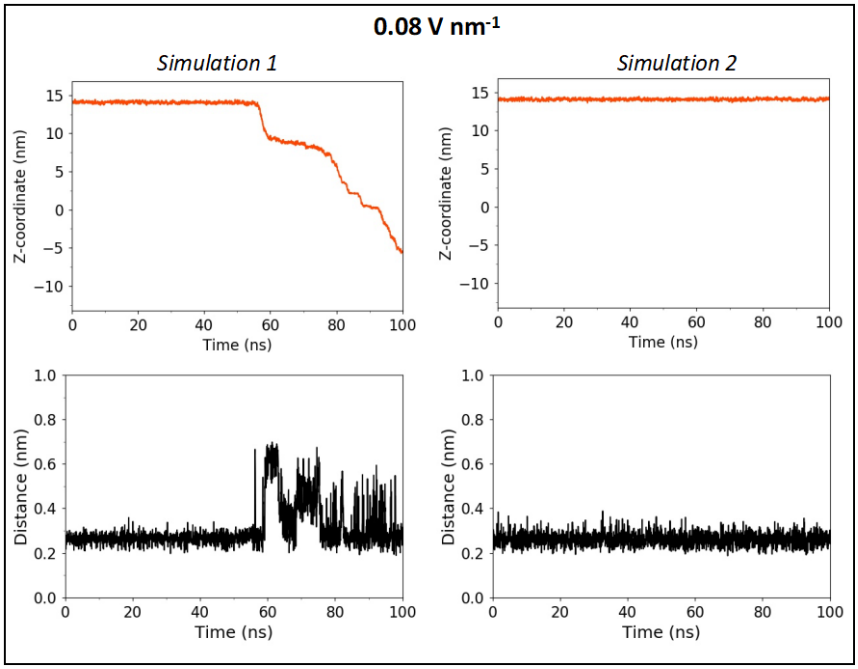

Figure 3.20: DNA translocation through 14Fx2, with continuous tensioned ssDNA pre-threaded through the pore, is measured as the Z coordinate over time of the centre of mass of the 3' terminal nucleotide starting furthest away from constriction 1 (orange). The distance over time between two PHE residue sidechains forming the gate is plotted (black). Data is from two simulations in 0.08 V nm^{-1}.

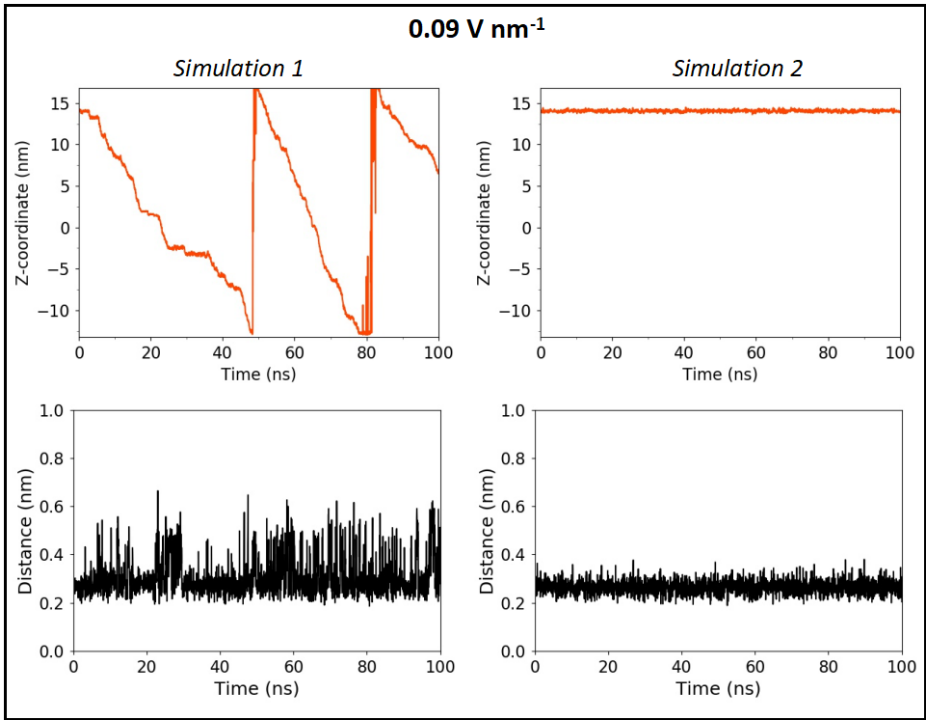

Figure 3.21: DNA translocation through 14Fx2, with continuous tensioned ssDNA pre-threaded through the pore, is measured as the Z coordinate over time of the centre of mass of

the 3’ terminal nucleotide starting furthest away from constriction 1 (orange). The distance over time between two PHE residue sidechains forming the gate is plotted (black). Data is from two simulations in 0.09 V nm^{-1}.

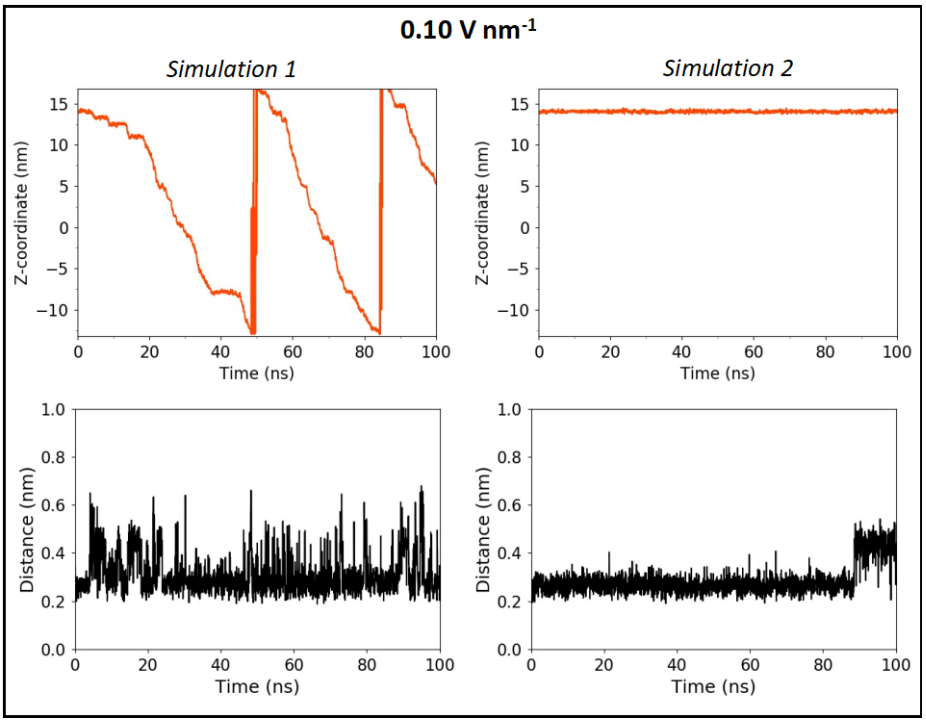

Figure 3.22: DNA translocation through 14Fx2, with continuous tensioned ssDNA pre-threaded through the pore, is measured as the Z coordinate over time of the centre of mass of the 3’ terminal nucleotide starting furthest away from constriction 1 (orange). The distance over time between two PHE residue sidechains forming the gate is plotted (black). Data is from two simulations in 0.10 V nm^{-1}.

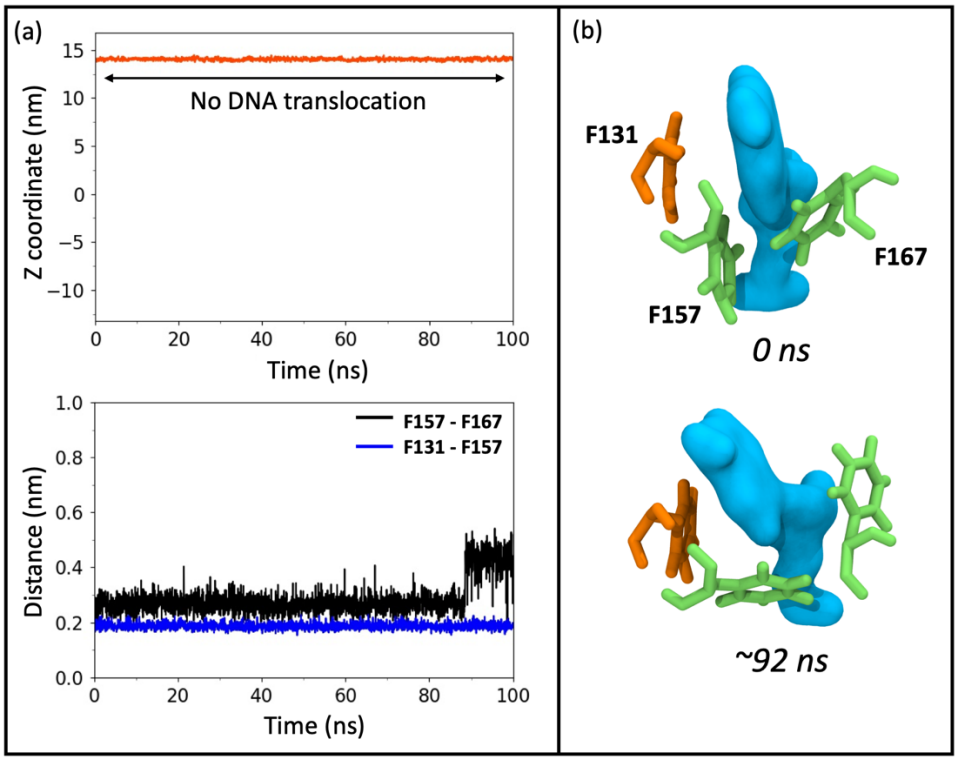

Figure 3.23: The translocation of continuous tensioned ssDNA through 14Fx2 is halted for 100 ns in 0.10 V nm^{-1}. (a, top) Z coordinate over time of the centre of mass of the 3' terminal nucleotide starting furthest away from constriction 1. (a, bottom) The distance over time between pairs of PHE residue sidechains forming the gate in constriction 1. (b) Molecular view of the two PHE residues forming the gate at the beginning of the simulation (green), and another PHE residue participating in the gate (orange), interacting with a DNA nucleotide (cyan) halted in the constriction for 100 ns.

For the 16-stranded pores, the DNA retention times also correlated with the interactions between DNA and the aromatic residues of the constriction regions. Like in 14Fx2, the aromatic residues formed tight pockets around the DNA nucleotides *via* non-specific steric interactions (Figure 3.24). The 'steps' in the Z coordinate of DNA during translocation (Figure 3.16) corresponded to the nucleotide moving out of a pocket, whilst other nucleotides are retained within the same or the other constriction (Figure 3.25). DNA moved rapidly through these pores in some translocation events, with complete translocation occurring within ~ 40 ns, like through 14LLx2. Nucleotides were unable to move into pockets formed in the constriction regions during these events. In some cases, a nucleotide moving into a pocket was not retained, possibly due to the initial nucleotide-pocket interaction being sub-optimal. As the DNA backbone is conformationally inflexible, it could not coil to optimise this interaction, and hence DNA translocated rapidly through the pore under the influence of the electric field.

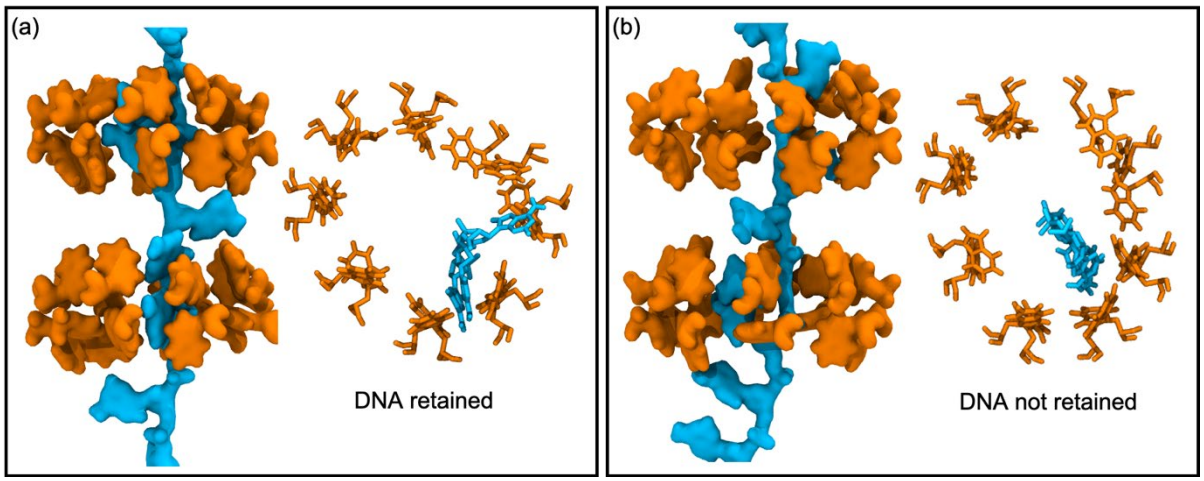

Figure 3.24: Two types of translocation events of continuous tensioned ssDNA were observed through pores with aromatic residues in the constriction region. (a) When DNA translocation is slowed, the nucleotides (cyan) are retained in a pocket (TRP residues in constriction 2 of 16WWx2 are shown in orange). (b) DNA translocation is rapid when nucleotides are unable to move into a pocket.

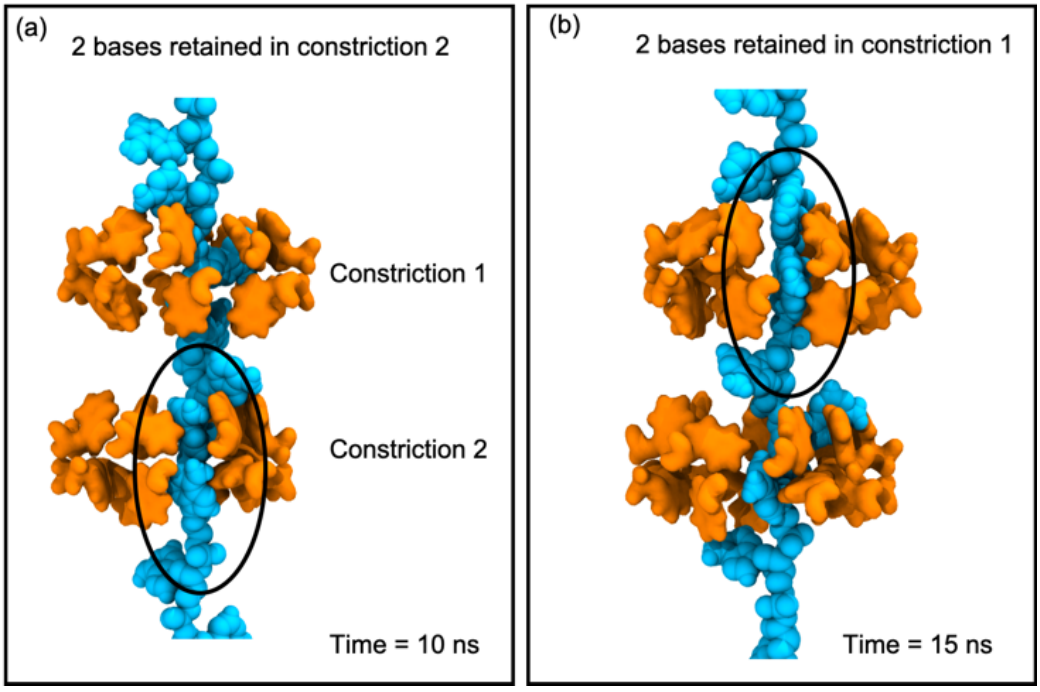

Figure 3.25: Stepwise translocation of continuous tensioned ssDNA. (a) Two DNA nucleotides (cyan, circled) are caught in a pocket formed by TRP residues (orange) in constriction 2 of 16WWx2. (b) The system in (a) 5 ns later, when two nucleotides (circled) are caught in a pocket formed by TRP residues in constriction.

Interestingly, DNA was retained for longer periods in 16FFx2 compared to 16WWx2. This is likely due to the PHE sidechains being more flexible than the bulkier TRP sidechains (Figure 3.26), so the former can rapidly rearrange to form pockets around the DNA nucleotides before they translocate further through the pore.

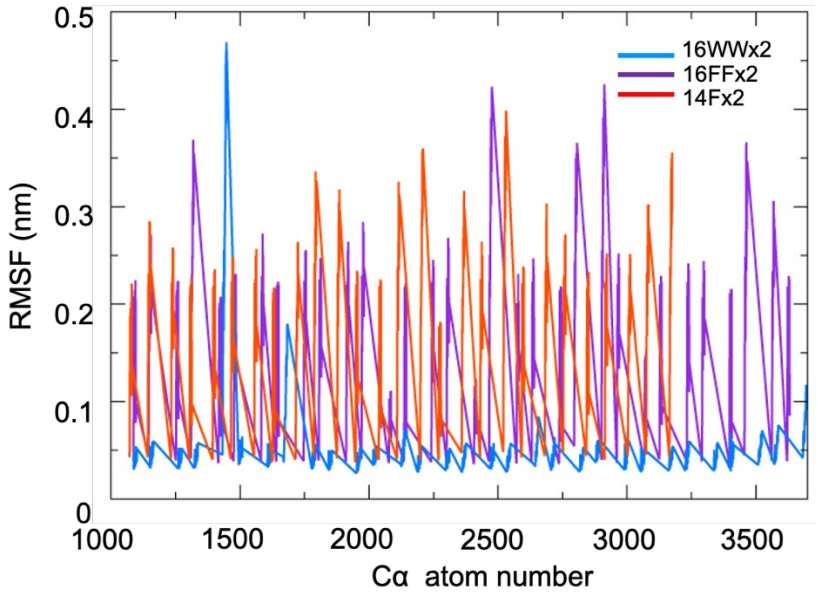

Figure 3.26: RMSF of aromatic residue sidechains within the constriction regions of 16WWx2, 16FFx2 and 14Fx2. Lower RMSF values of residues in 16WWx2 indicate lower flexibility of the TRP sidechains compared to PHE sidechains in 16FFx2 and 14Fx2.

3.3.5 Electrowetting behaviour of nanopores in different forcefields

In the initial simulations of the model nanopores without DNA and an applied electric field of 0.15 V nm^{-1}, water and ions did not enter 14LLx2, and the pore remained dry for 20 ns in three simulations. Previously, Trick et al. have shown that hydrophobic regions in model nanopores act as barriers to water and ion permeation. The introduction of 2 or 3 rings of LEU residues in a 14-stranded nanopore resulted in the pore dewetting when simulated in the absence of an electric field [154]. The behaviour of the 14LLx2 and model nanopores in the study performed by Trick et al. is akin to hydrophobic gating exhibited by some proteins, such as ion channels, whereby the open or closed states of the pores depend on the inclusion or exclusion of water by the hydrophobic region without steric occlusion [174, 175].

Water transport is also influenced by the size of the channel formed by the pore. The dynamics of water when nanoconfined have been extensively studied *via* simplified nanopores that are

proteinaceous [174, 176], synthetic such as carbon nanotubes [177], and also protein channels with hydrophobic gates [178, 179]. In hydrophobic nanopores < ~ 0.55 nm in diameter, the 'vapour' state of water is energetically more favourable than the 'liquid' state, therefore, the pore remains dry [180]. The strong hydrogen bonds between the water molecules cause the water to recede from the hydrophobic region in narrow nanopores. Additionally, water transport can also be influenced by applying a voltage. The 14-stranded model nanopores with 2 or 3 rings of LEU residues and a propensity to remain dry were observed to electrowet in voltages higher than 1 V [181]. Electrowetting occurs in an electric field due to the voltage-induced alignment of water dipoles, and the water molecules in the hydrophobic region maintain their interactions [181]. Therefore, the hydrophobic gating of nanopores can potentially be modulated by applying a voltage [182].

One challenge when investigating water dynamics *via* MD simulations is that the observations greatly depend on the water model used [183]. The choice of water model may be critical for simulating channels with hydrophobic gates on the 'edge' of wetting [179]. Therefore, further simulations were conducted of the 14-stranded model pores using the SPC water model with the GROMOS 53A6 forcefield and the TIP3P water model with CHARMM36m forcefield to elucidate the behaviour of water inside the hydrophobic nanopores as a function of the water models used. 14LLx2 and 14Fx2 were simulated in 0 V and 0.15 V nm^{-1}, with the pores either dry or wet at 0 ns. The data from these simulations are provided in Table 3.3. Starting with 14LLx2, in the absence of an electric field, the initially wet pore was observed to dewet by ~ 0.4-0.6 ns in three simulations with the GROMOS 53A6 forcefield (Figure 3.27). This was not the case in simulations with the CHARMM36m forcefield, in which the pore remained wet during 20 ns in three simulations. However, when 14LLx2 was initially dry, it stayed dry during 50 ns in simulations of both forcefields. In 0.15 V nm^{-1}, although 14LLx2 wetted by ~ 0.8-1.6 ns in three simulations using the CHARMM36m forcefield, it remained dry in three simulations with the GROMOS 53A6 forcefield, even when the simulations were extended to 100 ns. Interestingly, no differences in the behaviour of water in 14Fx2 were observed in simulations with the two forcefields, despite the pore dimensions being similar to 14LLx2.

Table 3.3. Summary of 14-stranded pore electrowetting behaviour in simulations using GROMOS 53A6 or CHARMM36m forcefields, in 1 M NaCl and 310 K.

Model pore	**Forcefield**	**Water model**	**0 V**		**0.15 V nm^{-1}**	
			Dry at t=0 ns	**Wet at t=0 ns**	**Dry at t=0 ns**	**Wet at t=0 ns**
14LLx2	CHARMM36m	TIP3P	Dry (3 x 50 ns)	Wet (3 x 20 ns)	Wet by ~ 0.8-1.6 ns (3 x 20 ns)	Wet (3 x 20 ns)
	GROMOS 53A6	SPC	Dry (3 x 50 ns)	Dewetted by ~ 0.4-0.6 ns (3 x 20 ns)	Dry (3 x 100 ns)	Wet (3 x 20 ns)
14Fx2	CHARMM36m	TIP3P	Dry (3 x 50 ns)	Wet (3 x 20 ns)	Wet by ~ 0.3-0.7 ns (3 x 1 ns)	Wet (3 x 20 ns)
	GROMOS 53A6	SPC	Dry (3 x 50 ns)	Wet (3 x 20 ns)	Wet by ~ 0.7-3.0 ns (3 x 20 ns)	Wet (3 x 20 ns)

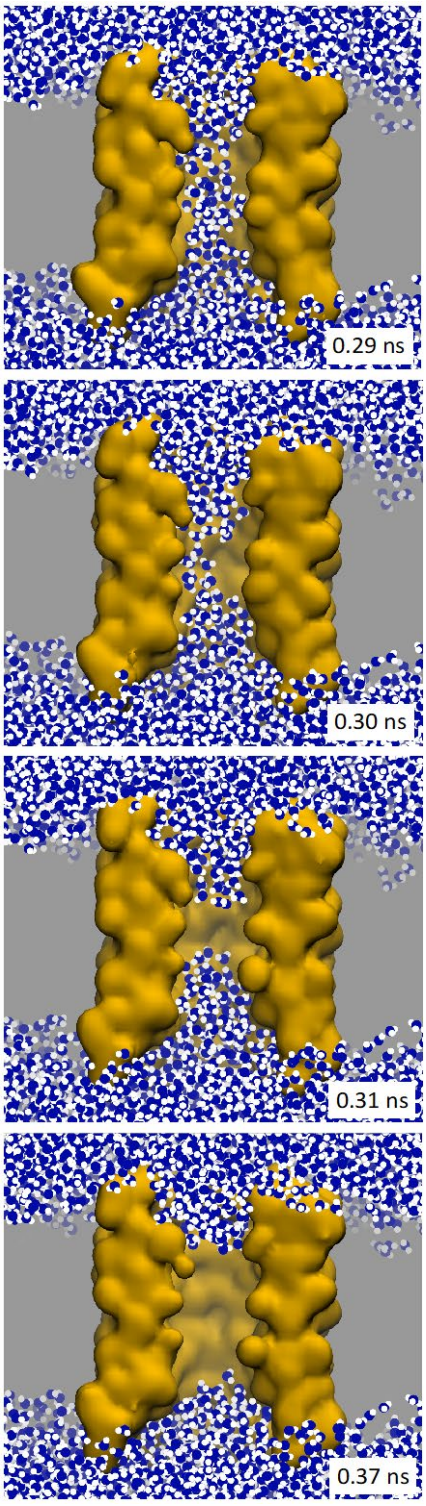

Figure 3.27: 14LLx2 dewetted in 0 V when simulated using the GROMOS 53A6 forcefield. The pore cross-section is shown in surface representation, and the water molecules are shown as blue (oxygen) and white (hydrogen) spheres. The area occupied by the lipid bilayer is shaded in grey.

3.4 Conclusions

In this chapter, model nanopores which are β-barrels modified to incorporate two hydrophobic constriction regions were studied to explore the effect of the chemical nature and geometry of nanopores on DNA translocation. The pores were initially simulated without DNA under an applied electric field. As expected, the mean flux of water and ions was lower for 14-stranded pores compared to the 16-stranded pores. However, no flux was observed through 14LLx2. Simulations of the 14-stranded pores showed that water dynamics inside 14LLx2 differ according to the forcefield used, however this was not the case in 14Fx2, despite both pores being similar in diameter. It would be of interest to further investigate the impact of pore geometry on water and ion dynamics inside nanometer-sized pores as a function of the ionic solution used, the strength of the applied electric field, or the water model used [183, 184], as this is relevant for simulations of narrow protein channels [179].

The model nanopores were simulated with a short flexible ssDNA and a longer tensioned ssDNA, both with polyA sequences. Short ssDNA readily entered the wider 16-stranded pores but could not enter the 14-stranded pores under an applied electric field, indicating that the pore width strongly influences the entry of DNA into these nanopores. However, it is important to note that the proteins used for DNA sequencing contain large vestibule regions from which the DNA strand the narrower pore, unlike the model nanopores in which DNA enters directly from the bulk solution. The cut-off size for easy entry of ssDNA into the model nanopores is predicted to be ~ 16 strands.

The translocation of short ssDNA is rapid and unhindered through the 14LLx2 pore, during which DNA is retained in an extended conformation. However, when LEU residues of the constriction regions are replaced with aromatic PHE residues (in 14Fx2), DNA translocation is slower, and the strand deviates from a linear conformation. This was also observed for wider 16-stranded pores, which indicates that aromatic residues slow the translocation rate of short ssDNA in the constriction regions. Hence, DNA translocation is strongly influenced by the chemical nature of residues of the constriction region compared to the pore width. There is no correlation between the width of the model nanopores and the translocation rate; DNA translocation was the slowest through 16WWx2, which is intermediate in width between 14Fx2 and 16FFx2.

The translocation of long tensioned ssDNA is the fastest through 14LLx2, as is the case for short ssDNA. Pores with aromatic residues in the constriction regions, especially PHE residues, slowed DNA translocation, as they retained DNA for extended periods by forming non-specific steric interactions with the nucleotides. However, when the pockets formed by aromatic residues do not capture DNA, DNA translocation is rapid and at a similar rate as within 14LLx2.

The results in this chapter suggest that nanopores containing aromatic TRP or PHE residues can slow down the translocation rate of flexible or tensioned ssDNA, respectively. The 14- and 16-stranded model nanopores with PHE residues forming the constriction regions showed the most promise, as they slowed down DNA translocation and, especially 14Fx2, also restricted the range of conformations adopted by the strand. While the 16-stranded pores have the advantage of easier DNA entry, the 14-stranded pores retained DNA in a more extended conformation, and 14Fx2 slowed down the translocation of both short flexible and longer tensioned ssDNA. To facilitate DNA entry, adding a vestibule region to 14Fx2 would be advantageous, similar to a-hemolysin, which also contains a 14-stranded β-barrel and is known to allow DNA entry into the pore.

Chapter 4 DNA translocation through the *E. coli* proteins CsgG and CsgF

4.1 Introduction

Proteins that form naturally occurring nanometer-sized channels have been popularly used as nanopores for DNA sequencing. Over the years, several proteins such as α-hemolysin [80, 88, 91, 152], MspA [162], and CsgG [43], as well as hybrid protein-synthetic pores [152], have been studied and optimised for DNA sequencing, with a focus on optimising of the sensing region of the nanopore [1, 60].

A mutant of *Escherichia coli* (*E. coli)* CsgG has been used in nanopore DNA sequencing devices since 2016 [43]. CsgG is an outer membrane lipoprotein and a component of the curli biogenesis system, which is multi-protein machinery that facilitates the secretion of curli subunits and their assembly into highly aggregative amyloid fibres associated with biofilm formation in Gram-negative bacteria. CsgG is a nonameric protein containing a 36-stranded transmembrane β-barrel, connected to a large solvent-accessible vestibule that opens into the periplasmic space. The β-barrel and the vestibule regions are partitioned by a ~ 1 nm wide constriction region [74, 75]. CsgG works in conjunction with other Csg proteins, including CsgA and CsgB, which form extracellular fibres, and CsgE and CsgF, which coordinate the secretion and assembly of the curli subunits [73]. CsgG permits ungated diffusion of peptides through the channel across the outer membrane. However, it becomes substrate-specific upon binding CsgE and CsgF accessory proteins during the curli biogenesis [185-187]. CsgE binds to the CsgG vestibule, forming a 'cap' that effectively gates the channel at the periplasmic side [74]. CsgF is secreted to the extracellular milieu through CsgG. The C terminus region of CsgF remains outside of CsgG, to which the growing amyloid fibre attaches during its secretion [75, 124, 125]. After folding into an α-helix, its N terminus remains in the CsgG β-barrel to form a second constriction region in the channel (Figure 4.1).

Previously, studies have shown the potential advantages of nanopores with two constriction regions for DNA sequencing [123, 151, 153]. In this study, the CsgG protein and the CsgG-CsgF complex are characterised using advanced molecular dynamics simulations. The stability of the two structures is assessed under applied electric fields; this is relevant for DNA sequencing, during which an electric field is applied to facilitate DNA movement through the nanopore. The translocation of DNA through CsgG and the CsgG-CsgF complex is investigated to ascertain the

impact of the nanopore geometry (single- or dual-constrictions) on DNA translocation rate and conformations.

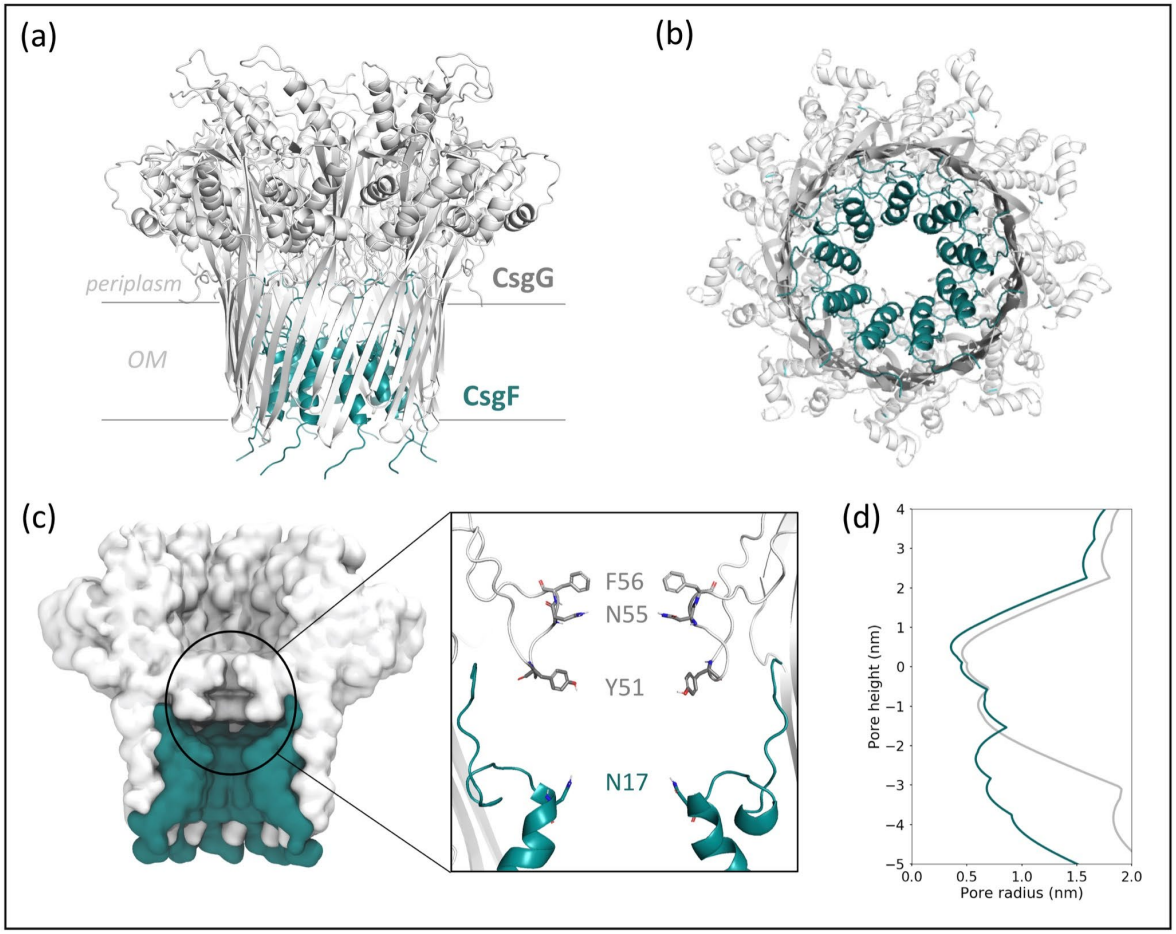

Figure 4.1: (a) The CsgG-CsgF complex is shown from the side in ribbon representation, with the outer membrane (OM) and the periplasm labelled. (b) Extracellular view of the CsgG-CsgF complex in ribbon representation. (c) A cross-sectional side view of the CsgG-CsgF complex is shown in surface representation. A close-up view of the constriction regions is also shown, with the labelled residues labelled in stick representation. (d) Pore radius profiles of CsgG and the CsgG-CsgF complex, coloured as labelled in (a).

Notes

This chapter is based on the publication "Atomistic level characterisation of ssDNA translocation through the *E. coli* proteins CsgG and CsgF for nanopore sequencing", published in the Computational and Structural Biotechnology Journal [188]. The figures used in this chapter are reproduced from ref. [188]

4.2 Methods

4.2.1 Preparation of protein structures

The protein structures used in this study are as follows: CsgG-1 - CsgG crystal structure (PDB 4UV3, 3.59 Å); CsgG-2 - CsgG taken from the electron cryo-EM structure of the CsgG-CsgF complex with CsgAN6 peptide (PDB 6L7C, 3.34 Å); CsgG-CsgF-1 – the CsgG-CsgF complex electron cryo-EM structure (PDB 6SI7, 3.4 Å); and CsgG-CsgF-2 – the CsgG-CsgF complex taken from the electron cryo-EM structure of the CsgG-CsgF complex with CsgAN6 peptide (PDB 6L7C, 3.34 Å). Two experimental structures were used as the starting conformations of CsgG and the CsgG-CsgF complex each, as each is one of the many conformational states adopted by the protein *in vivo* and therefore is a starting point to sample more of the conformational behaviour of the protein.

For CsgG-1, the missing loops (residues 144, 193-199) were built using Coot [189] by fitting the structure into the map density obtained at a higher resolution (PDB 4Q79, 3.1 Å). For CsgG-CsgF-1, the missing residues (residues 1-9, 103-110) were built using Modeller 9.02 *via* sequence alignment and fitting with CsgG-1 [155].

4.2.2 Simulations of systems without DNA

The N termini cysteine residues in CsgG were lipidated in all structures before inserting in a membrane composed of 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipids (CsgG: 1026 lipids; CsgG-CsgF: 1106) using CHARMM-GUI membrane builder [167, 190-192]. The systems were immersed in 1 M KCl [163] along with additional ions for neutralising the systems.

The equilibration of the systems involved running multiple simulations so that the strength of positional restraints applied to the protein and the POPC lipids could be reduced in a stepwise manner. Initial equilibration simulations were performed for 385 ps in total, with the system temperature maintained at 303.15 K to improve the packing of the membrane around the lipid anchors of the protein. The temperature was increased to 310 K in subsequent equilibration simulations of 225 ps.

4.2.3 Steered molecular dynamics simulations with DNA

The models of ssDNA with polyA and polyC sequences were generated using the 3DNA package [157]. DNA was added to CsgG-1 and CsgG-CsgF-1 systems, such that the 5’ terminal nucleotide was positioned inside the CsgG eyelet loop region at least 0.2 nm away from the protein residues to prevent steric clashes. The resultant systems were solvated, with ions added to a concentration

of 0.15 M alongside additional ions for neutralising the systems. The systems were equilibrated as described earlier (section 4.2.2) in 310 K. Positional restraints were applied to the DNA 5' terminal nucleotide during equilibration with a force constant of 1000 kJ mol^{-1} nm^2.

DNA was pulled through CsgG and the CsgG-CsgF complex in constant velocity (CV) steered MD simulations, as done previously for studying translocation through a-hemolysin [93, 94, 193], MspA [89], and solid-state nanopores [194, 195]. The 5' terminal nucleotide was set as the pull group, and CsgG Pro-52 was selected as the reference group. Both groups were linked *via* a harmonic spring with a spring constant of 1000 kJ mol^{-1} nm^2. The reaction coordinate, along which the 5' terminal nucleotide was pulled, was set as a vector parallel to the protein axis and in -Z direction.

4.2.4 Conductance of CsgG and the CsgG-CsgF complex in the presence of immobilised DNA

The starting structures of CsgG and the CsgG-CsgF complex with DNA were extracted from steered MD simulations. Positional restraints of 1000 kJ mol^{-1} nm^2 were applied to DNA backbone phosphorus atoms. In simulations where the CsgG eyelet loops were restrained, positional restraints of 500 kJ mol^{-1} nm^2 were applied to backbone atoms of residues 47-58.

4.2.5 Simulation protocol and analyses

All simulations were performed using GROMACS 2018.3 [159] and the CHARMM36m force field [196], and the TIP3P water model was used for solvating the systems [163]. Systems were simulated in NPT ensemble, with the pressure maintained semi-isotropically using the Parrinello−Rahman barostat at 1 bar and a time constant of 5 ps [145], and the temperature sustained using the velocity-rescale thermostat and a coupling constant of 0.1 ps. The lengths of all bonds were constrained using the LINCS algorithm enabling a timestep of 2 fs [134]. The Particle Mesh Ewald (PME) method was used to treat long-range electrostatic interactions with a short range cut-off of 1.4 nm [140]. The van der Waals interactions were curtailed at 1.4 nm, with long-range dispersion corrections applied to the pressure and energy. A constant voltage drop across the simulation cell in Z dimension imposed an electric field. The periodic boundary conditions were applied to all systems in three dimensions, like in previous studies [81, 88, 89, 162]. Replicate simulations were initiated using coordinates extracted at random time points from the last 100 ps of the equilibration run. The initial coordinates and velocities differ for each replicate simulation for the systems.

Analyses were performed using GROMACS utilities and locally written code. Clustering analysis was performed using the linkage method implemented in GROMACS. Trajectories from independent simulations were concatenated before clustering DNA conformations with a RMSD cut-off of 0.28 nm. Pore radius profiles of the proteins were calculated using HOLE [164]. The ionic density maps were generated using the Visual Molecular Dynamics (VMD) package [165]. Porcupine plots for the Principal Components Analysis were generated in PyMOL [156]. The molecular graphics images were generated using VMD and PyMOL. The ionic current and the mean water and ion flux rates were calculated as described previously ([80], [122]).

4.2.6 Statistical analysis

The unpaired t-test was used to assess the significance of the differences in DNA translocation rates, mean water and ion flux rates, and ionic currents. $p < 0.05$ was regarded as statistically significant.

4.3 Results and Discussion

The simulations of the CsgG and the CsgG-CsgF complex systems in this study are presented in Table 4.1.

Table 4.1. Summary of the simulations discussed in this chapter.

System	DNA	Simulations			
		0 V	**0.05 V nm^{-1}**	**0.075 V nm^{-1}**	**Steered MD**
CsgG	no	4 x 100 ns 2 x 200 ns	6 x 100 ns	6 x 50 ns	-
	yes	-	6 x 50 ns	-	8 x 70 ns
CsgG-CsgF complex	no	4 x 100 ns 2 x 200 ns	6 x 100 ns	6 x 100 ns	-
	yes	-	9 x 50 ns	-	8 x 70 ns

4.3.1 Conformational dynamics of CsgG and the CsgG-CsgF complex

Firstly, the conformational dynamics of the uncomplexed CsgG and the CsgG-CsgF complex were investigated. The simulation systems comprised proteins embedded in the POPC lipid bilayer and

submerged in 1 M KCl. Six independent simulations were performed for each system (2 x 200 ns and 4 x 100 ns).

To assess the conformational drift of CsgG during the simulations, the root mean square deviation (RMSD) of the protein backbone (Cα atoms) from its initial conformation was calculated (Figure 4.2 and Table 4.2). The uncomplexed CsgG exhibited reduced conformational drift during the simulations (RMSD ~ 0.18-0.22 nm) compared to when in complex with CsgF (RMSD ~ 0.22-0.30 nm). The same was the case for the eyelet loops that form the CsgG eyelet loop region (residues 47-58); the RMSD remained ~ < 0.10 nm in five simulations of uncomplexed CsgG, compared to ~ 0.25-0.35 nm in simulations of the CsgG-CsgF complex. The RMSD of CsgF in the CsgG-CsgF complex plateaued to values of ~ 0.32-0.42 nm in all simulations.

To assess the flexibility of the protein domains, the root mean square fluctuation (RMSF) and B-factors of individual residues were calculated during the last 100 ns of the 200 ns simulations, as the proteins exhibit the least conformational drift during this time (RMSD has plateaued). In both uncomplexed CsgG and the CsgG-CsgF complex, the loops in CsgG were the most flexible domains. The RMSF of residues forming the extended loops in the vestibule mouth (residues 102-112) and the short loops near the C termini (residues 240-246) was as high as ~ 0.60 nm in some simulations. The eyelet loops forming the CsgG constriction region were also comparatively more flexible than the rest of the protein, with RMSF of residues ranging between ~ 0.05-0.17 nm. The least flexible region in both uncomplexed CsgG and the CsgG-CsgF complex was the transmembrane β-barrel. The short turns connecting the β-sheets (residues 191-198) were noticeably more rigid in the CsgG-CsgF complex (RMSF ~0.07-0.20 nm) than in uncomplexed CsgG (RMSF ~ 0.20-0.35 nm).

In CsgF, the C termini were the most flexible, with RMSF of residues 30-35 progressively increasing from ~ 0.30-0.90 nm (Figure 4.3). This is expected as the C termini protrude from the CsgG β-barrel and thus are unrestrained and disordered.

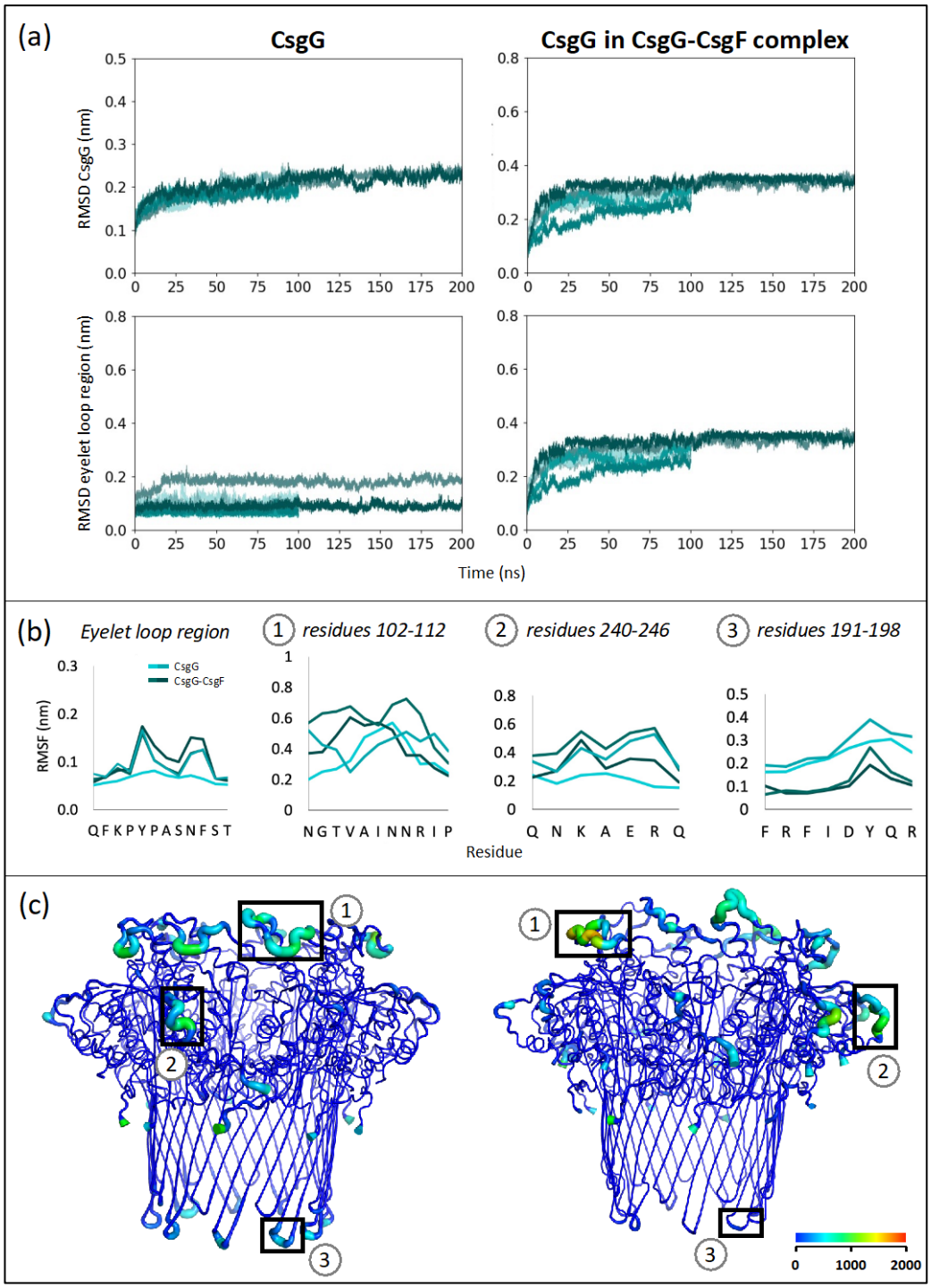

Figure 4.2: Conformational drift and flexibility of CsgG when uncomplexed and in the CsgG-CsgF complex, in 0 V. (a) RMSD of CsgG and the eyelet loop region compared to their initial conformation at 0 ns (backbone Cα atoms) is plotted over time for six independent simulations. (b) RMSF of residues in domains labelled in panel (c) during 100-200 ns in two simulations of uncomplexed CsgG and the CsgG-CsgF complex. RMSF of the eyelet loop region residues are average values for nine monomers. (c) CsgG coloured according to B-factor values of residues during 100-200 ns in a simulation of uncomplexed CsgG (left) and the CsgG-CsgF complex (right). The widening of the tube also indicates regions with higher B-factor values.

Table 4.2. RMSD of the protein backbone (Cα atoms) from its initial conformation at 100 ns in six independent simulations in 0 V.

System	RMSD CsgG (nm)	RMSD CsgG eyelet loops (nm)	RMSD CsgF (nm)
CsgG	~ 0.18-0.22	~ 0.05-0.18	-
CsgG-CsgF complex	~ 0.22-0.30	~ 0.25-0.35	~ 0.32-0.42

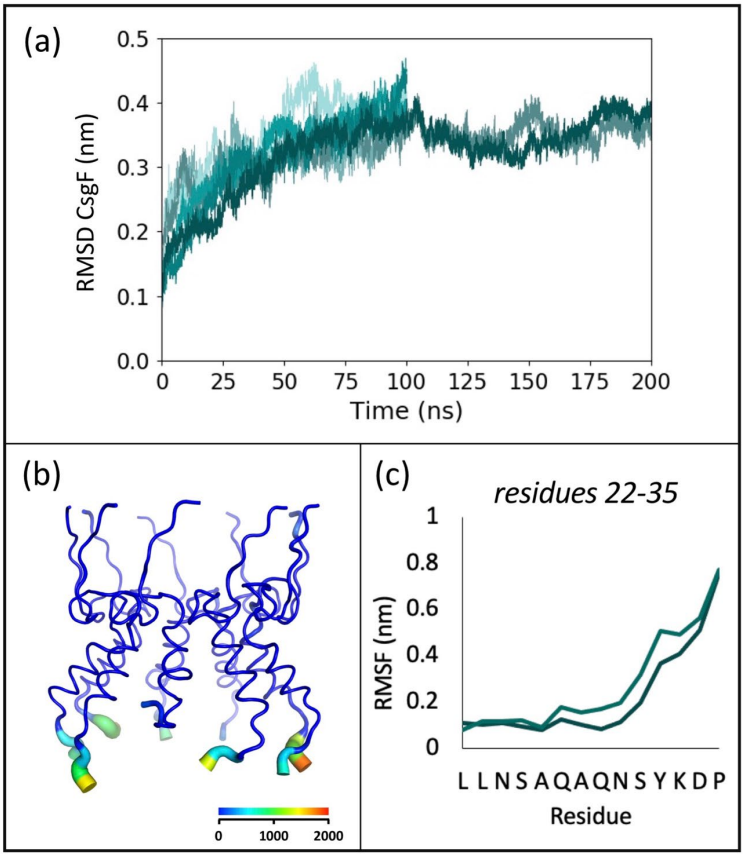

Figure 4.3: Conformational drift and flexibility of CsgF in the CsgG-CsgF complex, in 0 V. (a) RMSD of CsgF compared to its initial conformation at 0 ns (backbone Cα atoms) is plotted over time for six independent simulations. (b) CsgF coloured according to B-factor values of residues during 100-200 ns in a simulation of the CsgG-CsgF complex. The widening of the tube also indicates regions with higher B-factor values. (c) RMSF of residues in the C terminus of a CsgF monomer during 100-200 ns in two simulations of the CsgG-CsgF complex.

4.3.2 Stability of CsgG and the CsgG-CsgF complex under an applied electric field

When employing nanopores for DNA sequencing, the pores must withstand the electric field that propels DNA through the pore. The stability of uncomplexed CsgG and the CsgG-CsgF complex was thus evaluated under applied electric fields of 0.05 V nm^{-1} and 0.075 V nm^{-1}, equivalent to 0.9 V and 1.6 V, respectively, across the membrane. The electric fields are roughly five and nine times higher than 0.18 V used for DNA sequencing and were used to amplify any variations in protein dynamics.

Under an applied electric field of 0.05 V nm^{-1}, the uncomplexed CsgG showed a slightly higher deviation from its initial conformation than in the absence of an electric field, with plateau RMSD values of ~ 0.20-0.27 nm in six independent simulations (Table 4.3 and Figure 4.4). This was mainly due to the eyelet loops becoming more flexible; up to two eyelet loops were observed to move upwards into the vestibule of CsgG to varying degrees, which resulted in the RMSD of the eyelet loop region increasing to ~ 0.25-0.40 nm by 100 ns in five simulations (compared to ~ < 0.10 nm in 0 V). In one simulation, an eyelet loop flipped upwards during ~ 48-60 ns, and the RMSD of the eyelet loop region during this period increased from ~ 0.48 nm to ~ 0.65 nm. The loop conformation remained unchanged following this, hence the RMSD plateaued to ~ 0.65 nm by 100 ns.

Table 4.3. Summary of RMSD of the protein backbone Cα atoms at 100 ns in 0.05 V nm^{-1}, in six independent simulations.

System	RMSD CsgG (nm)	RMSD CsgG eyelet loops (nm)	RMSD CsgF (nm)
CsgG	~ 0.20-0.27	~ 0.25-0.40	-
CsgG-CsgF complex	~ 0.17-0.23	~ 0.16-0.24	~ 0.26-0.34

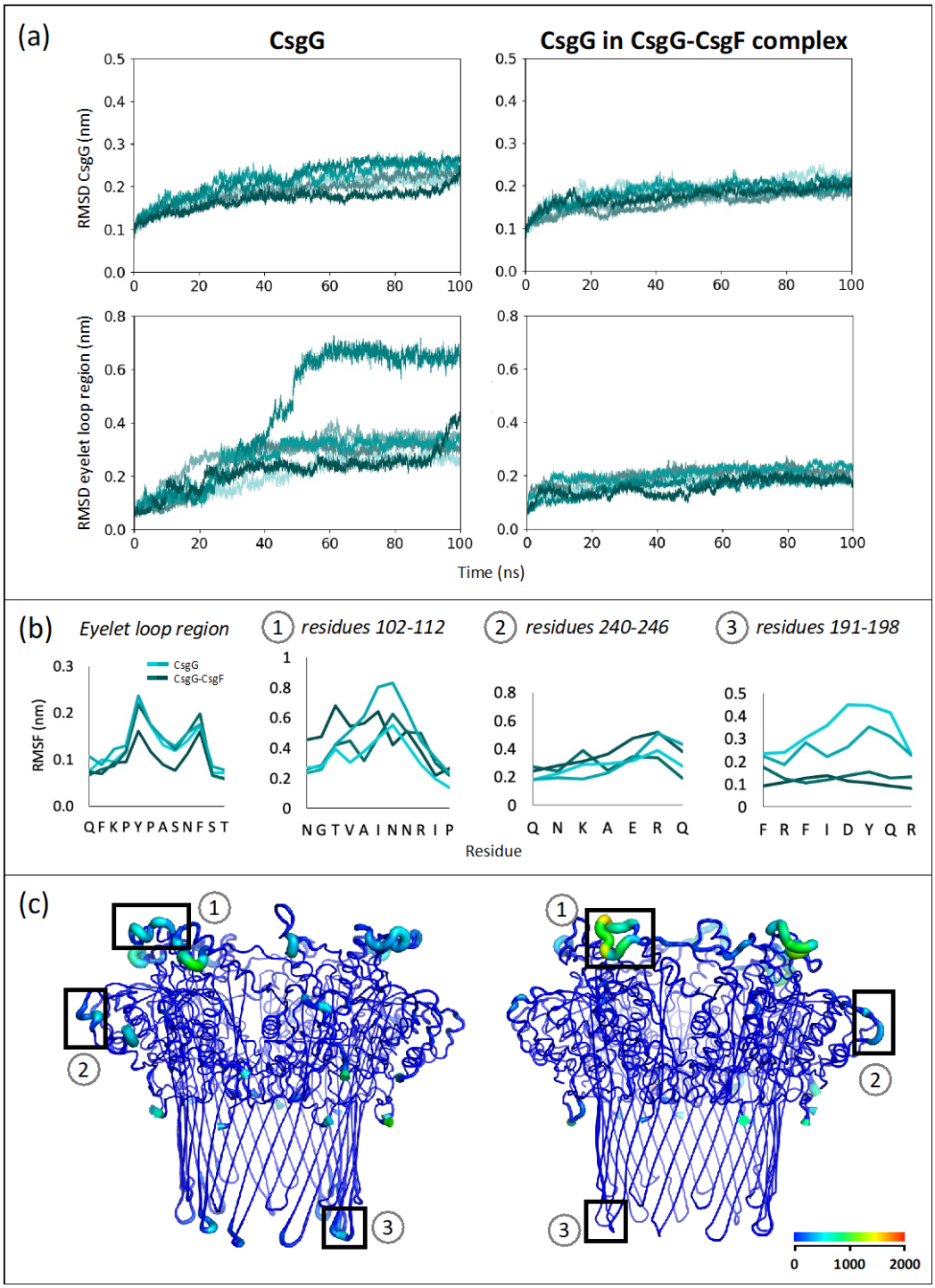

Figure 4.4: Conformational drift and flexibility of CsgG when uncomplexed and in the CsgG-CsgF complex, in 0.05 V nm^{-1}. (a) RMSD of CsgG and the eyelet loop region compared to their initial conformation at 0 ns (backbone Cα atoms) is plotted over time for six independent simulations. (b) RMSF of residues in domains labelled in panel (c) during 50-100 ns in two simulations of uncomplexed CsgG and the CsgG-CsgF complex. RMSF of the eyelet loop region residues are average values for nine monomers. (c) CsgG coloured according to B-factor values of residues during 50-100 ns in a simulation of uncomplexed CsgG (left) and the CsgG-CsgF complex (right). The widening of the tube also indicates regions with higher B-factor values.

The impact of the flipping of the eyelet loops on the dynamics of CsgG was ascertained by principal component analysis (PCA). PCA is a linear dimension-reduction method that extracts global collective motions of the protein by mapping its coordinates to a linear combination of orthogonal vectors known as principal components (PCs). The motions of the CsgG backbone were characterised before and following the flipping of the eyelet loops (referred to as 'pre-eyelet loop flipping' and 'post-eyelet loop flipping' from here on). The first ten principal components (PCs) were sufficient to describe > 90 % of the total backbone fluctuations of the protein.

The porcupine plots in Figure 4.5 show the direction and the magnitude of the movement of CsgG domains described by the 1st PC, which accounts for the most variance, i.e., the largest uncorrelated motion observed. The 1st PC accounts for ~ 45 % and ~52 % of the variance for pre- and post-eyelet loop flipping, respectively. The movement of the loops forming the vestibule mouth, the loops near the C termini in the vestibule region, and the short turns of the β-barrel dominated the dynamics of CsgG, both pre- and post-eyelet loop flipping. The direction and the scale of these motions varied amongst independent simulations; however, generally, more monomers were observed to participate in these motions post-eyelet loop flipping. The conformations of CsgG projected on the subspace spanned by the first two PCs show that the protein was able to explore more of the conformational space post-eyelet loop flipping (Figure 4.5). These results indicate that CsgG is more conformationally labile following the flipping of the eyelet loops. However, an examination of the secondary structure of components shows that this does not affect the CsgG structure in any significant way (Figure 4.6).

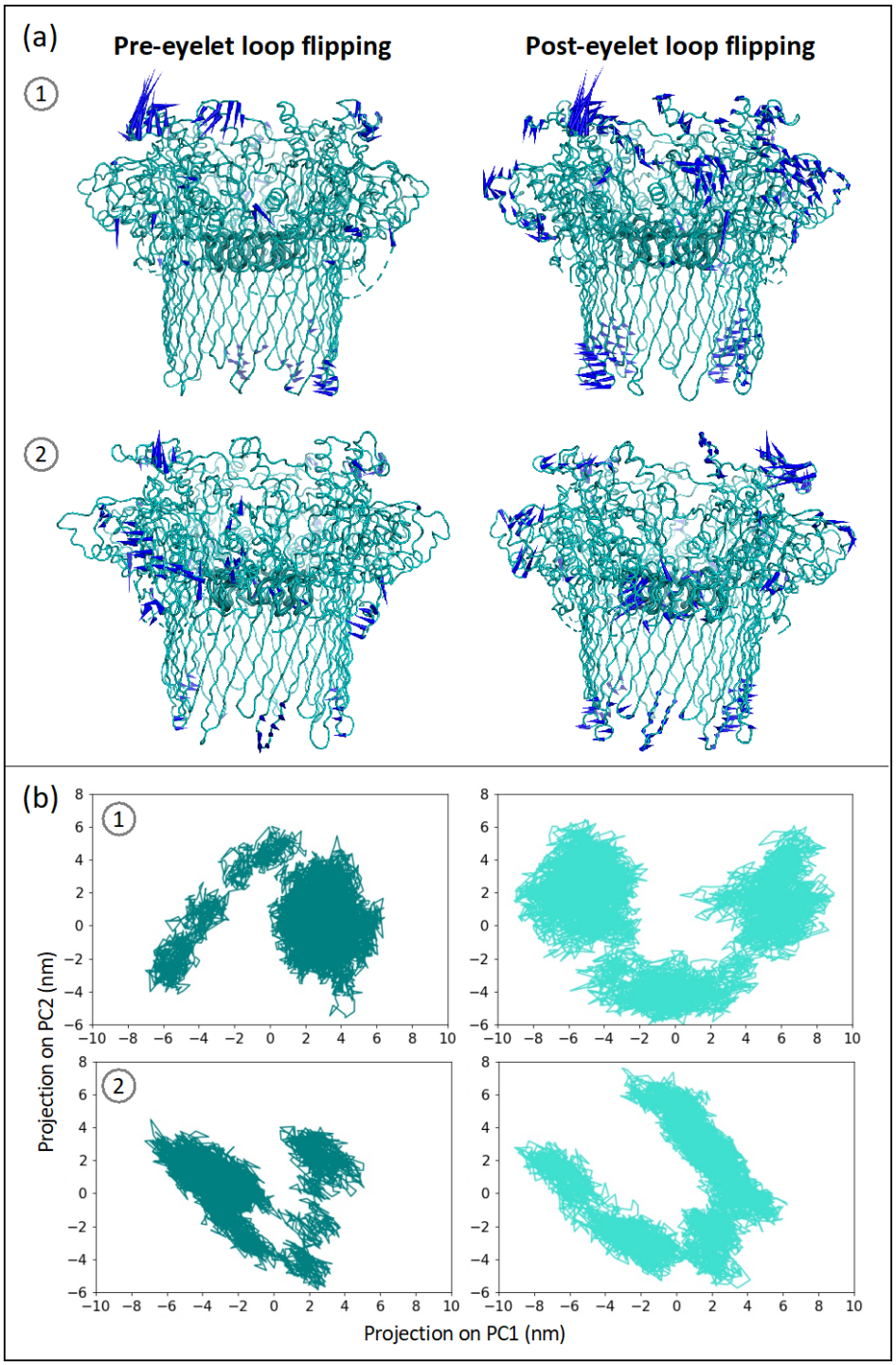

Figure 4.5: Principal components analysis performed for the CsgG backbone, from two simulations in 0.05 V nm^{-1} divided into conformations before and following the flipping of the eyelet loops (pre-eyelet loop flipping and post-eyelet loop flipping, respectively). (a) The motions described by PC1 are shown as arrows in the porcupine plots. The direction and the width of the arrows represent the direction and movement (> 0.3 nm) of the CsgG domains. (b) Projection of the conformations of CsgG on the subspace spanned by the first two PCs.

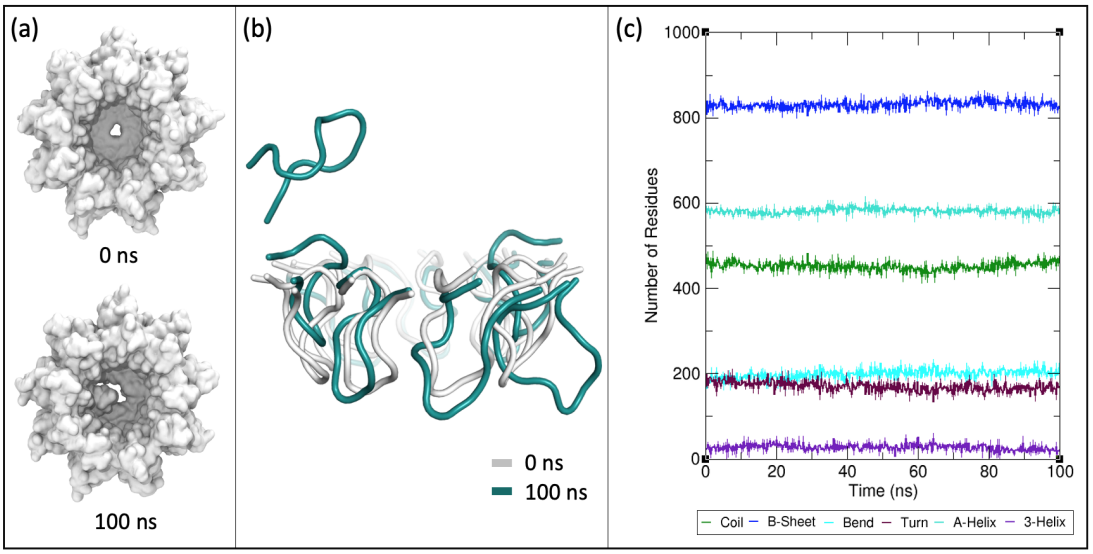

Figure 4.6: CsgG in 0.05 V nm^{-1}. Panel (a) shows the periplasmic view of CsgG at 0 ns and 100 ns after an eyelet loop moved upwards in the CsgG vestibule. A closer side-view of the eyelet loop region is shown in panel (b). Panel (c) shows the time evolution of secondary structure components of CsgG.

Interestingly, the upwards movement of the eyelet loops observed in simulations of CsgG was not observed in the CsgG-CsgF complex. Unlike uncomplexed CsgG, the CsgG-CsgF complex exhibited reduced conformational drift from its initial structure in 0.05 V nm^{-1} compared to in the absence of an electric field, with a lower plateau RMSD of ~ 0.17-0.23 nm for CsgG, ~ 0.16-0.24 nm for the CsgG eyelet loop region (Figure 4.4), and ~ 0.26-0.34 nm for CsgF (Figure 4.7). Additionally, whilst the flexibility of the short turns linking the β-sheets (residues 191-198) increased in uncomplexed CsgG (RMSF ~ 0.20-0.50 nm), it was not greatly affected by the electric field in the CsgG-CsgF complex (RMSF ~ 0.10-0.20 nm). The flexibility of the loops in the CsgG vestibule was not greatly affected by the electric field in both uncomplexed CsgG and the CsgG-CsgF complex (Figure 4.4).

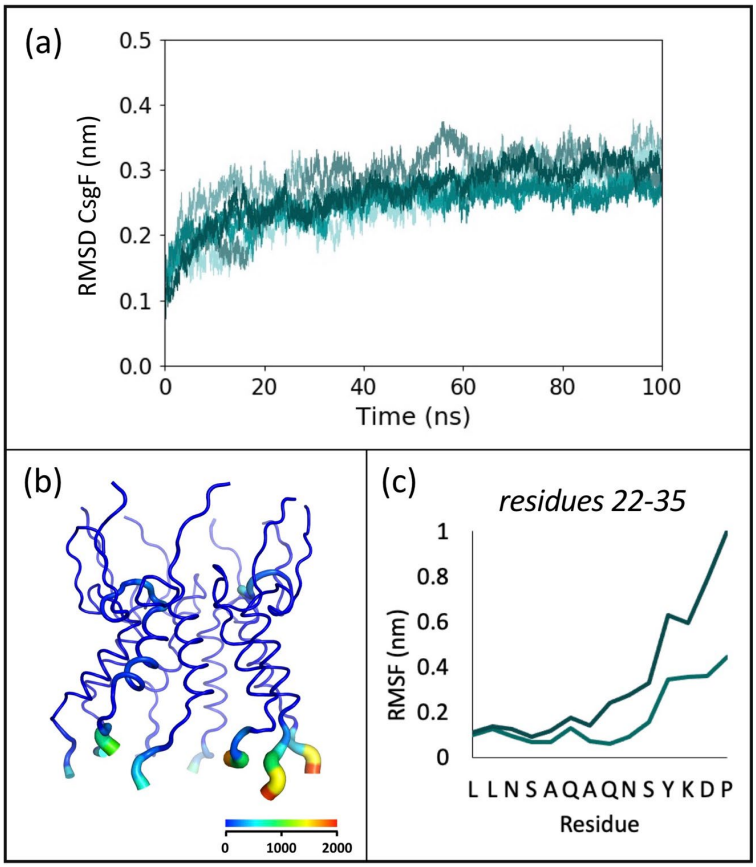

Figure 4.7: Conformational drift and flexibility of CsgF in the CsgG-CsgF complex, in 0.05 V nm^{-1}. (a) RMSD of CsgF compared to its initial conformation at 0 ns (backbone Cα atoms) is plotted over time for six independent simulations. (b) CsgF coloured according to B-factor values of residues during 50-100 ns in a simulation of the CsgG-CsgF complex. The widening of the tube also indicates regions with higher B-factor values. (c) RMSF of residues in the C terminus of a CsgF monomer during 50-100 ns in two simulations of the CsgG-CsgF complex.

CsgG was unstable under a higher electric field strength of 0.075 V nm^{-1} (Figure 4.8). The inter-monomer hydrogen bonds between the backbone atoms of two monomers were observed to break in the CsgG transmembrane β-barrel within 20-30 ns in six independent simulations. However, the interactions between these monomers in the vestibule region persisted following β-barrel destabilisation. Specifically, electrostatic interactions between N terminus Glu-8 in monomer 1 and Lys-179 of monomer 2 persisted for as long as 50 ns in all simulations. The inter-monomer interactions were further stabilised by a salt bridge formed between Lys-179 and Glu-210 in the same monomer. Additionally, the N terminus of monomer 1 remained wrapped around the adjacent monomer 2 in a conformation similar to that adopted in stable CsgG.

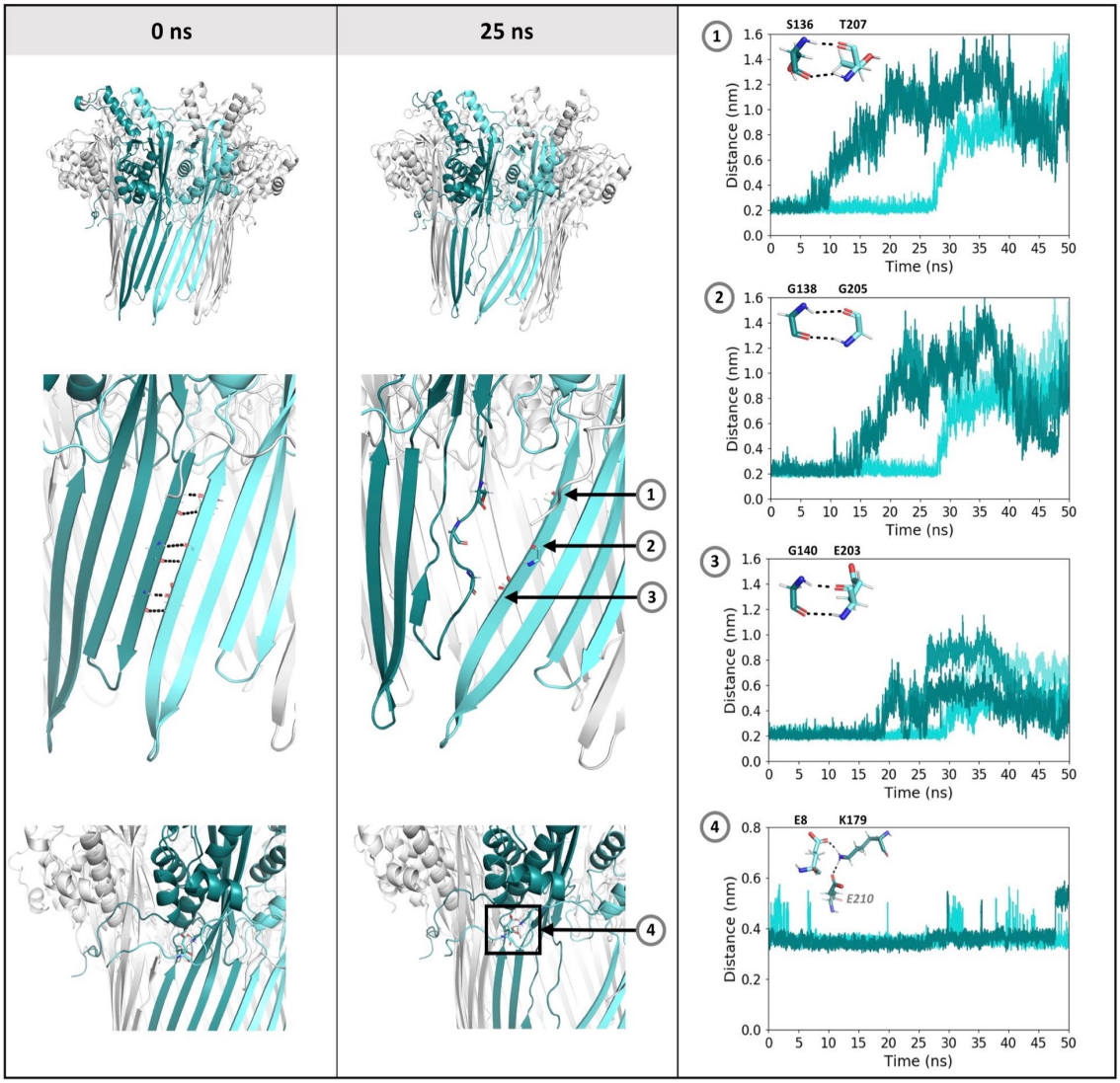

Figure 4.8: CsgG is unstable in 0.075 V nm^{-1}. The CsgG structure in one simulation at 0 ns and 25 ns is shown, with the two monomers that separate coloured in cyan and teal. The change in the inter-monomer interactions is illustrated by plotting the distance over time between the backbone atoms that form hydrogen bonds (plots 1-3), and the ammonium and carboxylate groups of lysine and glutamate residues that form electrostatic interactions between monomers. Data is from two independent simulations, plotted in teal and cyan. The inter-monomer interactions between residues are shown in the inset, with the hydrogen bonds marked by dashed lines.

The CsgG-CsgF complex remained stable in 0.075 V nm^{-1} in six independent simulations (Figure 4.9). The flexibility of the CsgG β-barrel, specifically the short turns that link the β-sheets, did not greatly differ compared to in the absence of an applied electric field. Additionally, CsgF also was more stable in an applied electric field. B-factor and RMSF analyses revealed that the CsgG vestibule loops were significantly less flexible when an electric field was applied. The RMSF values

for residues forming the short turns of the CsgG β-barrel did not greatly differ between simulations with and without an applied electric field, indicating that the CsgG β-barrel in the CsgG-F complex remained remarkably stable in high electric field strengths. The RMSD of CsgF was ~ 0.28-0.32 nm, similar to when under the electric field of 0.05 V nm^{-1} (Figure 4.10).

The interactions between CsgG and CsgF were characterised to determine the origins of the stability of CsgG when in complex with CsgF under a high electric field. CsgF is slotted inside the CsgG β-barrel in the CsgG-CsgF complex, with the N terminus of CsgF monomers lying close to the β-barrel surface, which kinks into the lumen due to the folding of conserved NPXFGG motif (residues 9–14), forming the CsgF constriction. This is followed by a 13-residue helix that angles outwards towards the CsgG β-barrel exit (Figure 4.1). The CsgF N terminus was observed to form a network of hydrogen bonds with the β-barrel residues of CsgG monomers, which remained stable in 0.075 V nm^{-1} (Figure 4.11). One CsgF monomer was observed to interact simultaneously with two of the CsgG monomers; the kinking of the N terminus enabled Asn-11 to form hydrogen bonds with Arg-142 in the β-sheet of an adjacent CsgG monomer, which stabilised the inter-monomer interactions in CsgG β-barrel (between Arg-142 and Glu-201 residues). The C terminus of CsgF also interacts with CsgG residues near the short turns of the β-barrel, which were comparatively less flexible than in uncomplexed CsgG.

It was observed that some of the interactions between CsgG and CsgF that were reported in the starting structures were not present in all monomer pairs in 0.075 V nm^{-1} (Figure 4.12). These include hydrogen bonds between CsgF Gly-1 and CsgG Asn-133 or Gln-153 residues, CsgF Thr-2 or Thr-4 and CsgG Thr-207, and CsgF Gln-6 and CsgG Gln-187 residues. Concurrently, additional interactions, including hydrogen bonds between residues in the CsgF C terminus and near the CsgG β-barrel exit, were observed in the simulations but not in the starting structure. Consequently, the short turns of the β-barrel (residues 191-198) remained ordered, with RMSF similar to when in 0 V and 0.05 V nm^{-1} (~ 0.05-0.20 nm). In addition to hydrogen bonding, Arg-8 in the CsgF N terminus also forms stabilising electrostatic interactions with Asp-149 and Glu-185 residues in CsgG (Figure 4.11).

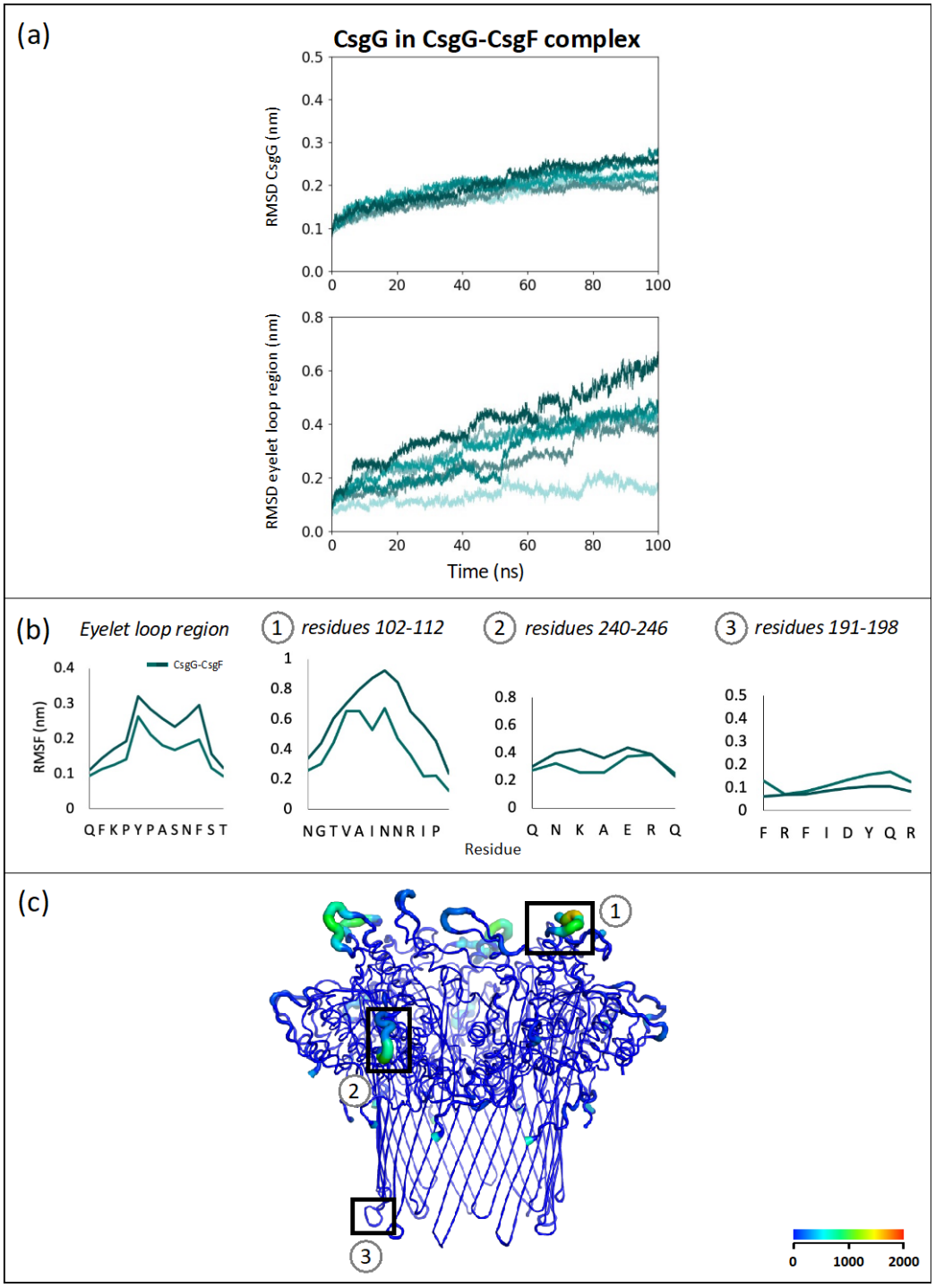

Figure 4.9: Conformational drift and flexibility of CsgG in the CsgG-CsgF complex, in 0.075 V nm^{-1}. (a) RMSD of CsgG and the eyelet loop region compared to their initial conformation at 0 ns (backbone Cα atoms) is plotted over time for six independent simulations. (b) RMSF of residues in domains labelled in panel (c) during 50-100 ns in two simulations of the CsgG-CsgF complex. RMSF of the eyelet loop region residues are average values for nine monomers. (c) CsgG coloured according to B-factor values of residues during 50-100 ns in a simulation of the CsgG-CsgF complex. The widening of the tube also indicates regions with higher B-factor values.

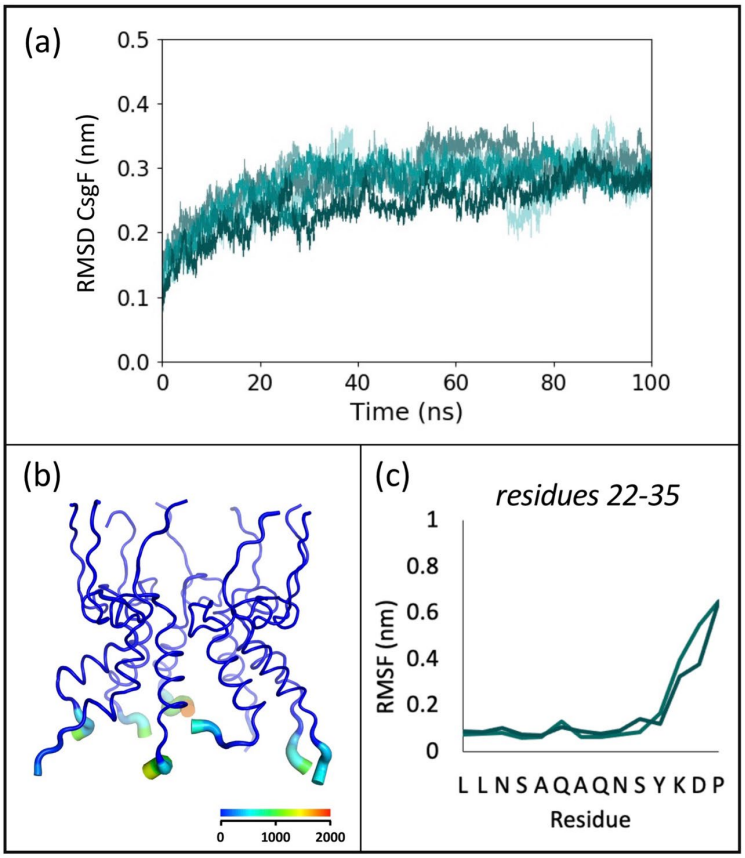

Figure 4.10: Conformational drift and flexibility of CsgF in the CsgG-CsgF complex, in 0.075 V nm^{-1}. (a) RMSD of CsgF compared to its initial conformation at 0 ns (backbone Cα atoms) is plotted over time for six independent simulations. (b) CsgF coloured according to B-factor values of residues during 50-100 ns in a simulation of the CsgG-CsgF complex. The widening of the tube also indicates regions with higher B-factor values. (c) RMSF of residues in the C terminus of a CsgF monomer during 50-100 ns in two simulations of the CsgG-CsgF complex.

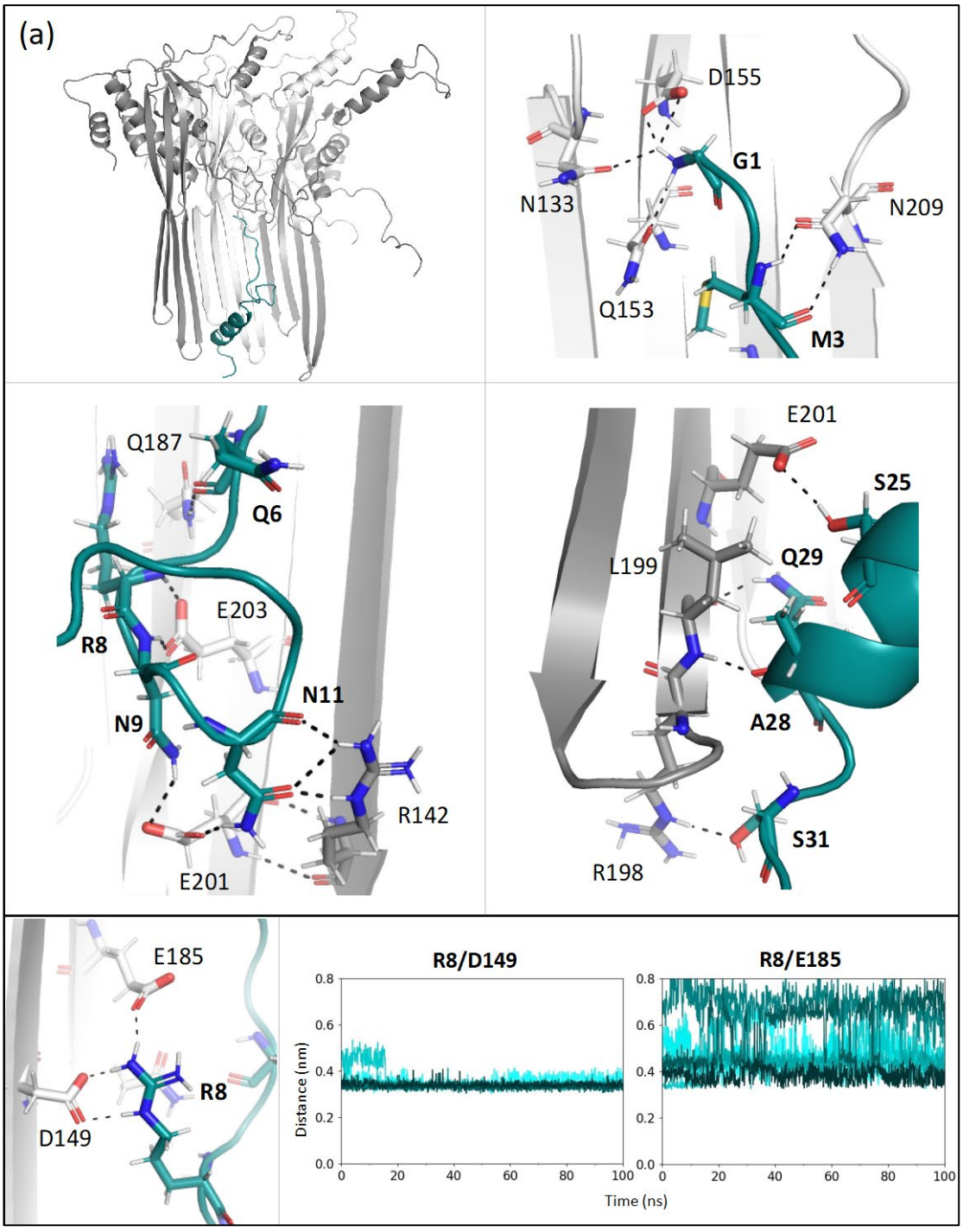

Figure 4.11: (a) In the CsgG-CsgF complex, residues in a CsgF monomer form hydrogen bonds with residues in three CsgG monomers. Hydrogen bonds are marked by dashed lines (< 0.32 nm). (b) CsgF Arg-8 forms electrostatic interactions with CsgG Asp-149 and Glu-185 residues. The bond distance is plotted over 100 ns for nine monomers in 0.075 V nm^{-1}.

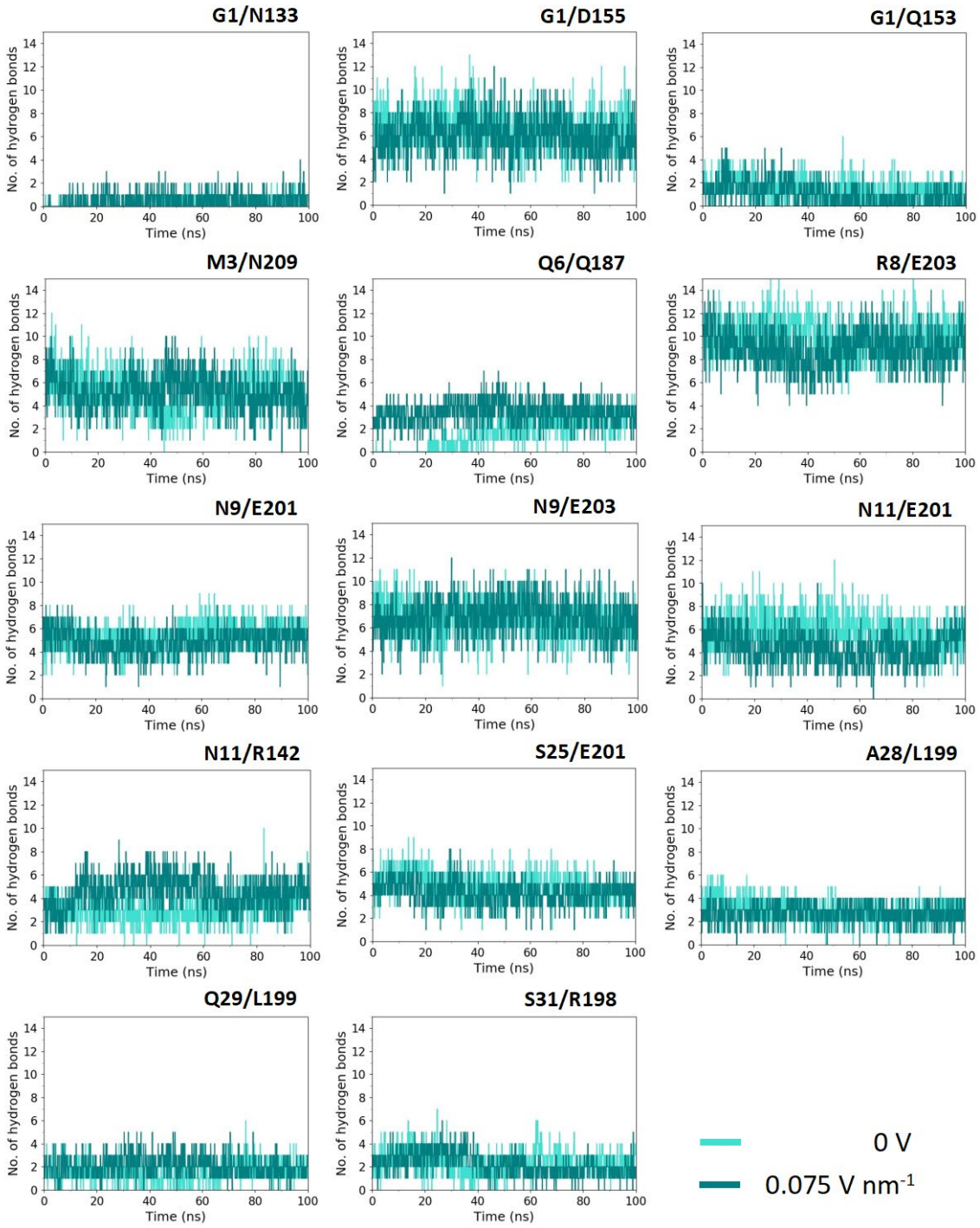

Figure 4.12: Hydrogen bonds between CsgF and CsgG monomers in the CsgG-CsgF complex, in 0 V and 0.075 V nm^{-1}. The number of hydrogen bonds between CsgF and CsgG residue pairs are plotted over 100 ns simulation.

4.3.3 DNA translocation

The translocation of DNA through uncomplexed CsgG and the CsgG-CsgF complex was investigated to study the rate of translocation and the conformations adopted by DNA in the two pores.

4.3.3.1 Benchmark steered MD simulations

DNA translocation was simulated by pulling a short ssDNA strand through the protein pores in steered MD simulations. To determine the pulling rate relevant for studying DNA translocation, simulations were performed in which a 20-nucleotide polyA ssDNA strand was pulled through uncomplexed CsgG and the CsgG-CsgF complex by applying a force on the 5' terminal nucleotide at rates of 0.25 nm ns^{-1} or 0.50 nm ns^{-1}. The 5' terminal nucleotide was initially positioned inside the CsgG eyelet loop region to mimic DNA translocation during DNA sequencing in which the strands remain threaded through the CsgG eyelet loop region.

The position of the DNA nucleotides as a function of time was plotted to compare DNA translocation at the two pull rates (Figure 4.13). DNA moved rapidly through uncomplexed CsgG and the CsgG-CsgF complex in simulations of 0.50 nm ns^{-1} pull rate, with up to 6 nucleotides exiting the pores by 20 ns. In uncomplexed CsgG, some nucleotides were halted inside the eyelet loop region for up to ~ 2 ns, but the DNA movement was unhindered and not affected by the strand interacting with residues of the pore lumen.

In the CsgG-CsgF complex, DNA nucleotides were halted inside the eyelet loop region more frequently than uncomplexed CsgG; however, this was for very short durations of up to ~ 3 ns. Some nucleotides were also briefly halted for up to ~ 2 ns in the CsgF constriction region, but despite this, DNA translocation did not greatly differ in CsgG and the CsgG-CsgF complex. Overall, when DNA was pulled at the rate of 0.50 nm ns^{-1}, the translocation of the strand was rapid through the pores as it was unable to interact with residues of the pore lumen due to the high pulling rate.

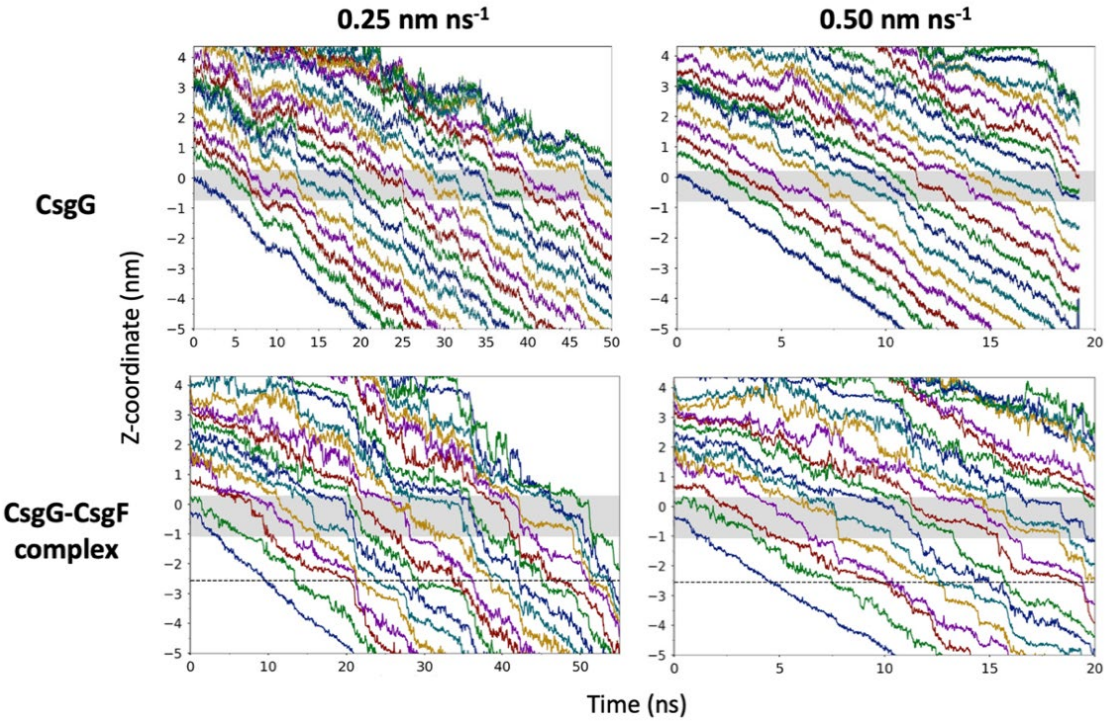

Figure 4.13: The translocation of 20-nucleotide polyA ssDNA through uncomplexed CsgG and the CsgG-CsgF complex, with the DNA pulled through at the rates of 0.25 nm ns^{-1} and 0.50 nm ns^{-1}, is measured as the Z coordinate of the centre of mass of nucleotides over time. The eyelet loop region is shaded in grey, and a dashed line marks the CsgF constriction.

DNA movement was comparatively slower when it was pulled through the pores at the rate of 0.25 nm ns^{-1}, with one nucleotide exiting the pores by 20 ns. The nucleotides were halted inside the eyelet loop region more frequently and for longer durations of ~ 3-5 ns. There were no substantial differences in DNA translocation between uncomplexed CsgG and the CsgG-CsgF complex; within 50 ns, 11 nucleotides were observed to exit uncomplexed CsgG, whilst 12 nucleotides were observed to exit the CsgG-CsgF complex. Thus, the rate of DNA pulling was not slow enough to resolve differences in DNA translocation through the pores that arise due to DNA-protein interactions. Additionally, the 3' terminal DNA segment was observed to coil above and in the vestibule during DNA translocation, which can affect the entry of nucleotides into the CsgG eyelet loop region and, therefore the DNA translocation rate. During DNA sequencing, long DNA strands are driven through the nanopores by a motor protein; hence the coiling of the DNA strand in the vestibule region observed in these simulations is not relevant to DNA sequencing. Therefore, all steered MD simulations from here on were run with a shorter 12-nucleotide DNA pulled through the pores at a slower rate of 0.15 nm ns^{-1}.

4.3.3.2 Translocation of short polyA ssDNA through CsgG and the CsgG-CsgF complex

A 12-nucleotide polyA ssDNA was pulled through the protein pore at a rate of 0.15 nm ns^{-1}. Four independent simulations were performed for uncomplexed CsgG and the CsgG-CsgF complex. The position of the centre of mass of the DNA nucleotides as a function of time was calculated to characterise the DNA translocation through the pore (Figures 4.14 and 4.15). Overall, DNA exited uncomplexed CsgG by ~ 60-70 ns and the CsgG-CsgF complex by ~ 73-75 ns. DNA moved through the CsgG eyelet loop constriction in a stepwise manner, with some nucleotides halted for as long as ~ 5 ns, due to the motion of nucleotides being hindered when translocating past the side chains of Phe-56, Asn-55, and Tyr-51 residues. Once DNA exited the eyelet loop region, its movement was unhindered in uncomplexed CsgG. In the CsgG-CsgF complex, DNA translocation was slower past the CsgF constriction due to nucleotides halting near Asn-17 residues for ~ 5 ns.

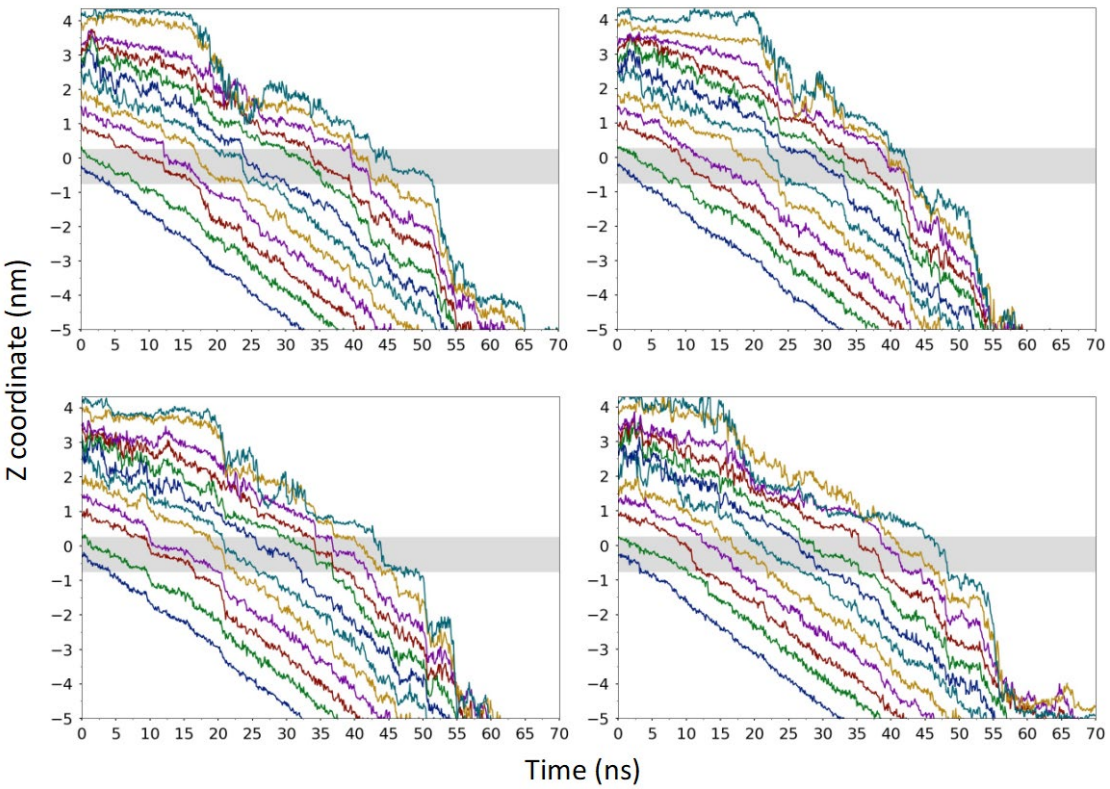

Figure 4.14: The translocation of polyA ssDNA through uncomplexed CsgG is measured as the Z coordinate of the centre of mass of nucleotides over time in four independent simulations. The eyelet loop region is shaded in grey.

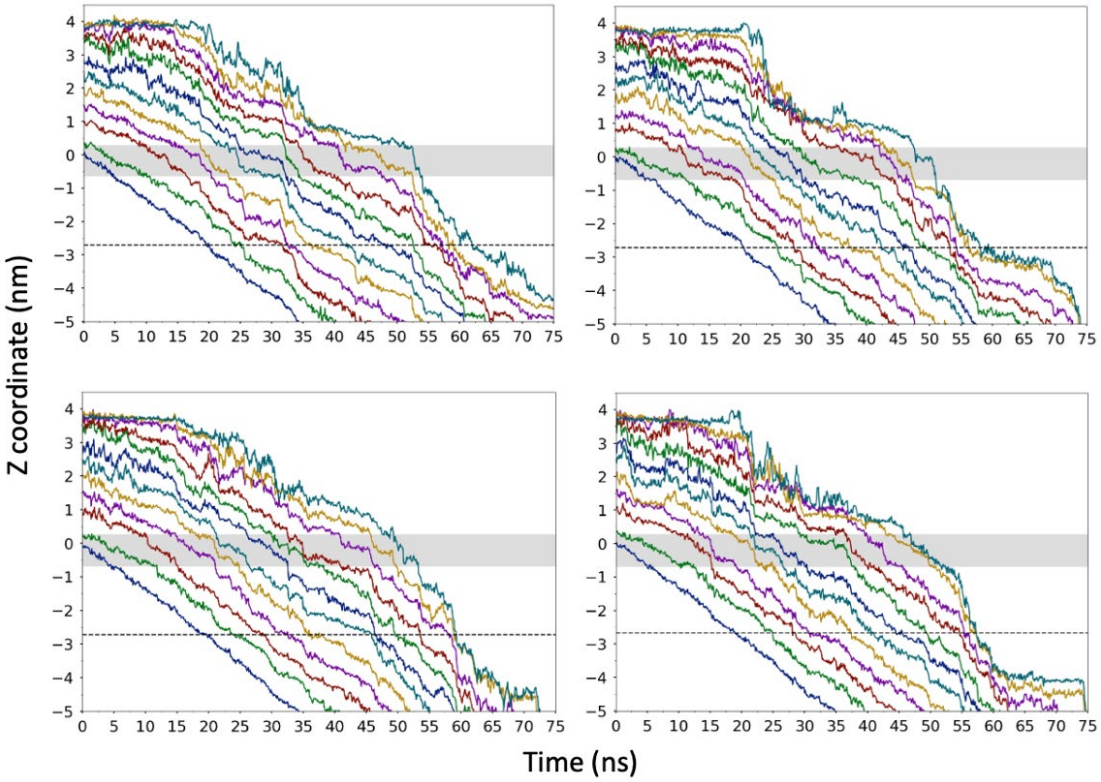

Figure 4.15: The translocation of polyA ssDNA through the CsgG-CsgF complex is measured as the Z coordinate of the centre of mass of nucleotides over time in four independent simulations. The eyelet loop region is shaded in grey, and a dashed line marks the CsgF constriction.

The rate of DNA translocation was calculated for regions of uncomplexed CsgG and the CsgG-CsgF complex as the distance spanned by a nucleotide per nanosecond. DNA translocation was divided into two regimes: (1) DNA is present inside the eyelet loop region, and (2) DNA is no longer present in the eyelet loop region (Table 4.4). Regime (1) mimics the behaviour of long DNA strands that remain threaded through the CsgG eyelet loop region during DNA sequencing.

In regime (1), the DNA translocation rate through the CsgG eyelet loop region is slower in the CsgG-CsgF complex (0.16 ± 0.07 nm ns^{-1}) than uncomplexed CsgG (0.18 ± 0.07 nm ns^{-1}), however the difference is not significant (p = 0.15). The translocation rate is similar (p = 0.25) through the CsgG β-barrel region (0.16 ± 0.01 nm ns^{-1}) and CsgF region in the CsgG-CsgF complex (0.15 ± 0.01 nm ns^{-1}), which indicates that DNA translocation rate is not greatly affected by CsgF.
DNA translocation through both uncomplexed CsgG and the CsgG-CsgF complex is substantially faster in regime (2), once DNA is no longer present in the CsgG eyelet loop region. DNA translocation is significantly slower (p = 0.0001) through the CsgF region (0.18 ± 0.04 nm ns^{-1}) compared to the CsgG β-barrel region (0.24 ± 0.09 nm ns^{-1}). Thus, while CsgF presents an

additional barrier to DNA translocation, the eyelet loop region has the greatest impact on the translocation rate of DNA.

Table 4.4. polyA ssDNA nucleotide translocation rates through regions of uncomplexed CsgG and the CsgG-CsgF complex.

Region	**Average DNA nucleotide translocation rate (nm ns^{-1}) ± SD**			
	Regime (1)		Regime (2)	
	CsgG	CsgG-CsgF complex	CsgG	CsgG-CsgF complex
CsgG eyelet loop region	0.18 ± 0.07	0.16 ± 0.07	N/A	N/A
CsgF region (β-barrel region in uncomplexed CsgG)	0.16 ± 0.01	0.15 ± 0.01	0.24 ± 0.09	0.18 ± 0.04

To investigate the differences in DNA translocation rate through uncomplexed CsgG and the CsgG-CsgF complex, the conformational behaviour of DNA and the DNA-protein interactions were next examined. The conformations adopted by the DNA strand in the pores were characterised *via* cluster analysis. In simulations of uncomplexed CsgG, the identified DNA conformations can be segregated into three groups, according to the position of the strand inside the pore: group 1, in which DNA is positioned in the vestibule region and the eyelet loop region, with the 5' terminus positioned below the eyelet loop region; group 2 in which the strand is positioned inside the eyelet loop region and the 5' terminus is in the β-barrel; and group 3 in which the strand is positioned in the β-barrel. As groups 1 and 3 are not relevant to DNA sequencing, the conformations of the strand in group 2 are considered (as done in a previous study of DNA translocation through model pores) [153]. Group 2 is formed by cluster populations comprising 59.3% of simulation time (~ 166 ns) (Figure 4.16). The DNA end-to-end distances of the conformations in this group ranged between ~ 5.7-6.9 nm, with the longer end-to-end distances corresponding to the more extended conformations of the strand as it translocated further downwards through the pore. The conformation of the DNA as it translocates uncomplexed CsgG is more extended compared to previously reported simulations of a model of the α-hemolysin β-barrel, in which the 12-nucleotide ssDNA was more coiled with end-to-end distances ranging from ~ 3.8-4.5 nm [81].

DNA was observed to form interactions with the residues in the eyelet loop region; Asn-55 and Tyr-51 residues formed hydrogen bonds with the nucleotides and backbone phosphate groups,

and Phe-56 and Tyr-51 residues interacted with the nucleotides *via* pi-stacking. During translocation, nucleotides interacting with Phe-56 residues were halted at the entrance of the eyelet loop region, whilst the nucleotides below formed interactions with Asn-55 and Tyr-51 residues. Once moving past Phe-56 residues, nucleotides were subsequently halted by Asn-55 and then by Tyr-51 residues during translocation. The combined effect of these interactions resulted in the DNA segment extending in the eyelet loop region. This is especially evident during the translocation of the DNA 3' terminus, which uncoiled as it moved through the eyelet loop region (Figure 4.17).

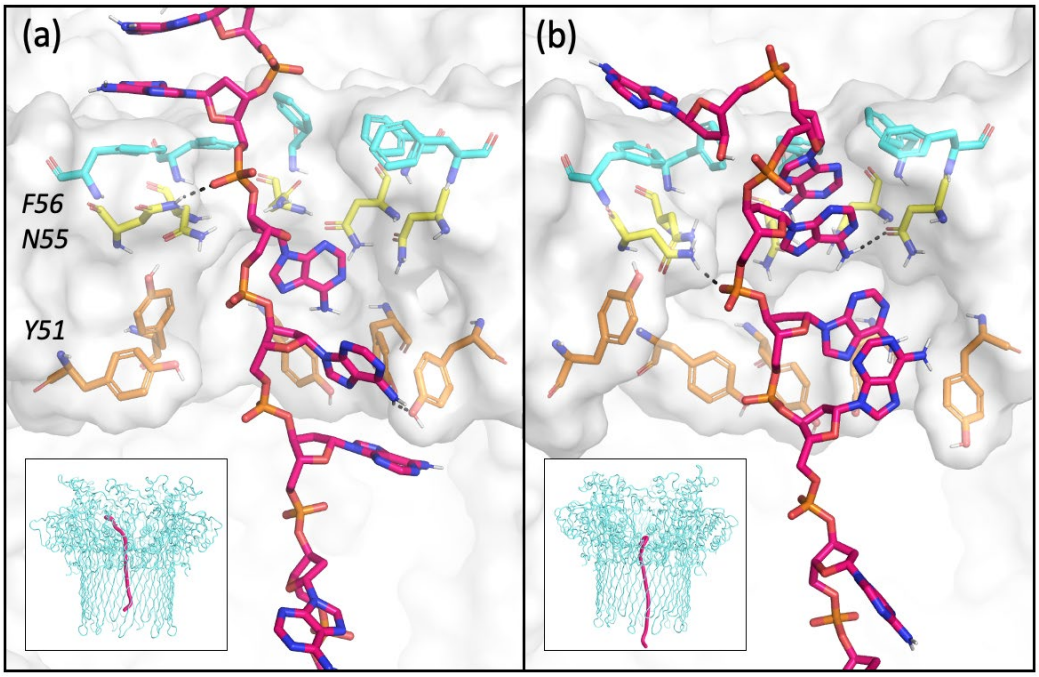

Figure 4.16: Representative polyA ssDNA conformations from two clusters of uncomplexed CsgG. DNA and the residues that interact with the nucleotides in the eyelet loop region are shown, with dashed lines marking hydrogen bonds. The inset shows the position of the DNA strand in uncomplexed CsgG.

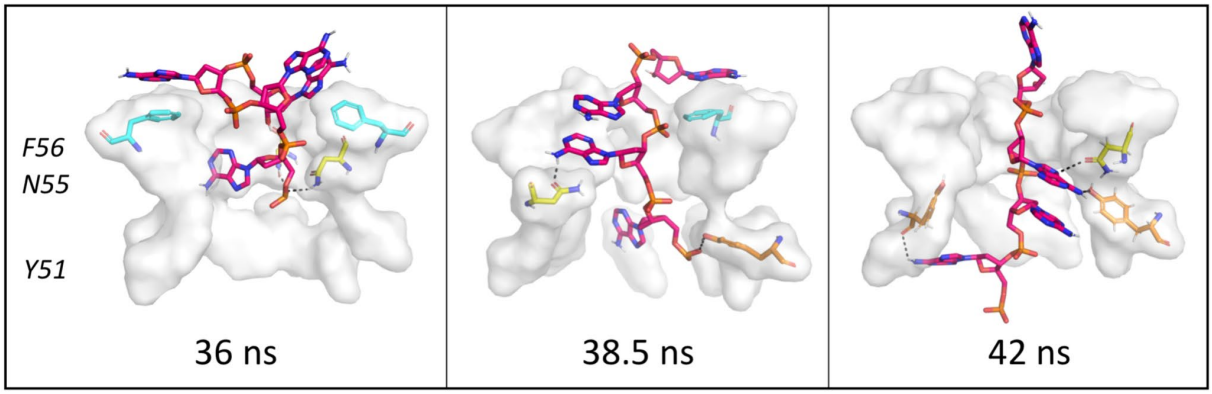

Figure 4.17: polyA ssDNA 3' terminus uncoiled as it translocated through the eyelet loop region in uncomplexed CsgG, due to interactions with residues in the eyelet loop region. The conformation of a 4-nucleotide segment at the DNA 3' terminus and the interacting protein residues are shown from one simulation, with dashed lines marking the hydrogen bonds.

In simulations of the CsgG-CsgF complex, the identified DNA conformations can be segregated into three groups, according to the position of the strand inside the pore: group 1, in which DNA is positioned in the vestibule region and the eyelet loop region, with the 5' terminus positioned below the eyelet loop region; group 2 in which the strand is positioned inside both CsgG and CsgF constriction regions; and group 3 in which the strand is positioned in CsgF. As group 1 is not relevant to DNA sequencing, the conformations of the strand in groups 2 and 3 are considered. Group 2 is formed by cluster populations comprising 50.7% of simulation time (~142 ns) (Figure 4.18a). The DNA end-to-end distances of the conformations in this group ranged between ~ 6.3-7.2 nm, with the longer end-to-end distances corresponding to the more extended conformations of the strand as it translocated further downwards through the pore.

DNA was observed to form interactions with the residues in the CsgG and the CsgF constriction regions, which resulted in the DNA segment retaining an extended conformation during translocation. As observed in uncomplexed CsgG, Asn-55 and Tyr-51 residues formed hydrogen bonds with the nucleotides and backbone phosphate groups, and Phe-56 and Tyr-51 residues interacted with the nucleotides *via* pi-stacking. Additionally, nucleotides in CsgF form hydrogen bonds with Asn-17 residues in the constriction region.

Group 3 consists of cluster populations comprising 22.1% simulation time (~ 62 ns), in which the DNA 3' terminus region is translocating through CsgF after exiting the CsgG eyelet loop region (Figure 4.18b). An Asn-24 residue also interacted with the DNA backbone phosphate group along with an Asn-17 residue in the CsgF constriction. DNA adopted group 3 conformations for a shorter duration than group 2, in which DNA is threaded through both CsgG and CsgF constrictions, which

is concurrent with the significant increase in translocation rate observed following DNA exit from the CsgG eyelet loop region.

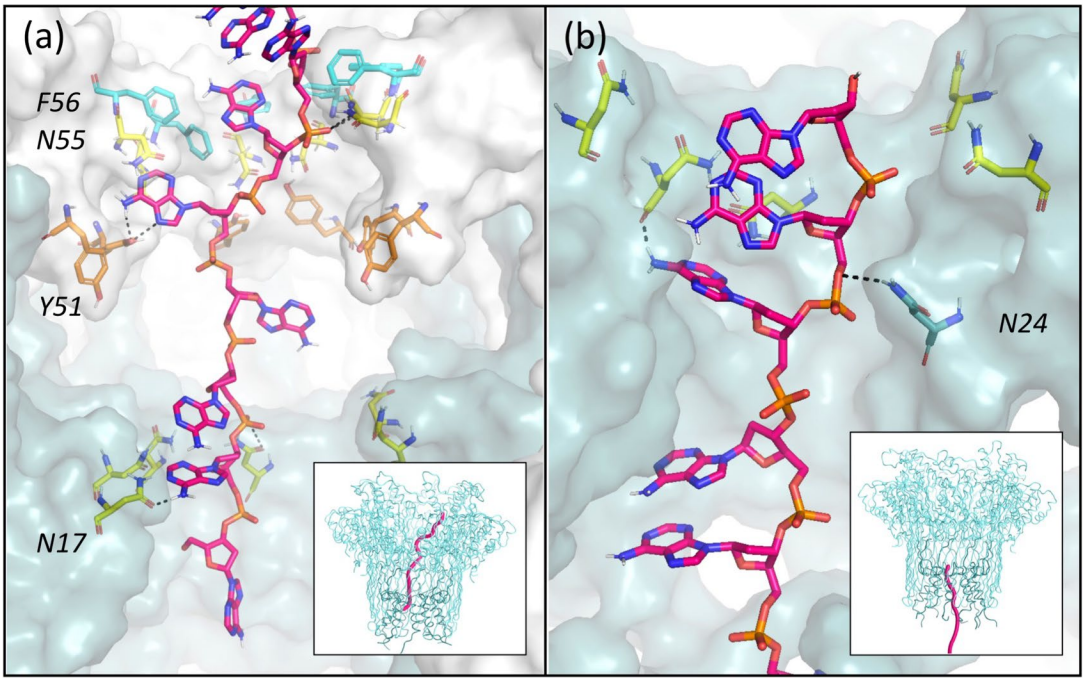

Figure 4.18: Representative polyA ssDNA conformations in two clusters of the CsgG-CsgF complex. DNA and the residues that interact with the nucleotides in the CsgG eyelet loop region and the CsgF constriction region are shown, with dashed lines marking hydrogen bonds. The inset shows the position of the DNA strand in the CsgG-CsgF complex.

To evaluate the interactions formed between DNA and the protein pores during translocation, the percentage of the simulation time that the DNA was within 0.4 nm of the residues was calculated (Figure 4.19). Overall, DNA primarily interacted with Phe-56, Asn-55, and Tyr-51 residues in the CsgG eyelet loop region in both uncomplexed CsgG and the CsgG-CsgF complex. DNA also interacted with Asn-17 residues in the CsgF constriction region, however, this did not occur as frequently as the interactions with the residues in the eyelet loop region. The higher frequency of DNA interactions with residues in the CsgG eyelet loop region than in CsgF suggests that the eyelet loop region has a greater impact on the DNA translocation rate.

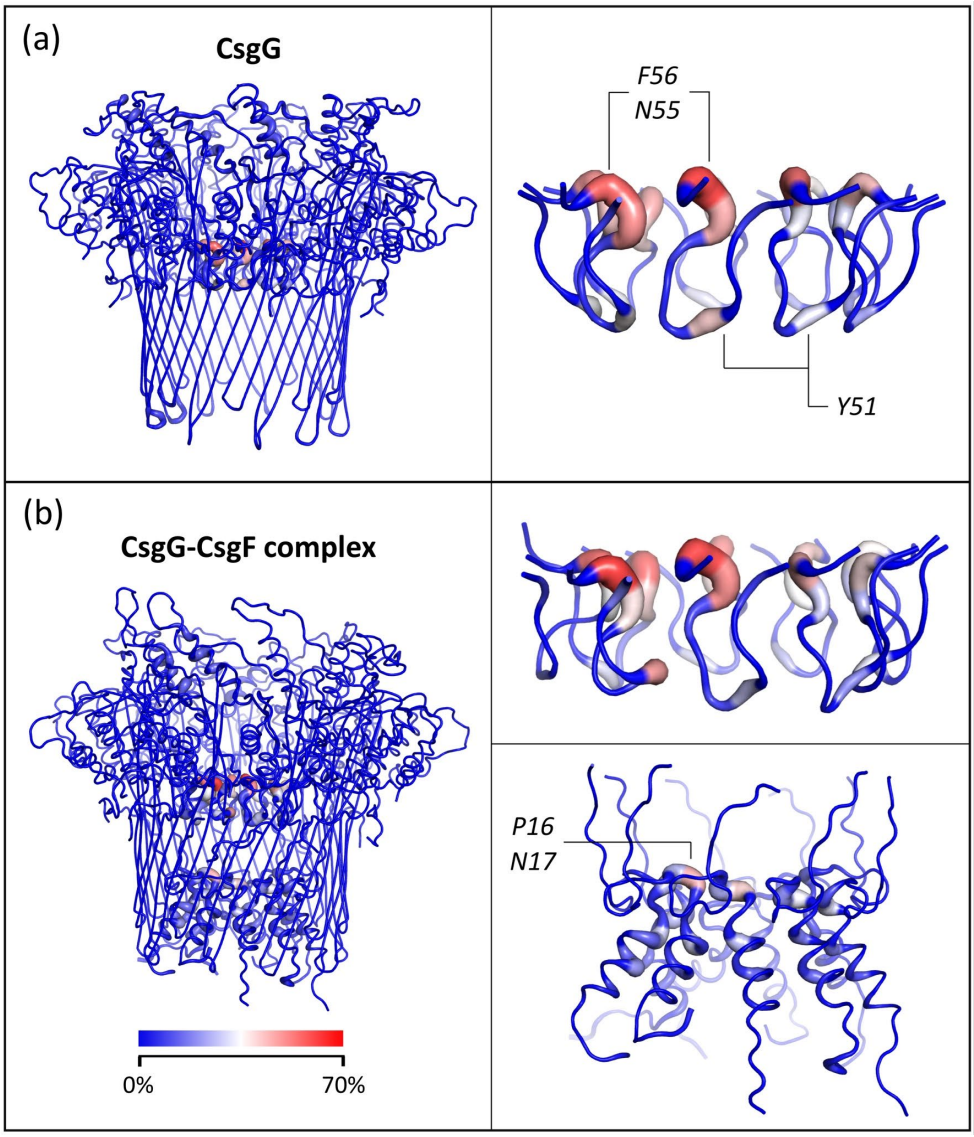

Figure 4.19: polyA ssDNA-protein interactions. CsgG (a) and the CsgG-CsgF complex (b) are coloured by the percentage of simulation time during which the residues interact with DNA in four independent simulations. An interaction is defined as an inter-atomic distance of < 0.4 nm. The eyelet loop region and CsgF are also shown.

During translocation, DNA was observed to retain a more extended conformation in the CsgG-CsgF complex compared to uncomplexed CsgG. The presence of CsgF increases the hydrophobicity of the channel formed by the CsgG-CsgF complex compared to uncomplexed CsgG, as it replaces the charged residues in the CsgG β-barrel with residues that are hydrophobic in nature (Figure 4.20). Previous studies have demonstrated that DNA retains a largely extended conformation during translocation through hydrophobic model pores based on a 14-stranded β-barrel architecture [122], and more so in pores containing two hydrophobic constriction regions compared to one [153]. Therefore, it is likely that DNA is retained in a more extended

conformation in the CsgG-CsgF complex due to the presence of two constriction regions and the hydrophobic nature of the pore.

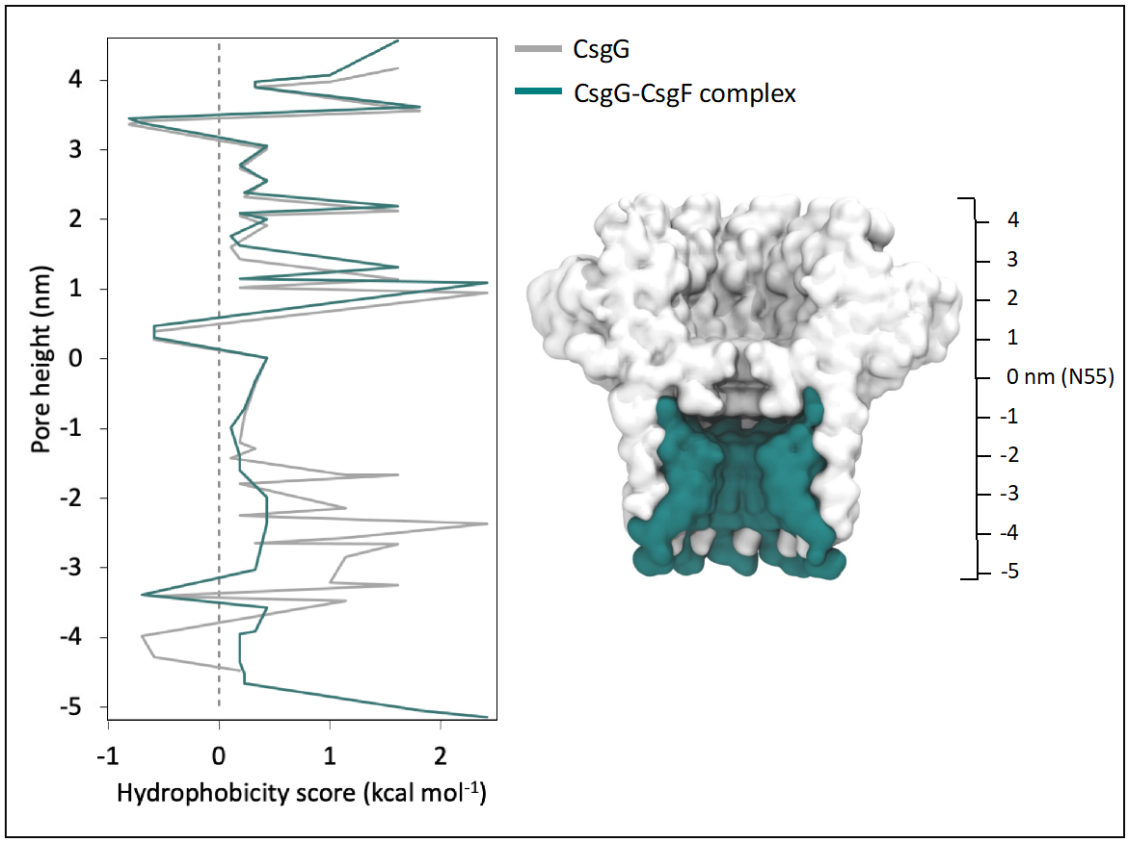

Figure 4.20: The hydrophobicity of the residues lining the pores formed by uncomplexed CsgG and the CsgG-CsgF complex is scored using the scale proposed by Wimley and White [197], which ranges from -0.81 kcal mol^{-1} for very hydrophobic residues to 2.41 kcal mol^{-1} for very hydrophilic residues. A cross-sectional side view of the CsgG-CsgF complex is shown on the right for reference.

4.3.3.3 Translocation of short polyC ssDNA through CsgG and the CsgG-CsgF complex

To evaluate the ability of uncomplexed CsgG and the CsgG-CsgF complex to discriminate between distinct nucleotides, a 12-nucleotide polyC ssDNA was also pulled through the protein pores. polyC exited uncomplexed CsgG by ~ 60 ns in three simulations and ~ 70 ns in one simulation and exited the CsgG-CsgF complex by ~ 65-70 ns in four simulations (Figures 4.21 and 4.22). polyC exited both pores earlier than polyA; the strand moved out of the CsgG-CsgF complex at most ~ 10 ns earlier (polyA = ~ 73-75 ns) and moved out of the eyelet loop region at most ~ 9 ns earlier (polyC = ~ 46-49 ns, polyA = ~ 55 ns). polyC also exited uncomplexed CsgG earlier than polyA in three simulations (polyA = ~ 60 ns in two simulations, ~ 65 ns in one simulation, and ~ 70 ns in one

simulation) but moved out of the eyelet loop region at a similar time as polyA in all simulations (~ 50 ns). As observed during polyA translocation, polyC traversed the eyelet loop constriction in a stepwise manner due to nucleotides being halted near Phe-56, Asn-55, and Tyr-51 residues for as long as 5 ns. In the CsgG-CsgF complex, DNA translocation was slower past the CsgF constriction due to nucleotides halting near Asn-17 residues for ~ 3-5 ns.

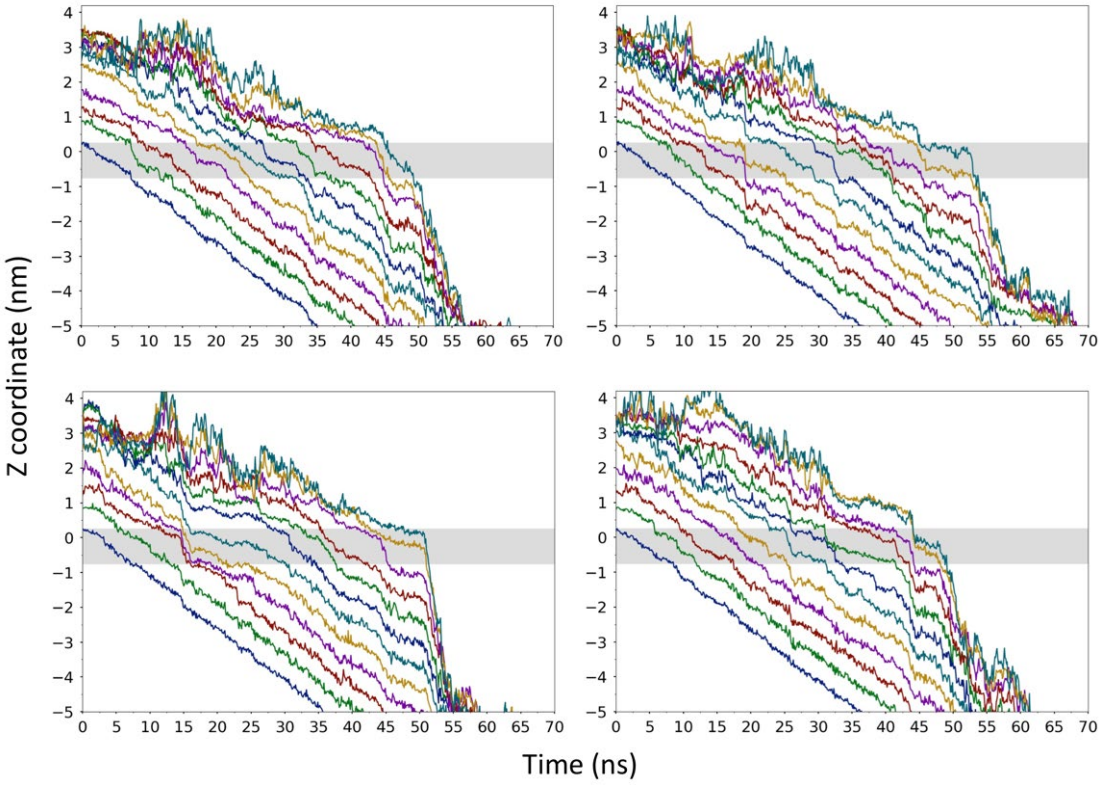

Figure 4.21: The translocation of polyC ssDNA through uncomplexed CsgG is measured as the Z coordinate of the centre of mass of nucleotides over time in four independent simulations. The eyelet loop region is shaded in grey.

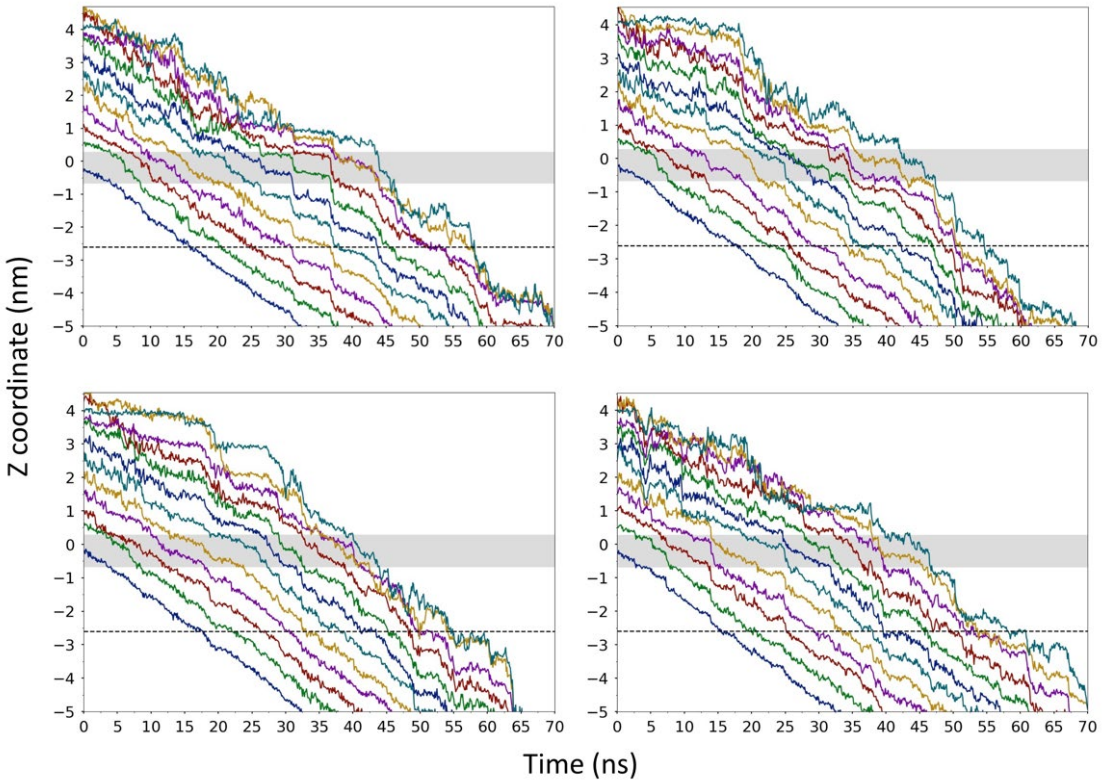

Figure 4.22: The translocation of polyC ssDNA through the CsgG-CsgF complex is measured as the Z coordinate of the centre of mass of nucleotides over time in four independent simulations. The eyelet loop region is shaded in grey, and a dashed line marks the CsgF constriction.

The translocation rates through regions of uncomplexed CsgG and the CsgG-CsgF complex were calculated for two regimes of DNA translocation (Table 4.5). In regime (1), during which DNA is translocating through the eyelet loop region, the translocation rate is not dramatically altered by CsgF. polyC translocated at the same rate through the eyelet loop region (0.17 ± 0.07 nm ns^{-1}) in uncomplexed CsgG and CsgG-CsgF complex, and the CsgG β-barrel and CsgF in the CsgG-CsgF complex (0.15 ± 0.01 nm ns^{-1}). Like polyA, the translocation of polyC in regime (2), after exiting the eyelet loop region, is substantially faster than in regime (1) through both protein pores. DNA translocation is significantly faster (p = 0.0022) through the CsgG β-barrel region (0.32 ± 0.24nm ns^{-1}) than through the CsgF region (0.18 ± 0.02 nm ns^{-1}). These results further prove that the CsgG eyelet loop region has the greatest impact on the translocation rate of DNA.

The translocation rates in regime (1) and through CsgF in regime (2) are similar for both polyA and polyC simulations. However, the centre of mass position of the DNA nucleotides (Figures 4.21 and 4.22) shows polyC exiting the protein pores earlier than polyA, especially the CsgG-CsgF complex. As the DNA 5' terminus is pulled through the pores at a constant rate of 0.15 nm ns^{-1}, the average

DNA nucleotide translocation rate is very similar for polyA and polyC despite the slight differences in velocities of nucleotides caused by their interactions with the protein. However, the force at which the DNA 5' terminus is pulled differs for both strands.

In constant velocity steered MD simulations, the force applied to the pull group is adjusted to maintain its translocation at a continuous rate. In simulations of uncomplexed CsgG and the CsgG-CsgF complex, the force experienced by the DNA 5' terminus is affected by the energy barriers encountered in its translocation path and the energy barriers encountered by other nucleotides above in the DNA strand. To overcome these barriers, a higher force is applied to the DNA 5' terminus to maintain a constant translocation rate.

The pulling force experienced by the DNA 5' terminus over time (average of four independent simulations) is shown in Figure 4.23. In simulations of uncomplexed CsgG, the pulling force during polyA translocation is ~ 50-100 kJ mol^{-1} nm^{-1} higher than polyC during ~ 10-40 ns. This corresponds to the time during which nucleotides at the centre of the strand translocated past the CsgG eyelet loop region. In simulations of polyA translocation through the CsgG-CsgF complex, the pulling force peaks at ~ 150-250 kJ mol^{-1} nm^{-1} on two occasions (~ 15-25 ns and ~ 40-50 ns), which correspond to the time during which nucleotides at the centre of the strand translocated past the eyelet loop region and the CsgF constriction. Overall, the pulling force was occasionally higher in polyA than in polyC systems, which indicates that polyA experienced greater barriers during translocation than polyC. This may be due to the larger steric bulk of the adenine in polyA compared to cytosine in polyC.

Table 4.5. polyC ssDNA nucleotide translocation rates through regions of uncomplexed CsgG and the CsgG-CsgF complex.

Region	**Average DNA nucleotide translocation rate (nm ns^{-1}) ± SD**			
	Regime (1)		Regime (2)	
	CsgG	CsgG-CsgF complex	CsgG	CsgG-CsgF complex
CsgG eyelet loop region	0.17 ± 0.07	0.17 ± 0.01	N/A	N/A
CsgF region (β-barrel region in uncomplexed CsgG)	0.15 ± 0.01	0.15 ± 0.01	0.32 ± 0.24	0.18 ± 0.02

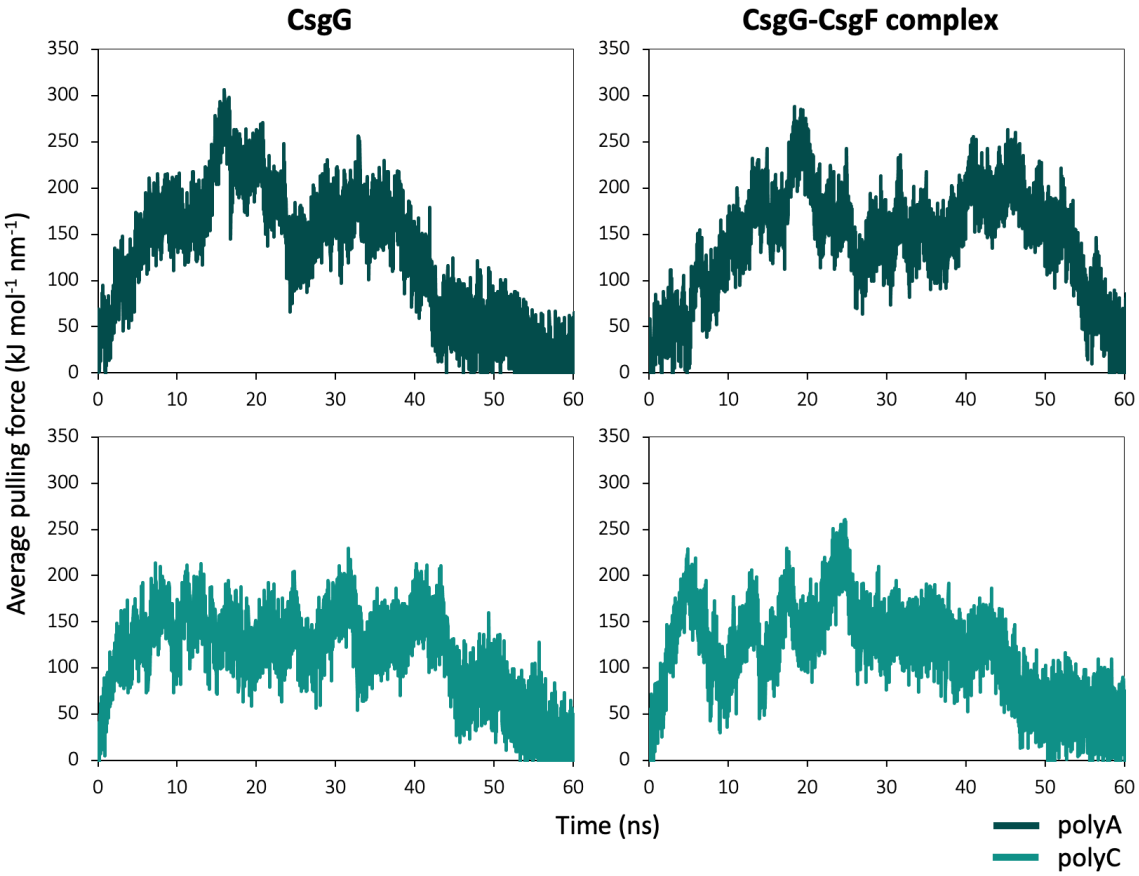

Figure 4.23: The pulling force experienced by DNA 5' terminus over time during polyA and polyC ssDNA translocation through uncomplexed CsgG and the CsgG-CsgF complex. The pulling force is an average of four independent simulations of each system.

Next, the conformational behaviour of DNA and the DNA-protein interactions during translocation was examined to investigate the differences in DNA translocation rate through uncomplexed CsgG and the CsgG-CsgF complex. Cluster analysis revealed that polyC is more coiled than polyA during translocation. In simulations of uncomplexed CsgG, group 2 cluster populations (relevant to DNA sequencing as the strand is translocating through the CsgG eyelet loop region) comprised 54.3% of simulation time (~152 ns). The end-to-end distances of DNA conformations in these clusters was shorter than polyA (polyC = ~ 5.7-6.5 nm, polyA = ~ 5.7-6.9 nm). polyC formed interactions with Phe-56, Asn-55, and Tyr-51 residues in the eyelet loop region in a similar manner to polyA (Figure 4.24).

For the CsgG-CsgF complex, group 2 is formed by cluster populations comprising 39.6% of simulation time (~111 ns); as a reminder, DNA is threaded through both CsgG and CsgF constriction regions in this group (Figure 4.25a). The DNA end-to-end distances of the conformations ranged between ~ 5.8-6.6 nm (compared to polyA = ~ 6.3-7.2 nm). DNA is retained

in a more extended conformation compared to uncomplexed CsgG due to the interactions formed between the nucleotides and residues in the CsgG and CsgF constriction regions.

Group 3 consists of cluster populations comprising 31.8% simulation time (~ 89 ns), in which the DNA 3' terminus region is translocating through CsgF after exiting the CsgG eyelet loop region (Figure 4.25b). DNA predominantly formed interactions with Asn-17 residues in the CsgF constriction. Like polyA, DNA adopted group 3 conformations for a shorter duration than group 2, which is consistent with the increase in translocation rate observed following DNA exit from the CsgG eyelet loop region into the CsgF constriction. Additionally, group 3 accounts for a longer duration than simulations of polyA, which is consistent with polyC exiting the eyelet loop region in the CsgG-CsgF complex earlier than polyA.

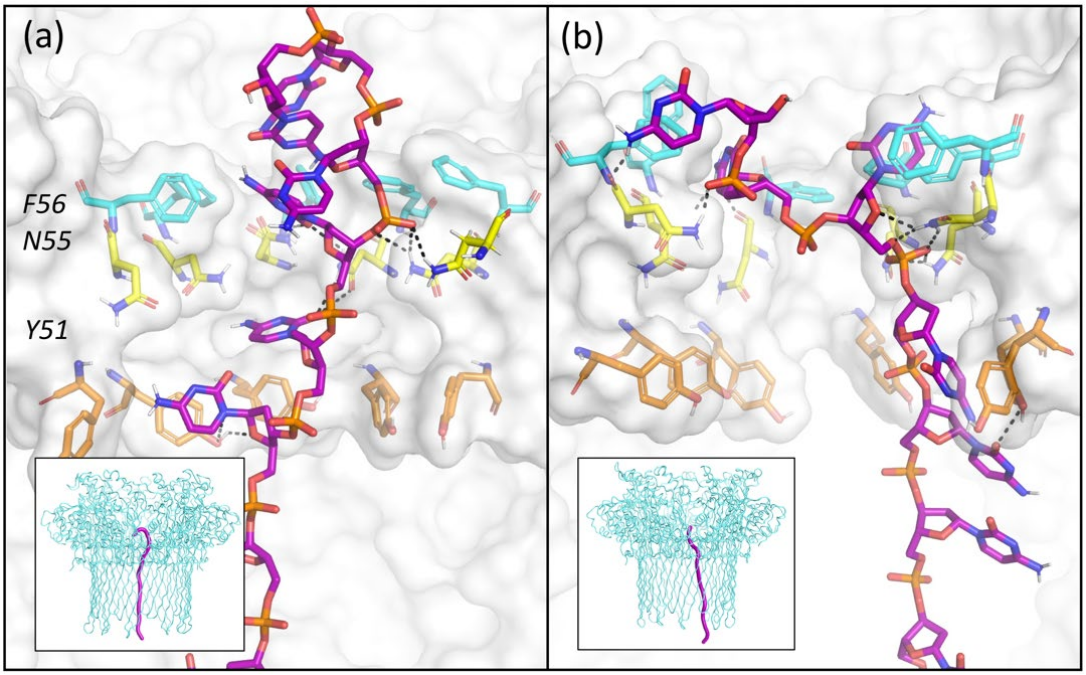

Figure 4.24: Representative polyC ssDNA conformations from two clusters of uncomplexed CsgG. DNA and the residues that interact with the nucleotides in the CsgG eyelet loop region are shown, with dashed lines marking hydrogen bonds. The inset shows the position of the DNA strand in uncomplexed CsgG.

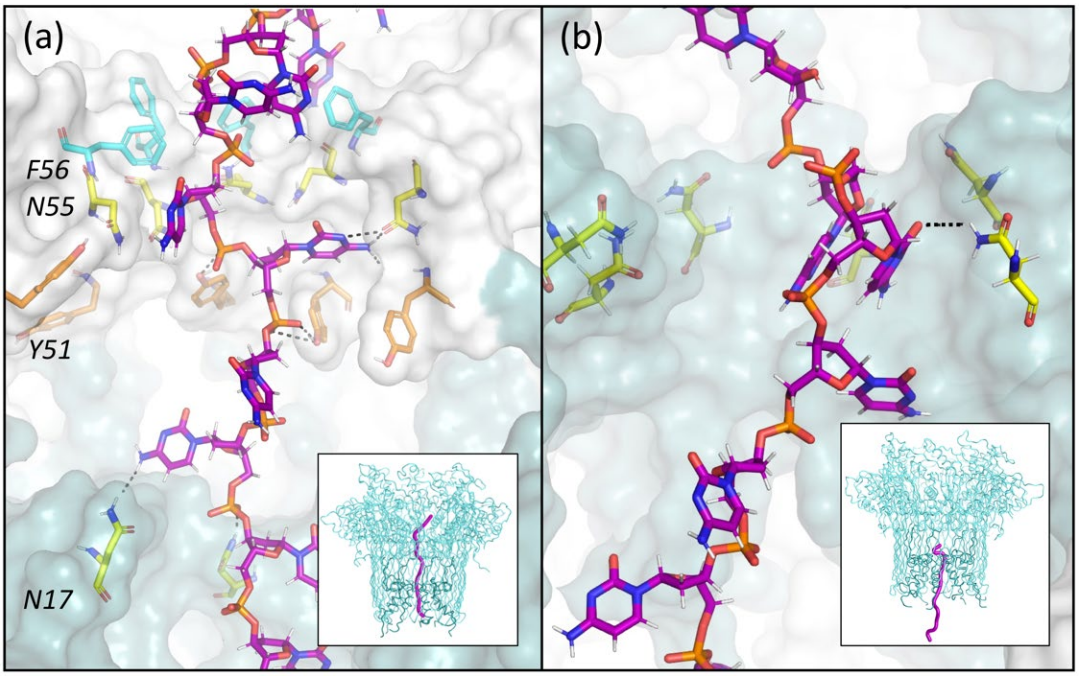

Figure 4.25: Representative polyC ssDNA conformations from two clusters of the CsgG-CsgF complex. DNA and the residues that interact with the nucleotides in the CsgG eyelet loop region and the CsgF constriction region are shown, with dashed lines marking hydrogen bonds. The inset shows the position of the DNA strand in the CsgG-CsgF complex.

To evaluate the interactions formed between DNA and the protein pores during translocation, the percentage of the simulation time that the DNA was within 0.4 nm of the residues was calculated (Figure 4.26). polyC formed interactions with Phe-56, Asn-55, and Tyr-51 residues in the CsgG eyelet loop region at a similar frequency as polyA in uncomplexed CsgG. In contrast, the frequency of interactions between polyC and the CsgG-CsgF complex is substantially lower than polyA, consistent with polyC exiting the pore earlier than polyA. Like polyA, polyC interacted more frequently with residues in the CsgG eyelet loop region than CsgF, which further indicates that the eyelet loop region has a greater impact on the DNA translocation rate.

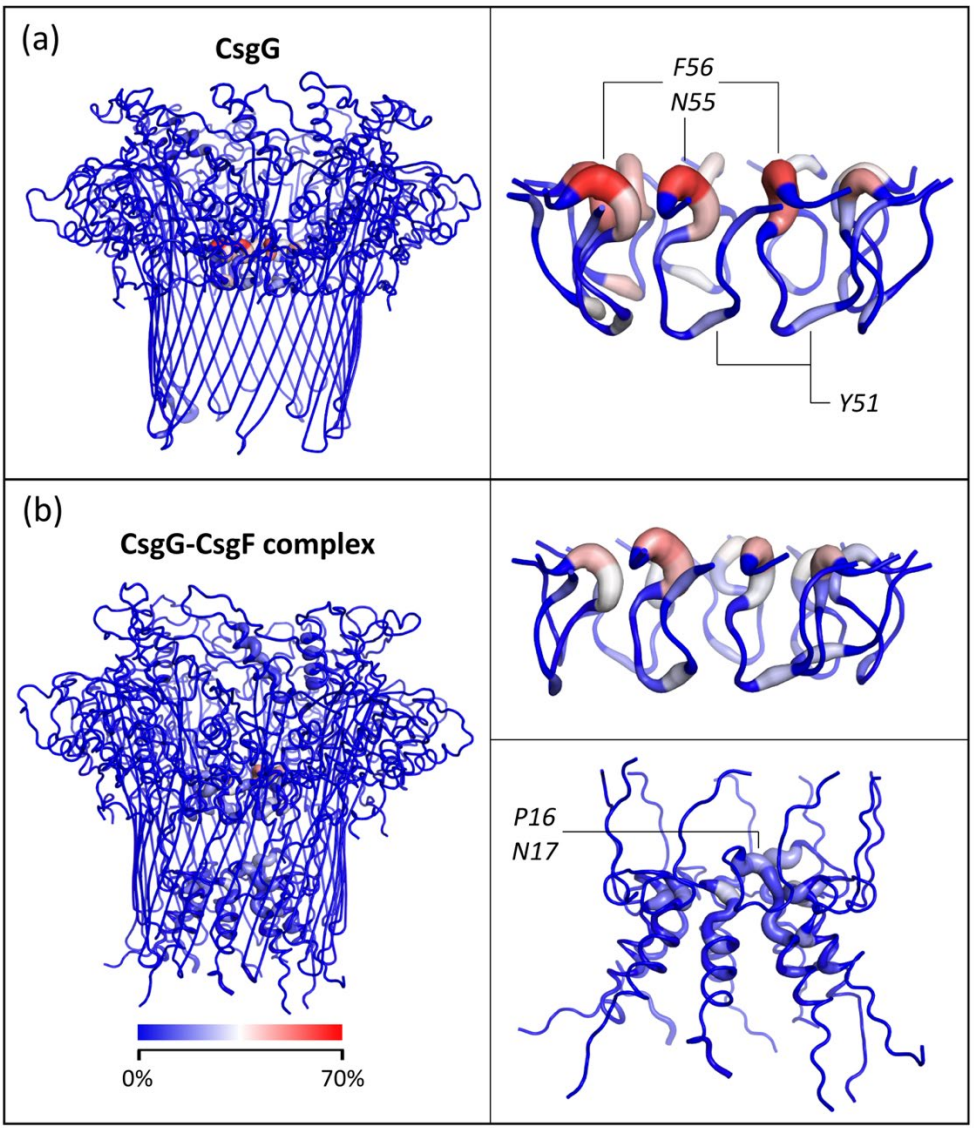

Figure 4.26: polyC ssDNA-protein interactions. CsgG (a) and the CsgG-CsgF complex (b) are coloured by the percentage of simulation time during which the residues interact with DNA in four independent simulations. An interaction is defined as an inter-atomic distance of < 0.4 nm. The eyelet loop region and CsgF are also shown.

4.3.4 Conductance of CsgG and the CsgG-CsgF complex in the presence of immobilised DNA

DNA sequencing relies on the characteristic reduction of the nanopore current caused by the translocating nucleotides. To characterise the influence of the additional constriction region formed by CsgF on the pore ionic current during DNA translocation, the DNA was threaded through and immobilised inside the protein pores and simulated under an applied electric field

equivalent to 0.9 V, which is five times higher than 0.18 V used for DNA sequencing [123]. The differences in water and ion dynamics in the pores were compared.

4.3.4.1 Water and ion dynamics in CsgG and the CsgG-CsgF complex

The mean bidirectional flux of water and the ionic currents through uncomplexed CsgG and the CsgG-CsgF complex in the presence of immobilised polyA ssDNA are reported in Table 4.6. Overall, the ionic current and the mean water flux were lower through the CsgG-CsgF complex compared to uncomplexed CsgG, as it contains an additional constriction region formed by CsgF. There was large variability in water flux through uncomplexed CsgG in three independent simulations (average: 993 ± 168 ns^{-1}) due to the motions of the eyelet loops in the CsgG constriction region. Up to three eyelet loops were observed to move upwards by 50 ns in one simulation, which resulted in the narrowing of the CsgG constriction region; hence, the water flux through the pore was significantly lower in this simulation (926 ns^{-1}) compared to the other two simulations (1034 ns^{-1} and 1020 ns^{-1}) (Figure 4.27). In the two simulations with higher water flux, the eyelet loops moved closer to the β-barrel wall, which widened the CsgG eyelet loop region. Conversely, as observed in simulations without DNA under an applied electric field, the eyelet loops did not move upwards in the CsgG-CsgF complex. The presence of CsgF resulted in reduced water flux and ionic current compared to uncomplexed CsgG.

Table 4.6. The mean bidirectional water flux and ionic currents through uncomplexed CsgG and the CsgG-CsgF complex with immobilised polyA ssDNA in 0.9 V.

System		**Mean water flux (ns^{-1})**	**I_{total} (pA)**	**I_K (pA)**	**I_{Cl} (pA)**
CsgG	Simulation 1	1034 ± 182	384 ± 74	378 ± 88	6.3 ± 21
	Simulation 2	926 ± 161	244 ± 172	222 ± 193	21 ± 27
	Simulation 3	1020 ± 159	371 ± 104	334 ± 92	37 ± 27
	Average ± SD	**993 ± 168**	**333 ± 124**	**311 ± 133**	**22 ± 25**
CsgG-CsgF complex	Simulation 1	145 ± 25	108 ± 88	102 ± 84	6.1 ± 19
	Simulation 2	134 ± 22	322 ± 68	306 ± 79	16 ± 35
	Simulation 3	122 ± 21	113 ± 73	110 ± 64	2.7 ± 17
	Average ± SD	**133 ± 23**	**181 ± 77**	**172 ± 76**	**8.3 ± 25**

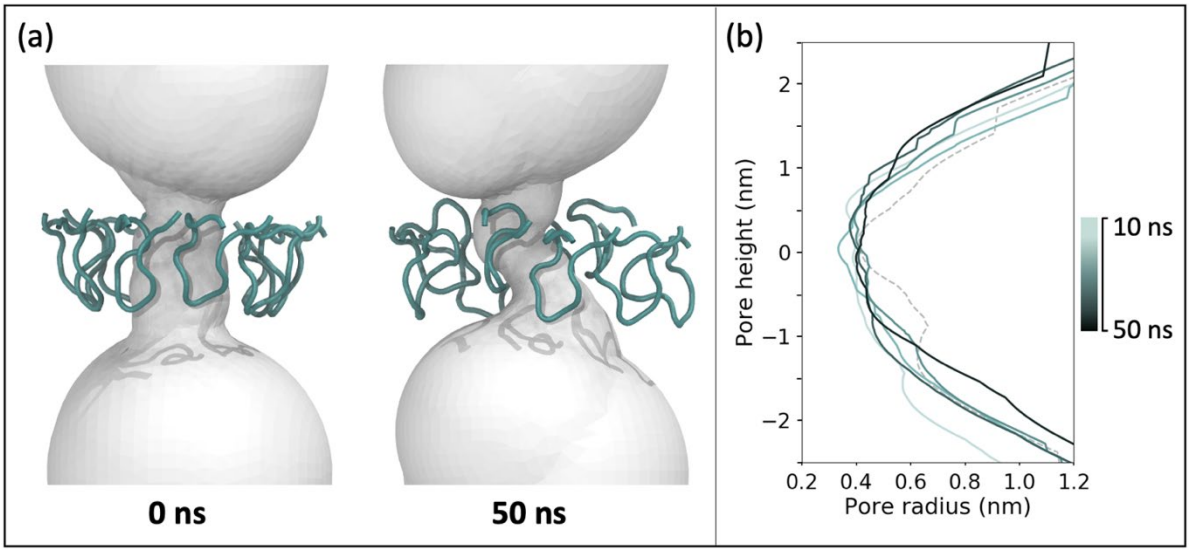

Figure 4.27: (a) The conformation of the eyelet loops and the shape of the pore formed by uncomplexed CsgG at 0 ns and 50 ns is shown from a simulation with immobilised polyA ssDNA, in which the lowest water flux was observed. (b) The pore radius is plotted over 50 ns simulation of uncomplexed CsgG with immobilised polyA ssDNA. The radius at 0 ns is plotted as a dashed line.

Next, the systems were simulated with positional restraints applied to CsgG eyelet loops (backbone atoms of residues 47-58) to prevent large loop motion. The mean bidirectional flux of water and the ionic currents through uncomplexed CsgG and the CsgG-CsgF complex for these simulations are reported in Table 4.7. The absence of large motions of the eyelet loops resulted in a lower and more consistent water flux amongst three independent simulations (average: 859 ± 140 ns^{-1}). Thus, a reduction in the motion of the eyelet loops may result in less noisy data generated during DNA sequencing. The water flux and the total ionic currents through the CsgG-CsgF complex are similar with and without positional restraints applied to the CsgG eyelet loops, as the eyelet loops did not move upwards in the presence of CsgF. The presence of CsgF resulted in a ~ 80% reduction in water flux through the CsgG-CsgF complex compared to uncomplexed CsgG.

4.3.4.2 Ionic current through the CsgG-CsgF complex in the presence of immobilised polyA and polyC ssDNA

To evaluate the ability of the CsgG-CsgF complex to discriminate between distinct nucleotides, the pore was simulated with immobilised polyC ssDNA. The mean bidirectional flux of water and the ionic currents through the CsgG-CsgF complex for these simulations are reported in Table 4.7. The average total ionic current through the CsgG-CsgF complex was ~ 1.7 x higher in the presence of

polyC than polyA. This can be explained by the larger steric bulk of the purine adenine physically occluding the pore to a greater extent than cytosine.

Table 4.7. The mean bidirectional water flux and ionic currents through uncomplexed CsgG and the CsgG-CsgF complex with immobilised ssDNA and CsgG eyelet loops, in 0.9 V.

System	DNA		Mean water flux (ns^{-1})	I_{total} (pA)	I_K (pA)	I_{Cl} (pA)
CsgG	polyA	Simulation 1	893 ± 142	389 ± 73	386 ± 94	3.0 ± 27
		Simulation 2	831 ± 142	205 ± 93	203 ± 114	1.9 ± 28
		Simulation 3	854 ± 137	218 ± 138	217 ± 143	0.8 ± 13
		Average ± SD	**859 ± 140**	**271 ± 105**	**269 ± 63**	**1.9 ± 23**
CsgG-CsgF complex	polyA	Simulation 1	156 ± 26	65 ± 63	64 ± 52	1.4 ± 17
		Simulation 2	140 ± 25	212 ± 61	211 ± 52	0.5 ± 45
		Simulation 3	122 ± 23	176 ± 70	172 ± 46	4.1 ± 30
		Average ± SD	**140 ± 25**	**151 ± 65**	**149 ± 50**	**2.0 ± 33**
	polyC	Simulation 1	156 ± 28	389 ± 61	389 ± 75	0.2 ± 19
		Simulation 2	155 ± 26	121 ± 61	120 ± 62	1.0 ± 25
		Simulation 3	143 ± 24	241 ± 44	241 ± 22	1.2 ± 34
		Average ± SD	**151 ± 26**	**251 ± 56**	**250 ± 58**	**0.7 ± 27**

It is challenging to draw direct comparisons between experimental studies due to differences in experimental conditions such as temperature, ion concentration, and the voltage applied. To circumvent this, blockage current ratios are calculated by dividing the ionic current in the presence of DNA by the open pore current. Therefore, a lower blockage current ratio corresponds to a greater degree of open pore current blocked by DNA. The blockage current ratios for the CsgG and the CsgG-CsgF complex are shown in Table 4.8, and the open pore currents are reported in Table 4.9. The total blockage current ratio in the presence of polyA is lower for the CsgG-CsgF complex (0.31) than for uncomplexed CsgG (0.46). Thus, the presence of CsgF resulted in a more distinct change in channel conductance in the presence of DNA.

Additionally, the blockage current ratio in the CsgG-CsgF complex is base size-dependent. The open pore current is impeded to a greater degree by larger purines in polyA compared to the smaller pyrimidines in polyC (the total blockage current ratio is 0.31 and 0.52, respectively). Interestingly, in wild-type α-hemolysin the current blockage is not solely base size-dependent; polyA blocked current to a lesser degree than polyC when translocating through the pore (blockage current ratio polyA = 0.15, polyC = 0.05, in 0.12 V) [51], and when immobilised inside the pore (blockage current ratio polyA = 0.18, polyC = 0.16, in 0.12 V) [198]. However, increasing the hydrophobicity of the α-hemolysin constriction region *via* mutagenesis (mutants E111N/K147N/M113X, X = V, L, or I) resulted in improved size-dependent base discrimination, i.e., polyA caused a greater current blockage than the smaller pyrimidines in polyC [199]. Furthermore, a hydrophobic solid-state nanopore was also shown to discriminate between polyA and polyC [200]. Similarly, in the simulations performed, the hydrophobic pore formed by the CsgG-CsgF complex (Figure 4.20) can discriminate between polyA and polyC on a base size basis.

Table 4.8. Blockage current ratios for uncomplexed CsgG and the CsgG-CsgF complex with immobilised ssDNA and CsgG eyelet loops, in 0.9 V. Ratios are calculated for average ionic currents, calculated from three independent simulations.

System	**DNA**	I_{total}	I_K	I_{Cl}
CsgG	polyA	0.46	0.64	0.01
CsgG-CsgF complex	polyA	0.31	0.48	0.01
	polyC	0.52	0.80	0.00

Table 4.9. The ionic currents through uncomplexed CsgG and the CsgG-CsgF complex with immobilised CsgG eyelet loops in 0.9 V, calculated for 50 ns.

System	I_{total} **(pA)**	I_K **(pA)**	I_{Cl} **(pA)**
CsgG	582 ± 70	423 ± 56	160 ± 26
CsgG-CsgF complex	480 ± 93	313 ± 62	167 ± 42

4.3.4.3 Ionic density through CsgG and the CsgG-CsgF complex

The ionic density maps were plotted for systems of uncomplexed CsgG and the CsgG-CsgF complex with and without immobilised DNA to identify regions where ions interacted strongly with the protein and the DNA in 0.9 V (Figure 4.28). Overall, the density of potassium (K^+) and chloride (Cl^-) ions is higher in uncomplexed CsgG than in the CsgG-CsgF complex. The density of Cl^- ions is lower than K^+ ions in all systems, which is concurrent with the lower Cl^- current observed in the simulations (Tables 4.7 and 4.9). In the presence of DNA, the density of K^+ ions is higher inside the pores compared to systems without DNA. This is likely due to the positively charged K^+ ions being attracted to DNA due to its negative charge. Concurrently, the repulsion between DNA and the negatively charged Cl^- ions results in a lower density of Cl^- ions in the presence of DNA compared to systems without DNA.

The positively charged K^+ ions accumulate near the CsgG β-barrel in uncomplexed CsgG. This is due to the presence of acidic sidechains of Asp-149, Asp-155, Glu-185, Glu-203, and Glu-201 residues lining the CsgG β-barrel. However, these residues form interactions with CsgF monomers in the CsgG-CsgF complex (Figure 4.11) and are replaced by hydrophobic residues (Figure 4.20) that have a low propensity to accumulate ions. Hence the ionic density is lower in the transmembrane region of the CsgG-CsgF complex compared to uncomplexed CsgG. Additionally, more K^+ ions accumulated in the eyelet loop region in uncomplexed CsgG than in the CsgG-CsgF complex. The higher density of K^+ ions in uncomplexed CsgG is consistent with the higher K^+ current observed through the pore compared to the CsgG-CsgF complex. The Cl^- ionic density is lower in the constriction regions of the CsgG-CsgF complex than in uncomplexed CsgG; however, the differences in Cl^- ionic densities between both pores are overall negligible, as was the case for Cl^- current measured. In uncomplexed CsgG, the Cl^- ions did not accumulate near the β-barrel surface like the K^+ ions because there is only one basic sidechain of residue Arg-142 lining the β-barrel.

The presence of polyC ssDNA inside the CsgG-CsgF complex resulted in a higher density of K^+ ions in the two constriction regions compared to the presence of polyA ssDNA. This is consistent with the ~ 1.7 times higher K^+ current observed through the pore in the presence of polyC compared to polyA. Overall, these density maps highlight that the presence of CsgF in the CsgG-CsgF complex affects the density of K^+ ions inside the pore, which is in line with the differences in ionic current observed through both pores mainly due to the cationic current.

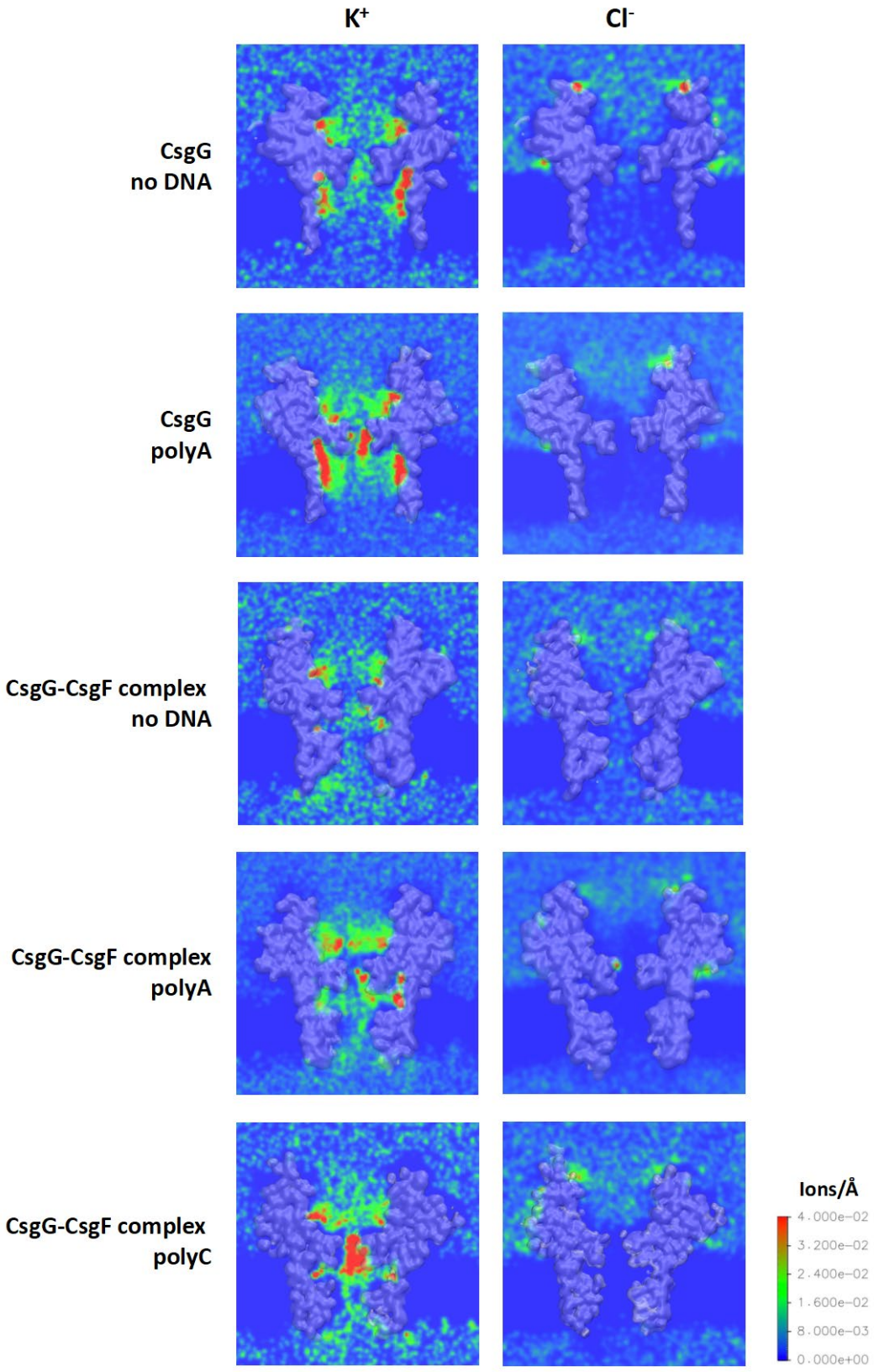

Figure 4.28: Ionic density maps of potassium and chloride ions (K^+ and Cl^-) for uncomplexed CsgG and the CsgG-CsgF complex, in 0.9 V.

4.4 Conclusions

In conclusion, the conformational behaviour of CsgG when uncomplexed and in the CsgG-CsgF complex was found to differ under an applied electric field. The eyelet loops forming the CsgG constriction region are more flexible in uncomplexed CsgG than in the CsgG-CsgF complex under an applied electric field. In uncomplexed CsgG, the eyelet loops were observed to 'flip' upwards into the vestibule region to varying degrees. Although this does not impact the protein's conformation or stability, it perturbs the geometry of the constriction region. The presence of CsgF stabilises the eyelet loops' conformation under an applied electric field in the CsgG-CsgF complex. In higher electric field strengths, whilst uncomplexed CsgG is unstable due to disruption of inter-monomer interactions, CsgF forms an extensive network of hydrogen bonds and electrostatic interactions with the CsgG β-barrel which results in the CsgG-CsgF complex remaining stable under the conditions.

The translocation of DNA with polyA and polyC sequence through uncomplexed CsgG and the CsgG-CsgF complex is slowed down due to the strand interacting with residues in the CsgG eyelet loop region and CsgF constriction region. DNA formed hydrogen bonds with Asn-55 and pi-stacking interactions with Phe-56 and Tyr-51 residues in the CsgG eyelet loop region, and transient hydrogen bonds with Asn-17 residues in the CsgF constriction, which resulted in the progressive movement of the strand through the pores. However, DNA interacted with residues in the eyelet loop region more frequently than in the CsgF constriction; hence DNA translocation is faster through the CsgF region after the strand exits the CsgG eyelet loop region. Therefore, the DNA translocation rate is influenced by the CsgG eyelet loop region, with the CsgF constriction playing a minor role. Additionally, a higher pulling force is occasionally required to pull polyA through the protein pores compared to polyC, which indicates that polyA experienced greater barriers during translocation than polyC. This is consistent with polyA interacting with the residues of the CsgG-CsgF complex more frequently than polyC.

The simulations of systems with the immobilised DNA threaded through uncomplexed CsgG and the CsgG-CsgF complex revealed that: (1) the change in the channel conductance in the presence of DNA is more distinct through the CsgG-CsgF complex compared to uncomplexed CsgG, and (2) the CsgG-CsgF complex channel conductance is sensitive to the size of the bases A and C.

In summary, the CsgG-CsgF complex forms a more hydrophobic and stable nanopore compared to uncomplexed CsgG. Although the second constriction formed by CsgF has a minor effect on the DNA translocation rate, it provides several notable advantages; firstly, DNA is retained in a more linear conformation during translocation compared to uncomplexed CsgG due to the dual-constriction and hydrophobic pore formed by the CsgG-CsgF complex, like the nanopores studied

in chapter 3 that also maintained the extended conformation of DNA during translocation. Secondly, the presence of CsgF results in a profound reduction in eyelet loop mobility, which has a concomitant decrease in ion flux with and without DNA. Lastly, the CsgG-CsgF complex channel conductance is sensitive to the size of the bases adenine (A) and cytosine (C). This work offers new structural and dynamic insights that will inform the future design of novel biology nanopores with improved performance.

Chapter 5 Characterisation of long ssDNA translocation through the *E. coli* proteins CsgG and CsgF

5.1 Introduction

In chapter 4, CsgG when uncomplexed and the CsgG-CsgF complex were characterised for nanopore DNA sequencing. The translocation of short polyA and polyC ssDNA were simulated by pulling the strands through the protein pores in steered MD simulations, which revealed that the DNA translocation rate is primarily influenced by the CsgG eyelet loop region. Simulations of DNA immobilised inside uncomplexed CsgG and the CsgG-CsgF complex under an applied electric field (as done previously for other proteins [4, 52, 151, 198, 199, 201]) revealed that the eyelet loop region plays a key role in modulating the conductance of the pores. The eyelet loops were found to be more mobile in uncomplexed CsgG compared to the CsgG-CsgF complex, which resulted in large variations in the ionic current amongst independent simulations. Lastly, the ionic current through the CsgG-CsgF complex was found to be more sensitive to the presence of ssDNA compared to uncomplexed CsgG and is sensitive to the size of adenine (A) and cytosine (C) bases.

Although the steered MD simulations in chapter 4 provided insights into DNA translocation through CsgG and the CsgG-CsgF complex, the simulations were not representative of nanopore sequencing experiments. DNA translocation during sequencing is facilitated by an electric field applied across the pore [1]. The electric field is known to influence the conformational dynamics of the protein [202], which can impact DNA translocation through the pore [203]. Additionally, the DNA fragments used are typically ultra-long that can exceed a megabase (1,000,000 bases) in length [42].

In this study, the translocation of long ssDNA through uncomplexed CsgG and the CsgG-CsgF complex is characterised under an applied electric field. As done in chapter 3, continuous ssDNA bound to itself across the periodic boundaries was used to mimic long DNA in nanopore sequencing experiments, and to eliminate artefacts that arise during the translocation of short strands, such as DNA coiling in the pore exit. Additionally, ssDNA with polyA and polyC sequences were used to investigate the rate of translocation, conformational dynamics, and the interactions that take place between the DNA and protein that influence its translocation through the pore. Multiple long (200 ns) simulations were performed to characterise the differences in the conductance of uncomplexed CsgG and CsgG-CsgF complex that arise due to polyA and polyC sequences.

5.2 Methods

5.2.1 Preparation of protein structures

The protein structures used in this study are CsgG (PDB 4UV3, 3.59 Å) and the CsgG-CsgF complex (PDB 6SI7, 3.4 Å). The missing loops in CsgG (residues 144, 193-199) were built using Coot [189] by fitting the structure into the map density obtained at a higher resolution (PDB 4Q79, 3.1 Å). For CsgG-CsgF, the missing residues (1-9, 103-110) were built using Modeller 9.02 *via* sequence alignment and fitting with CsgG. The N termini cysteine residues in CsgG were lipidated in all structures using CHARMM-GUI membrane builder [167, 190-192].

5.2.2 Generation of ssDNA

The models of ssDNA with polyA and polyC sequences were generated using the 3DNA package [157]. These were used to build continuous tensioned ssDNA, with no termini and bound to itself across periodic boundaries, as described in chapter 3 (section 3.2.2). To reduce the tension in the strand so that it is similar to experiment [123], ssDNA was simulated in NPT ensemble for 5 ns to reduce the length of the periodic image in the Z dimension.

5.2.3 Simulation protocol and analyses

The proteins were inserted in a phosphatidylcholine (POPC) membrane (CsgG: 1026 lipids; CsgG-CsgF: 1106) using CHARMM-GUI membrane builder [167, 190-192]. ssDNA was positioned inside CsgG and CsgG-CsgF complex, such that the strand was at least 2 Å away from the protein residues to prevent steric clashes. The resultant systems were solvated, with ions added to a concentration of 0.35 M alongside additional ions for neutralising the systems. The equilibration simulations were performed for ~ 65 ns total, with the system temperature maintained at 310 K. The equilibration of the systems involved running multiple simulations so that the strength of positional restraints applied to the protein and the POPC lipids could be reduced in a stepwise manner. Positional restraints were applied to the protein backbone and POPC headgroups with a force constant of 1000 kJ mol^{-1} nm^2, which was reduced to 500 kJ mol^{-1} nm^2. In simulations with an electric field applied, the positional restraints were applied to POPC headgroups with a force constant of 500 kJ mol^{-1} nm^2.

All simulations were performed using GROMACS 2021.3 [130, 159] and the CHARMM36m force field [196], and the TIP3P water model was used for solvating the systems [163]. Systems were simulated in NVT ensemble to retain the Z dimension of the periodic box at a constant value. This was to ensure that the ssDNA bound to itself across periodic images does not coil any further

during the simulations. The temperature was maintained using the velocity-rescale thermostat and a coupling constant of 0.1 ps. The lengths of all bonds were constrained using the LINCS algorithm enabling a timestep of 2 fs [134]. The Particle Mesh Ewald (PME) method was used to treat long-range electrostatic interactions with a short range cut-off of 1.4 nm [140]. The van der Waals interactions were curtailed at 1.4 nm, with long-range dispersion corrections applied to the pressure and energy. A constant voltage drop across the simulation cell in the Z dimension imposed an electric field. The periodic boundary conditions were applied to all systems in three dimensions, as done in previous studies [81, 88, 89, 162]. Replicate simulations were initiated using coordinates extracted at random time points from the last 100 ps of the equilibration simulation. The initial coordinates and velocities differ for each replicate simulation. The systems contain ~ 912,0000 atoms and were simulated in an applied electric field with a performance of ~ 24 ns/day.

Analyses were performed using GROMACS utilities and locally written code. Clustering analysis was performed using the linkage method implemented in GROMACS. Trajectories from independent simulations were concatenated before clustering DNA conformations with a Root mean square deviation cut-off of 0.15 nm, using the initial conformation of DNA inside the pore as a reference. Pore radius profiles of the proteins were calculated as an average across the simulations for a given pore using HOLE [164]. The molecular graphics images were generated using the Visual Molecular Dynamics (VMD) package [165] and PyMOL [156]. The ionic current was calculated as described previously [80], [122].

5.3 Results and Discussion

The translocation of continuous ssDNA through uncomplexed CsgG and the CsgG-CsgF complex was simulated by applying an electric field equivalent to 0.57 V across the membrane, as systems containing uncomplexed CsgG did not remain stable under higher electric field strengths. A summary of the simulation systems of the uncomplexed CsgG and the CsgG-CsgF complex in this chapter are presented in Table 5.1.

Table 5.1. Summary of the simulations discussed in this chapter.

System	DNA	Simulations (0.57 V)
CsgG	polyA	3 x 200 ns
	polyC	3 x 200 ns
CsgG-CsgF complex	no	3 x 100 ns
	polyA	3 x 200 ns
	polyC	3 x 200 ns

5.3.1 Translocation of polyA ssDNA

The translocation of polyA ssDNA through uncomplexed CsgG and the CsgG-CsgF complex was investigated to study the rate of translocation and the conformations adopted by the DNA in the pore. Three independent simulations were run for each protein pore. To characterise DNA translocation, the position of the centre of mass of the DNA nucleotides was calculated as a function of time (Figure 5.1). DNA translocation was faster through uncomplexed CsgG, with at most four nucleotides exiting the eyelet loop region by 100 ns in two simulations, compared to the CsgG-CsgF complex, in which two nucleotides exited the eyelet loop region by 100 ns in one simulation. Overall, nucleotides were more mobile in the vestibule region compared to the rest of the strand, as indicated by greater fluctuations in their Z coordinates over time. This is expected as the vestibule is the widest region of the protein pores and so the DNA is not confined in a narrow geometry.

The CsgG eyelet loop region was observed to greatly impact the translocation of DNA through both protein pores. In two simulations of uncomplexed CsgG, DNA translocation proceeded once the nucleotides halted in the eyelet loop region were released and rapidly moved downwards. DNA did not translocate at all during ~ 20-200 ns in one simulation; three nucleotides were halted in and near the eyelet loop region during this time. Although DNA remained halted inside the pore, nucleotides in the vestibule and the β-barrel were conformationally more labile (as indicated by greater fluctuations in their Z coordinate) compared to those in and near the eyelet loop region, which indicates that DNA translocation was largely controlled by the eyelet loop region.

Similarly, in the CsgG-CsgF complex, nucleotides were halted inside and above the eyelet loop region for as long as 100 ns before proceeding translocation through the pore. The CsgF region posed a second barrier to DNA translocation; in one simulation, two nucleotides exited the eyelet

loop region by ~ 90 ns, however the strand did not move downwards following this due to nucleotides being halted above the CsgF constriction region. Nucleotides were halted in the CsgF constriction region for similar timescales as in the CsgG eyelet loop region.

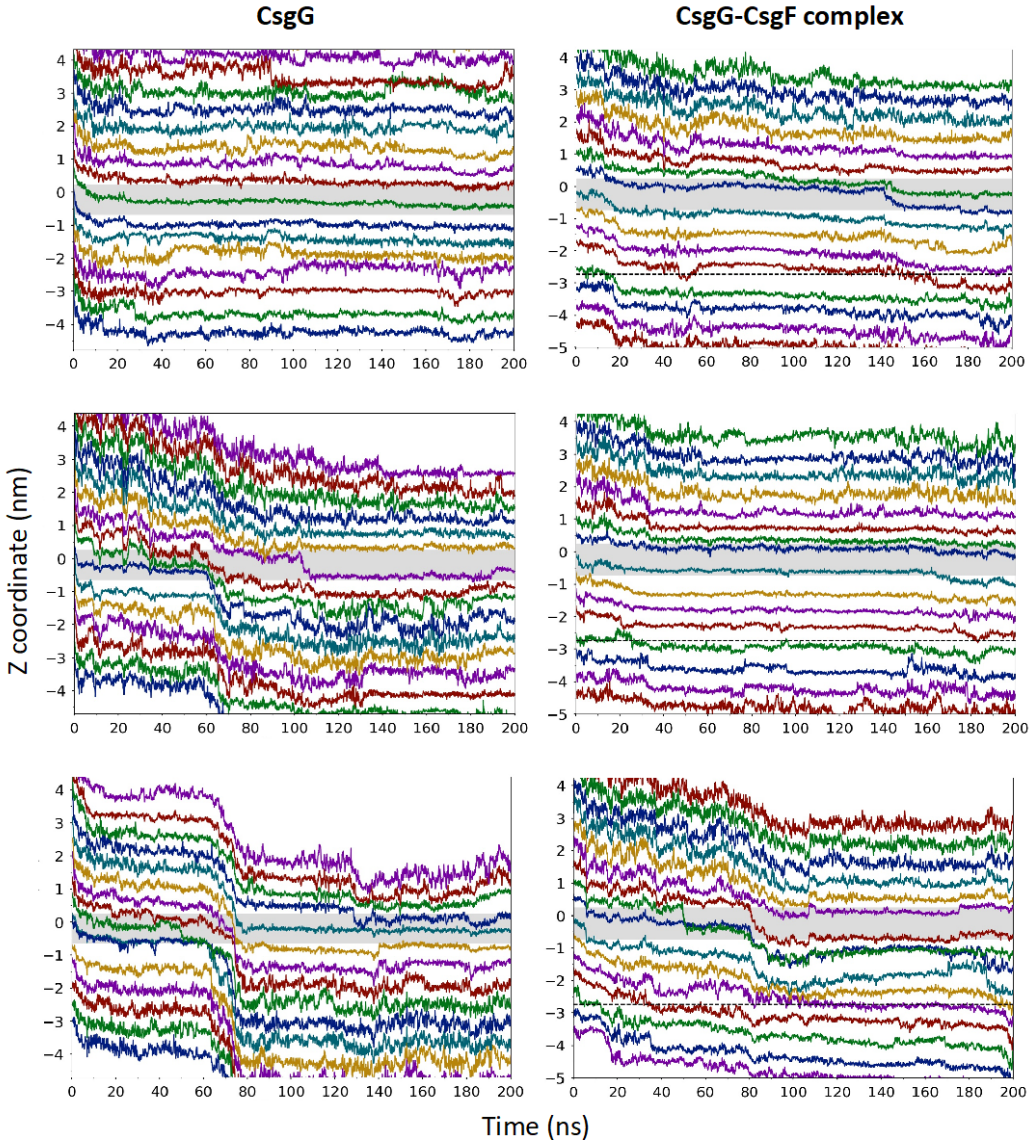

Figure 5.1: The translocation of polyA ssDNA through uncomplexed CsgG and the CsgG-CsgF complex is measured as the Z coordinate of the centre of mass of nucleotides over time in three independent simulations. The eyelet loop region is shaded in grey, and a dashed line marks the CsgF constriction.

The conformations of the DNA strand and the DNA-protein interactions were next examined to elucidate the origins of the differences in translocation through uncomplexed CsgG and the CsgG-CsgF complex. In simulations of uncomplexed CsgG, DNA remained in close proximity to the pore lining by 200 ns (Figure 5.2a). The DNA-protein interactions by 200 ns in simulation 1, in which

DNA translocated the least, are shown in Figure 5.2b. The phosphate groups of the nucleotide formed hydrogen bonds with charged residues (Lys-94 and Arg-97) in the vestibule region. In the eyelet loop region, the nucleotide backbone atoms formed hydrogen bonds with Ser-54 residues, whilst the rest of the nucleotides interacted with backbone atoms of Asn-55 residues. Nucleotides in the β-barrel region also formed hydrogen bonds with residues near the pore exit. Together, this resulted in DNA remaining associated with the pore for 200 ns in this simulation.

The DNA was retained in a more extended conformation in the CsgG-CsgF complex compared to uncomplexed CsgG (Figure 5.3a), due to the strand being held inside the pore by two constriction regions. The DNA-protein interactions by 200 ns in simulation 2, in which DNA translocated the least, are shown in Figure 5.3b. In this simulation, nucleotides formed hydrogen bonds with Tyr-51 and Asn-55 in the eyelet loop region. Nucleotides in the CsgF region formed hydrogen bonds with the Pro-16 backbone and two asparagine residues (Asn-17 and Asn-24).

To identify key residues that interact with the DNA and influence its translocation, the DNA-protein interactions were quantified by calculating the percentage simulation time that the residues were within 0.4 nm of the DNA strand in three independent simulations (Figure 5.4 and Figure 5.5). This showed that DNA primarily interacted with residues in the CsgG eyelet loop region (Phe-56, Asn-55, and Tyr-51) and CsgF constriction region (Asn-17, Phe-20, and Asn-24). DNA interacted with residues in the eyelet loop region more frequently in the CsgG-CsgF complex compared to uncomplexed CsgG, which is concurrent with the slower DNA translocation observed through the CsgG-CsgF complex. Although less frequently, DNA also formed interactions with residues in the vestibule region of both pores, which included charged residues such as arginine and lysine, and polar residues such as asparagine. In uncomplexed CsgG, DNA formed interactions with residues in the β-barrel near the pore exit, which included charged residues and residues with aromatic sidechains (tyrosine and phenylalanine). The DNA-protein interactions in the vestibule region and the β-barrel were formed as the strand coiled and moved close to the pore lining when DNA translocation was halted by the constriction regions(s).

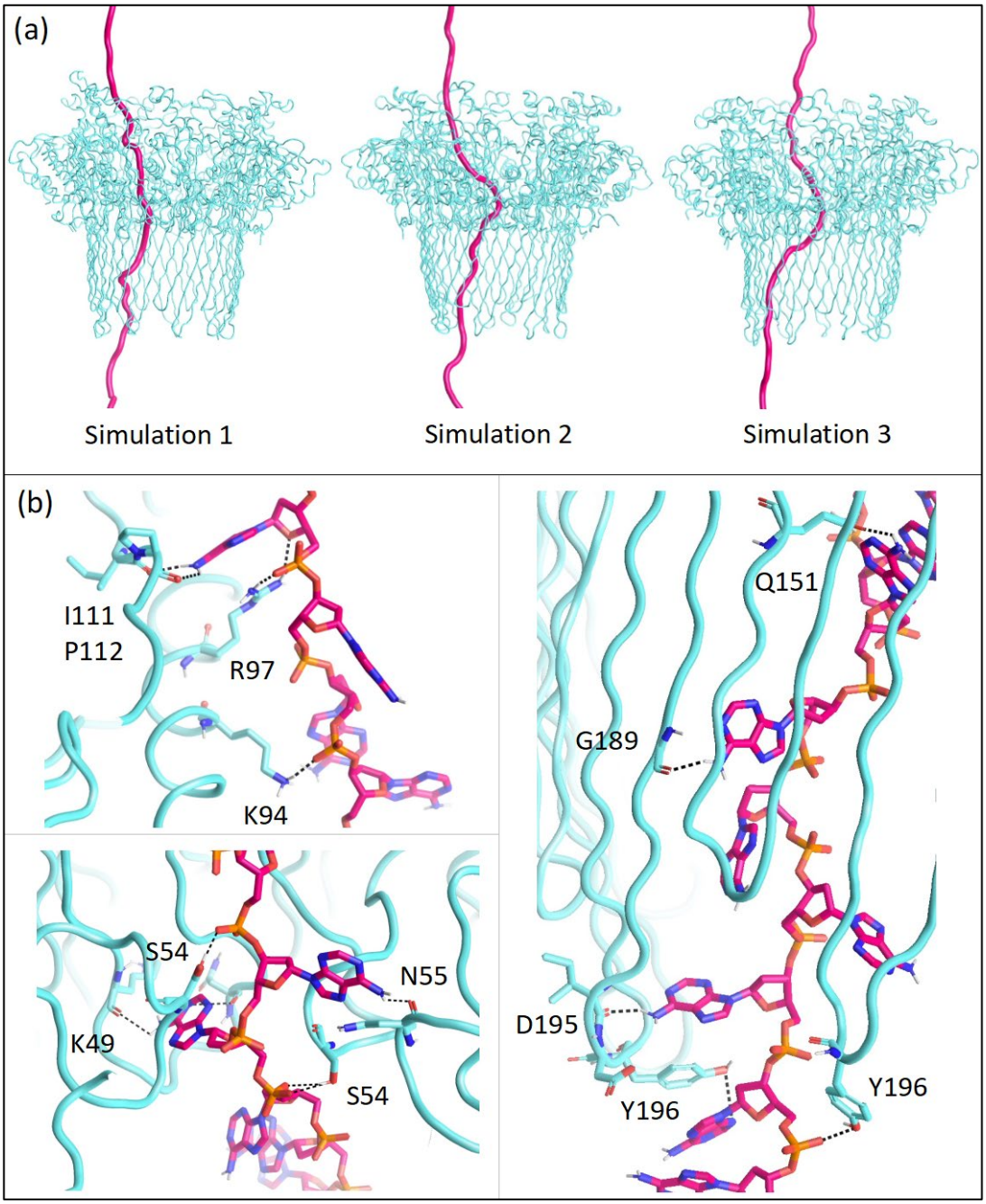

Figure 5.2: (a) The conformation of polyA ssDNA (pink) at 200 ns is shown for three independent simulations of uncomplexed CsgG in 0.57 V. (b) The interactions between polyA ssDNA and the residues in the eyelet loop region (left) and the β-barrel at 200 ns are shown. Hydrogen bonds are marked by dashed lines (< 0.32 nm).

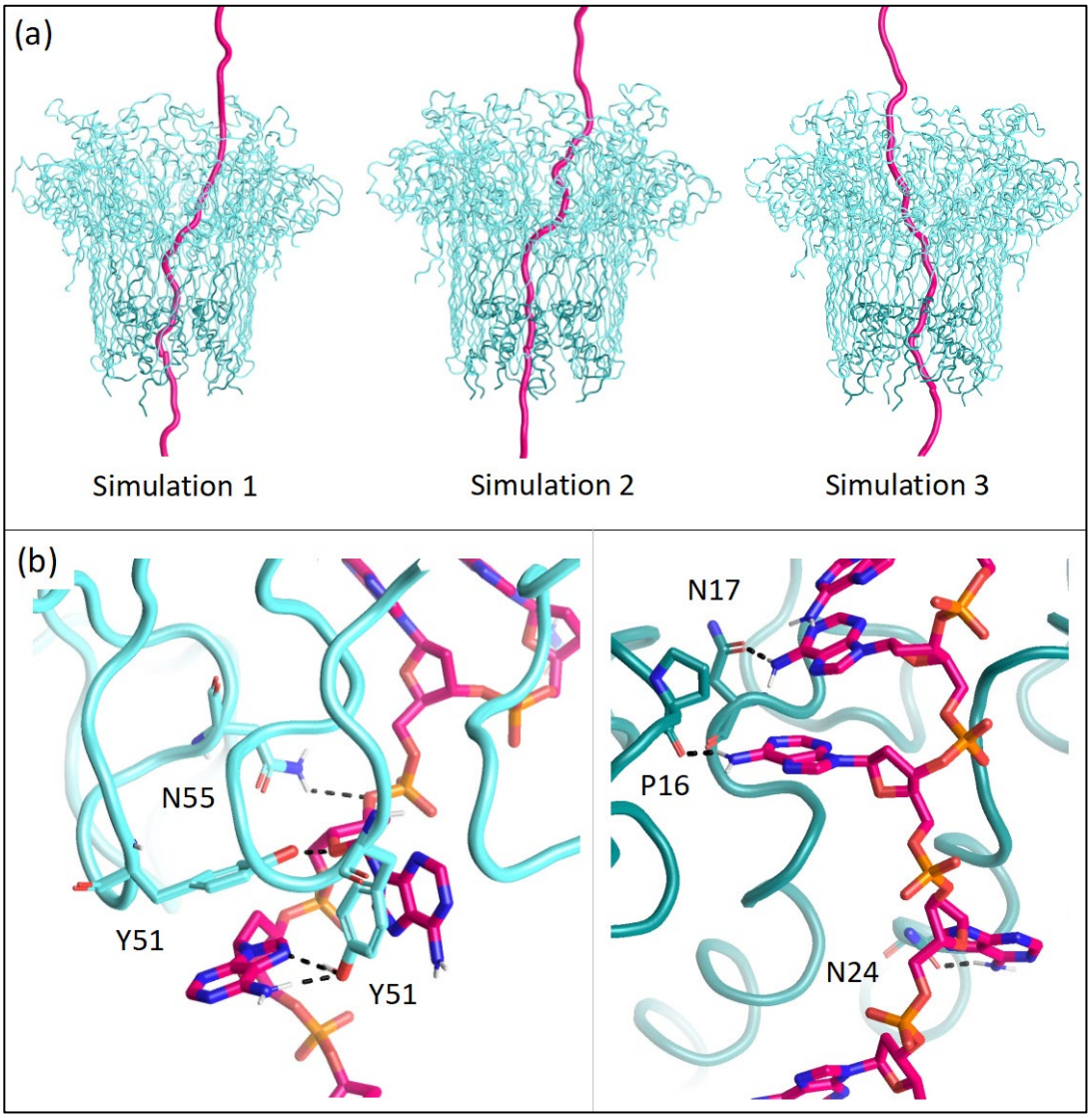

Figure 5.3: (a) The conformation of polyA ssDNA (pink) at 200 ns is shown for three independent simulations of the CsgG-CsgF complex in 0.57 V. (b) The interactions between polyA ssDNA and the residues in the eyelet loop region (left) and CsgF (right) at 200 ns are shown. Hydrogen bonds are marked by dashed lines (< 0.32 nm).

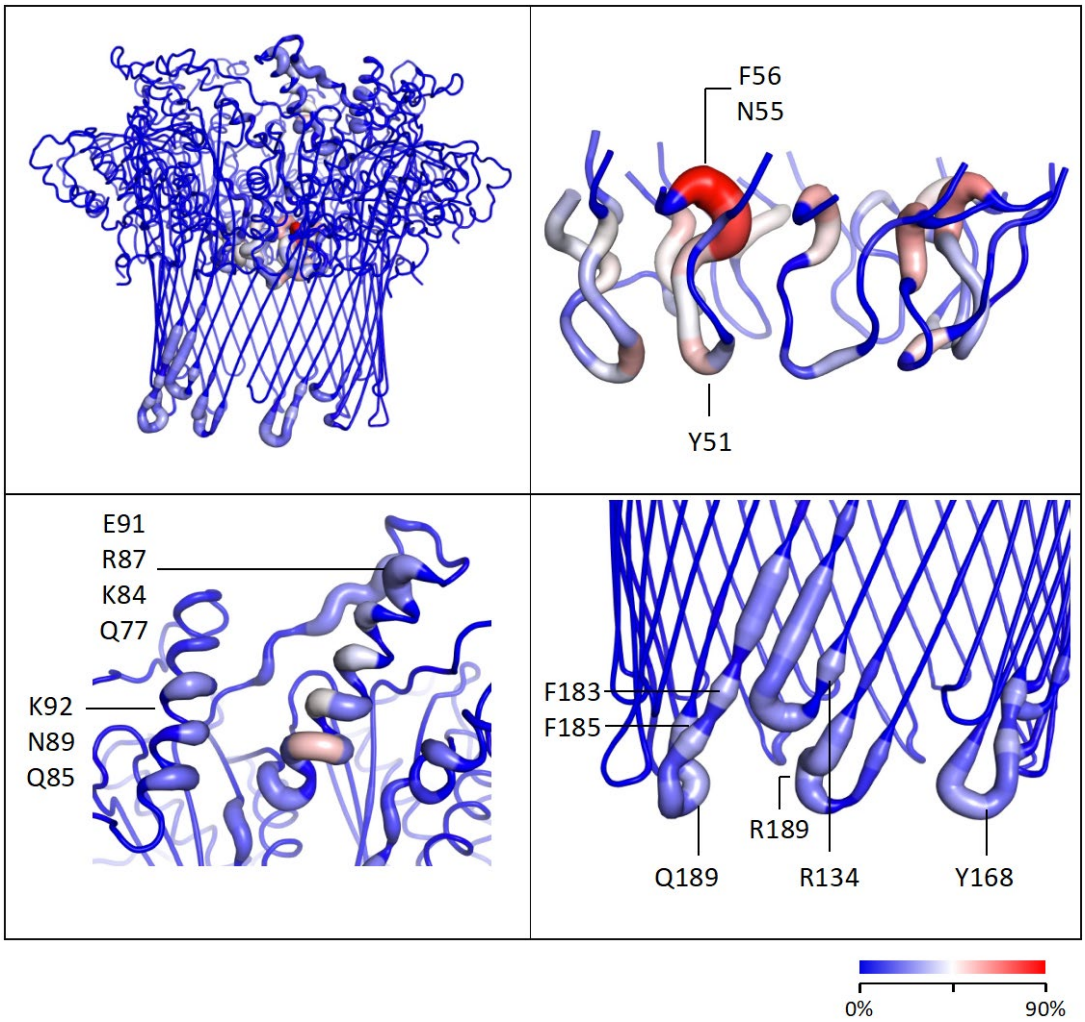

Figure 5.4: polyA ssDNA-protein interactions in simulations of uncomplexed CsgG are coloured by the percentage of simulation time for which the residues interact with the DNA nucleotides in three independent simulations. Interactions are defined as an inter-atomic distance of < 0.4 nm.

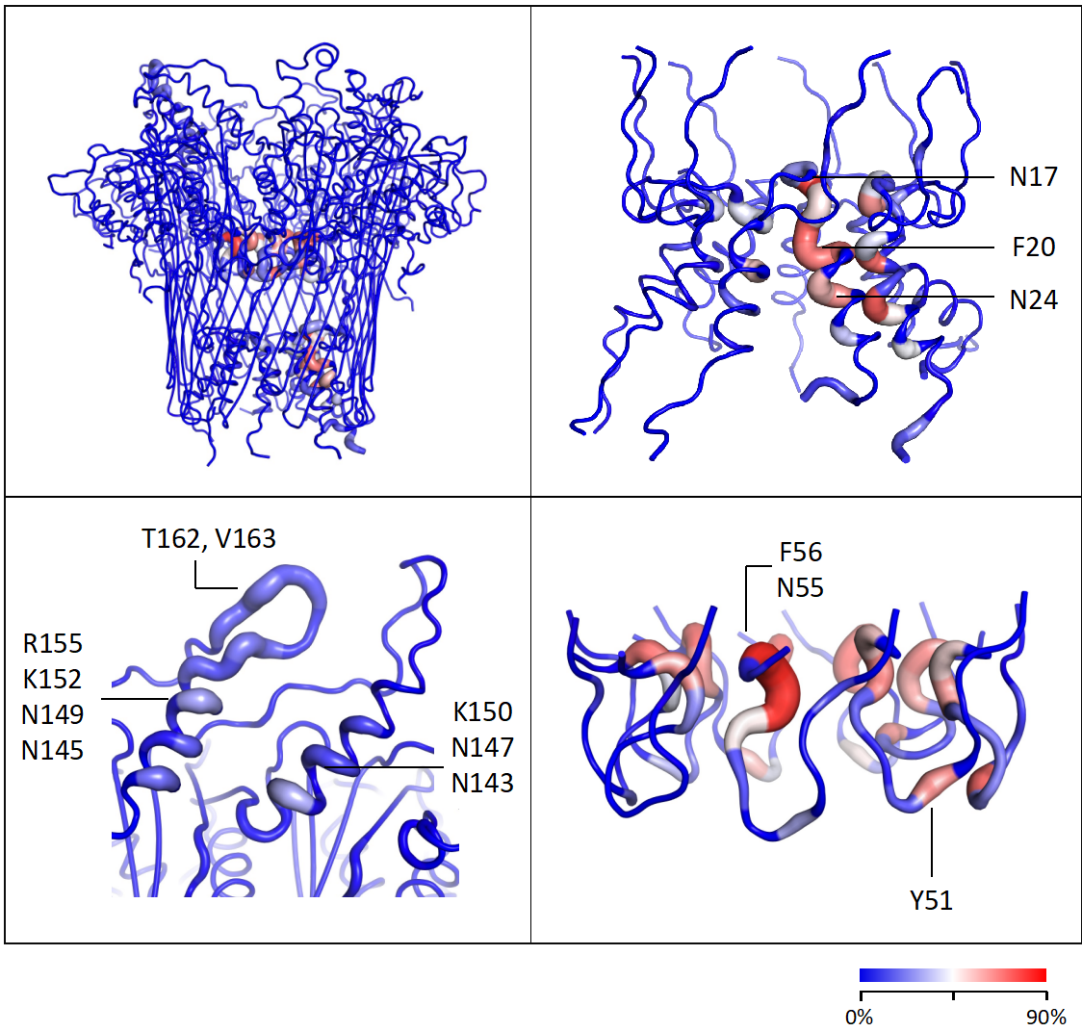

Figure 5.5: polyA ssDNA-protein interactions in simulations of the CsgG-CsgF complex are coloured by the percentage of simulation time for which the residues interact with the DNA nucleotides in three independent simulations. Interactions are defined as an inter-atomic distance of < 0.4 nm.

During DNA translocation, there was large variability in the motion of the eyelet loops forming the CsgG constriction region in both uncomplexed CsgG and the CsgG-CsgF complex. To assess the conformational drift of the eyelet loop region during the simulations, the root mean square deviation (RMSD) of the backbone Cα atoms (residues 47-58) from its initial conformation was calculated (Figure 5.6). The eyelet loop region exhibited a larger conformational drift in the CsgG-CsgF complex (RMSD ~ 0.20-0.30 nm) compared to uncomplexed CsgG (RMSD ~ 0.18-0.22 nm). The motility of eyelet loops resulted in the perturbation of the geometry of the constriction region during DNA translocation (Figure 5.6). This was more prominent in the CsgG-CsgF complex, in which an eyelet loop also flipped upwards into the vestibule region in the three simulations

(simulation 1 = ~ 55 ns, simulation 2 = ~ 70 ns, and simulation 3 = ~ 40 ns). However, DNA translocation was not markedly affected following this (Figure 5.1).

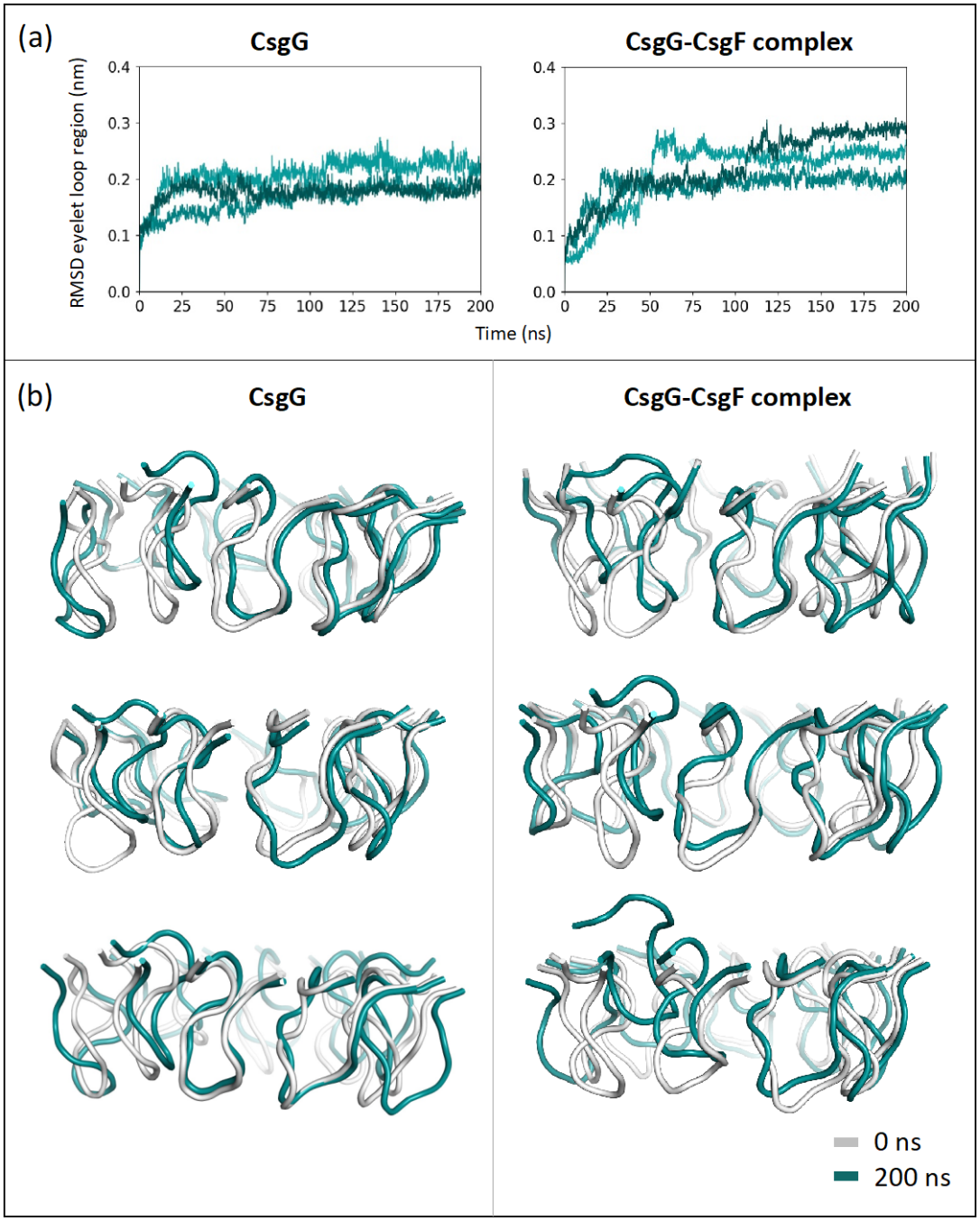

Figure 5.6: (a) RMSD of the eyelet loop region compared to its initial conformation at 0 ns (backbone Cα atoms) is plotted over time for three independent simulations of polyA ssDNA systems. (b) The conformation of the eyelet loop region at 0 ns and 200 ns is shown.

The ionic currents through the pores were measured for the simulations to ascertain the impact of eyelet loop mobility on the ionic current through the pore during DNA translocation. The current through uncomplexed CsgG was on average ~ 7.5 times higher than through the CsgG-CsgF complex (Table 5.2). This is expected, as uncomplexed CsgG lacks the additional CsgF constriction region and therefore is more open than the CsgG-CsgF complex.

The current through the pore during DNA translocation is expected to be similar amongst independent simulations, as the DNA consists of repeat units of the nucleotide which would block the pore to a similar degree. However, the ionic current during DNA translocation varied substantially amongst three independent simulations of each protein pore. The cumulative current over time is reported in Figure 5.7; a stationary current is indicated by a linear increase of the cumulative current with time. For uncomplexed CsgG, the cumulative current increased linearly in simulation 1, in which DNA translocated the least through the pore. The ionic current is the highest in simulation 1 due to the DNA remaining close to the pore lining during ~ 20-200 ns (Figure 5.5). However, in two simulations during which DNA translocation was observed, the current fluctuated over time, as indicated by the non-linear increase in the cumulative current. This was observed to a greater degree in the CsgG-CsgF complex, in which the slope of the cumulative current deviated substantially following the flipping of an eyelet loop (simulation 1 = ~ 60 ns, simulation 2 = ~ 70 ns, and simulation 3 = ~ 40 ns). These results indicate that the current through uncomplexed CsgG and the CsgG-CsgF complex is affected by the perturbation of the eyelet loops in the CsgG constriction during DNA translocation.

Table 5.2. The ionic current through uncomplexed CsgG and the CsgG-CsgF complex during polyA ssDNA translocation in 0.57 V, calculated for 200 ns in three independent simulations.

System	DNA		I_{total} (pA)	I_K (pA)	I_{Cl} (pA)
CsgG	polyA	Simulation 1	734 ± 44	556 ± 41	178 ± 49
		Simulation 2	511 ± 87	470 ± 63	41 ± 41
		Simulation 3	417 ± 81	384 ± 65	33 ± 18
		Average ± SD	**554 ± 73**	**470 ± 58**	**84 ± 39**
CsgG-CsgF complex	polyA	Simulation 1	70 ± 43	60 ± 36	9.8 ± 8.8
		Simulation 2	58 ± 29	56 ± 25	2.5 ± 7.8
		Simulation 3	78 ± 29	77 ± 22	0.6 ± 16
		Average ± SD	**69 ± 34**	**64 ± 28**	**4.3 ± 11**

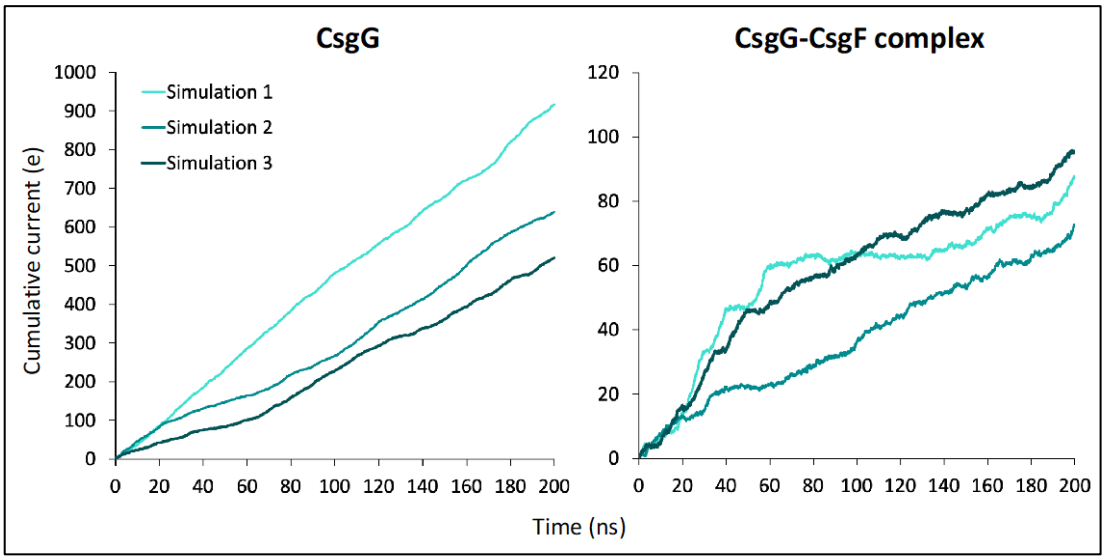

Figure 5.7: The cumulative current is plotted as a function of time in three independent simulations of polyA translocation through uncomplexed CsgG and the CsgG-CsgF complex in 0.57 V. A linear increase of the cumulative currents with time indicates stationary currents; a linear regression fit to these curves gives the average currents in Table 5.2. The cumulative currents are shown in the units of the unitary charge (e = 1.6 × 10^{-19} C).

To elucidate the origins of the variations in current, cluster analysis was performed to group similar conformations adopted by the DNA inside the constriction region(s) in three independent simulations, and the ionic current was measured for each cluster population. Following this, several measurements were conducted to quantify the properties of the DNA and the pore to elucidate the cause of the differences in currents measured for the cluster populations.

For uncomplexed CsgG, the ionic current for ten cluster populations ranged between ~ 245-941 pA (Table 5.3). To measure the degree of DNA coiling inside the pore, the DNA end-to-end distances were measured for a 9 nucleotide-long segment consisting of nucleotides in and near the CsgG eyelet loop region. A shorter end-to-end distance between the two termini of DNA corresponds to a higher degree of coiling [81, 88], which is expected to result in a lower current through the pore. However, this was not the case in these simulations; although the average DNA end-to-end distance inside the pore was the longest for cluster 11 (~ 5.50 nm), the current was the lowest for this cluster population (~ 245 pA). The average DNA end-to-end distance inside the pore was ~ 4.77 nm in three cluster populations (clusters 1, 7, 10), however the current through the pore ranged between ~ 443-941 pA in these clusters.

To investigate if the current through uncomplexed CsgG correlated to the degree of DNA coiling in the eyelet loop region, the DNA end-to-end distance was measured for a 4 nucleotide-long segment occupying the region for the cluster populations. Once again, this did not correlate with the pore current. Additionally, the average DNA end-to-end distance in the eyelet loop region did not greatly differ amongst cluster populations (~ 1.77-1.87 nm).

Although the DNA end-to-end distance is indicative of DNA coiling, it does not take into account the differences in the pore geometry that arise due to the perturbation of the eyelet loops in the CsgG constriction region. Hence, the solvent-accessible surface area (SASA) was calculated for the DNA segments; however, these also did not correlate with the pore current.

Table 5.3. The ionic current is calculated for conformations prominently adopted by polyA ssDNA during translocation through uncomplexed CsgG, obtained from cluster analysis. The DNA end-to-end distance and the SASA of DNA segments inside the pore, and the SASA of the CsgG eyelet loop region, are calculated as an average for each cluster population.

			DNA in CsgG eyelet loop region (4 nucleotides)		**DNA in the pore (9 nucleotides)**		
Cluster	**Duration (ns)**	I_{total} **(pA)**	DNA end-to-end distance (nm)	SASA (nm^2)	DNA end-to-end distance (nm)	SASA (nm^2)	**SASA CsgG eyelet loop region (nm^2)**
1	8.9	941 ± 166	1.85 ± 0.10	6.5 ± 0.45	4.77 ± 0.16	19.9 ± 0.88	24.8 ± 1.17
2	33.6	809 ± 144	1.77 ± 0.11	5.8 ± 0.62	5.28 ± 0.14	19.7 ± 1.25	27.5 ± 1.00
3	8.7	806 ± 147	1.82 ± 0.09	5.9 ± 0.46	5.34 ± 0.13	20.8 ± 1.10	28.2 ± 0.87
4	28.7	788 ± 255	1.83 ± 0.09	6.0 ± 0.56	4.76 ± 0.15	20.0 ± 0.79	24.7 ± 0.88
5	22.1	777 ± 171	1.80 ± 0.10	5.9 ± 0.47	5.20 ± 0.17	21.1 ± 0.92	27.0 ± 0.98
6	101.1	703 ± 63	1.77 ± 0.08	5.0 ± 0.54	5.10 ± 0.14	18.9 ± 1.37	27.9 ± 0.96
7	24.2	625 ± 218	1.83 ± 0.10	5.0 ± 0.36	4.77 ± 0.14	18.1 ± 0.95	27.5 ± 0.89
8	12.3	620 ± 174	1.82 ± 0.10	5.6 ± 0.34	4.70 ± 0.16	18.4 ± 0.73	26.3 ± 0.87
9	13.8	520 ± 143	1.80 ± 0.15	5.8 ± 0.40	4.64 ± 0.20	18.6 ± 0.84	26.8 ± 1.50
10	10.2	443 ± 170	1.84 ± 0.09	7.2 ± 0.43	4.77 ± 0.13	19.9 ± 0.74	24.7 ± 0.74
11	31.0	245 ± 127	1.87 ± 0.07	5.2 ± 0.87	5.50 ± 0.11	20.2 ± 1.17	27.5 ± 1.13

In simulations of the CsgG-CsgF complex, the ionic current for ten cluster populations ranged between ~ 7.6-292 pA (Table 5.4). As with the uncomplexed CsgG, the DNA end-to-end distances did not correlate with the ionic currents measured for the cluster populations. For example, although the DNA end-to-end distances in the eyelet loop region were ~ 1.72 nm in two clusters, the ionic current was substantially higher in cluster 1 (~ 292 pA) compared to cluster 10 (~ 7.6 pA).

The same was the case for the end-to-end distances of DNA segments threaded through both CsgG and CsgF constriction regions (cluster 8 total ionic current = ~ 45 pA and DNA end-to-end distance = ~ 4.22 nm, cluster 9 total ionic current = ~ 41 pA and DNA end-to-end distance = ~ 3.49 nm). The SASA of nucleotide segments also did not correlate with the ionic current. In summary, the ionic current through uncomplexed CsgG and the CsgG-CsgF complex was not found to be characteristic of the conformations adopted by the polyA ssDNA during translocation.

The pore radius was calculated for each cluster population of the CsgG-CsgF complex systems to compare the pore width to the current measured through the pore. This revealed that although the constriction regions in the CsgG-CsgF complex were wider in cluster 10 compared to cluster 7, the current was substantially lower for cluster 10. Additionally, the current was similar in clusters 7 and 8 despite the constriction regions being narrower in cluster 7 (Figure 5.8). However, when the representative structures of these clusters were simulated after removing DNA, the current through the open pore correlated with the pore radius (Table 5.5). This indicates that the current through the CsgG-CsgF complex during polyA ssDNA translocation is influenced by a complex interplay between the dynamics of the DNA and the eyelet loop region.

Table 5.4. The ionic current is calculated for conformations prominently adopted by polyA ssDNA during translocation through the CsgG-CsgF complex, obtained from cluster analysis. The DNA end-to-end distance and the SASA of DNA segments inside the pore, and the SASA of the CsgG eyelet loop region, are calculated as an average for each cluster population.

			DNA in CsgG eyelet loop region (4 nucleotides)		**DNA in the pore (9 nucleotides)**		
Cluster	**Duration (ns)**	**I_{total} (pA)**	DNA end-to-end distance (nm)	SASA (nm^2)	DNA end-to-end distance (nm)	SASA (nm^2)	**SASA CsgG eyelet loop region (nm^2)**
1	14.5	292 ± 159	1.72 ± 0.10	6.1 ± 0.60	4.62 ± 0.14	16.2 ± 0.79	24.2 ± 0.84
2	8.3	121 ± 77	1.60 ± 0.15	6.4 ± 0.62	4.31 ± 0.13	14.2 ± 0.78	23.4 ± 1.07
3	11.5	108 ± 63	1.50 ± 0.09	8.5 ± 0.48	4.03 ± 0.09	16.2 ± 0.67	21.1 ± 0.98
4	7.9	87 ± 43	1.57 ± 0.06	7.9 ± 0.43	3.90 ± 0.08	15.7 ± 0.61	24.3 ± 0.85
5	7.2	73 ± 75	1.49 ± 0.10	7.2 ± 0.58	3.49 ± 0.14	14.7 ± 0.90	25.2 ± 1.36
6	18.0	71 ± 34	1.74 ± 0.04	5.3 ± 0.40	4.32 ± 0.11	14.5 ± 0.76	22.6 ± 0.78
7	25.2	45 ± 78	1.80 ± 0.06	6.2 ± 0.68	4.42 ± 0.12	14.8 ± 0.87	22.6 ± 0.98
8	143.9	44 ± 33	1.63 ± 0.08	4.9 ± 0.53	4.26 ± 0.10	12.8 ± 1.04	23.1 ± 1.10
9	55.1	41 ± 44	1.39 ± 0.10	7.9 ± 0.54	3.89 ± 0.10	14.7 ± 0.80	22.5 ± 1.63
10	66.7	7.6 ± 97	1.72 ± 0.10	5.1 ± 0.48	4.40 ± 0.16	14.1 ± 0.78	21.4 ± 1.08

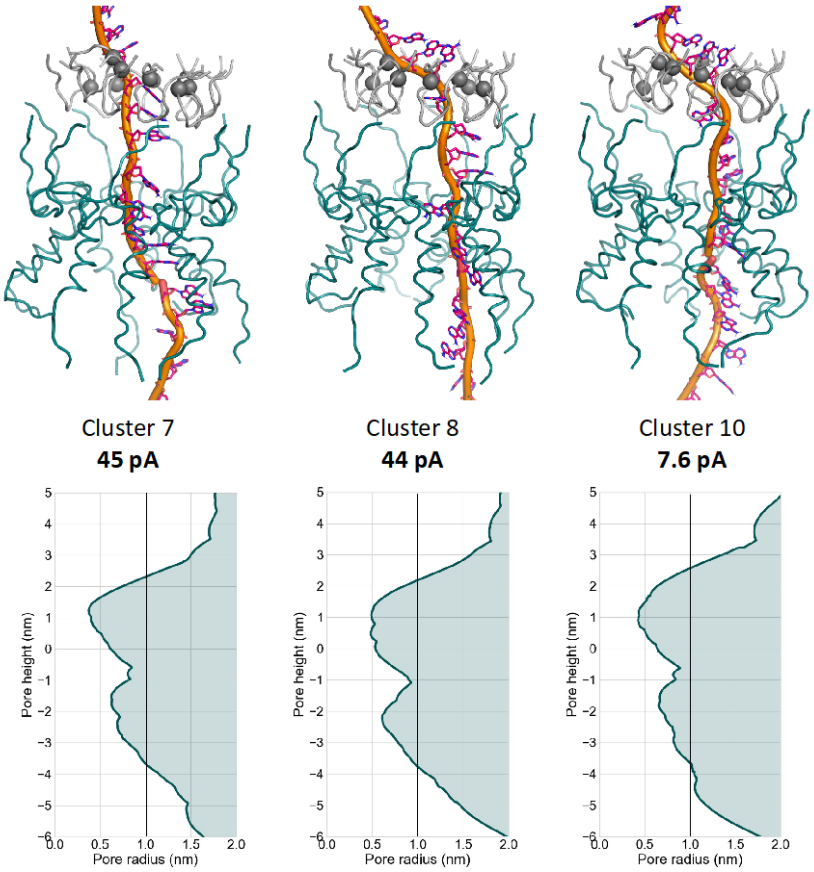

Figure 5.8: The conformation of polyA ssDNA, CsgG eyelet loop region, and CsgF in representative structures of three cluster populations are shown. The Cα atoms of Asn-55 residues in the eyelet loop region are shown as spheres. The pore radius profile for the structures is plotted. The pore height of 0 nm is equivalent to the Z coordinate of Asn-55 residues in the CsgG eyelet loop region.

Table 5.5. The ionic current through the CsgG-CsgF complex during polyA ssDNA translocation and after removing polyA ssDNA, in 0.57 V.

Cluster	I_{total} **(pA)**	
	DNA translocation	Without DNA (100 ns)
7	45 ± 78	377 ± 61
8	44 ± 33	390 ± 103
10	7.6 ± 97	439 ± 57

5.3.2 Translocation of polyC ssDNA

The translocation of polyC ssDNA through uncomplexed CsgG and the CsgG-CsgF complex was investigated to evaluate the impact of the nucleotide size on DNA translocation and the ionic current measured through the protein pores. The translocation of polyC ssDNA was faster compared to polyA ssDNA through uncomplexed CsgG, with at most seven nucleotides exiting the eyelet loop region by 100 ns in three simulations (Figure 5.9). This is concurrent with the smaller size of cytosine in polyC ssDNA compared to adenine in polyA ssDNA. As observed in simulations of polyA ssDNA, polyC ssDNA translocation was largely controlled by the CsgG eyelet loop region. In two simulations, nucleotides were halted in the eyelet loop region ~ 20-60 ns before they were released, which resulted in stepwise DNA translocation through the pore. DNA did not translocate during ~ 40-200 ns in one simulation, during the time in which three nucleotides were halted in and near the eyelet loop region.

In simulations of the CsgG-CsgF complex, the translocation of polyC ssDNA was slower compared to polyA ssDNA, which was opposite of what was observed in uncomplexed CsgG. Little to no polyC ssDNA translocation was observed during 200 ns in three independent simulations. The DNA nucleotides in and near the CsgG eyelet loop region were comparatively more stationary (lower fluctuations of their Z coordinate) compared to the nucleotides in the vestibule and the CsgF region during this time.

Overall, the eyelet loop region had the greatest impact on the translocation of polyC ssDNA and polyA ssDNA through the protein pores, which indicates that DNA translocation is largely controlled by the eyelet loop region regardless of the nucleotide sequence.

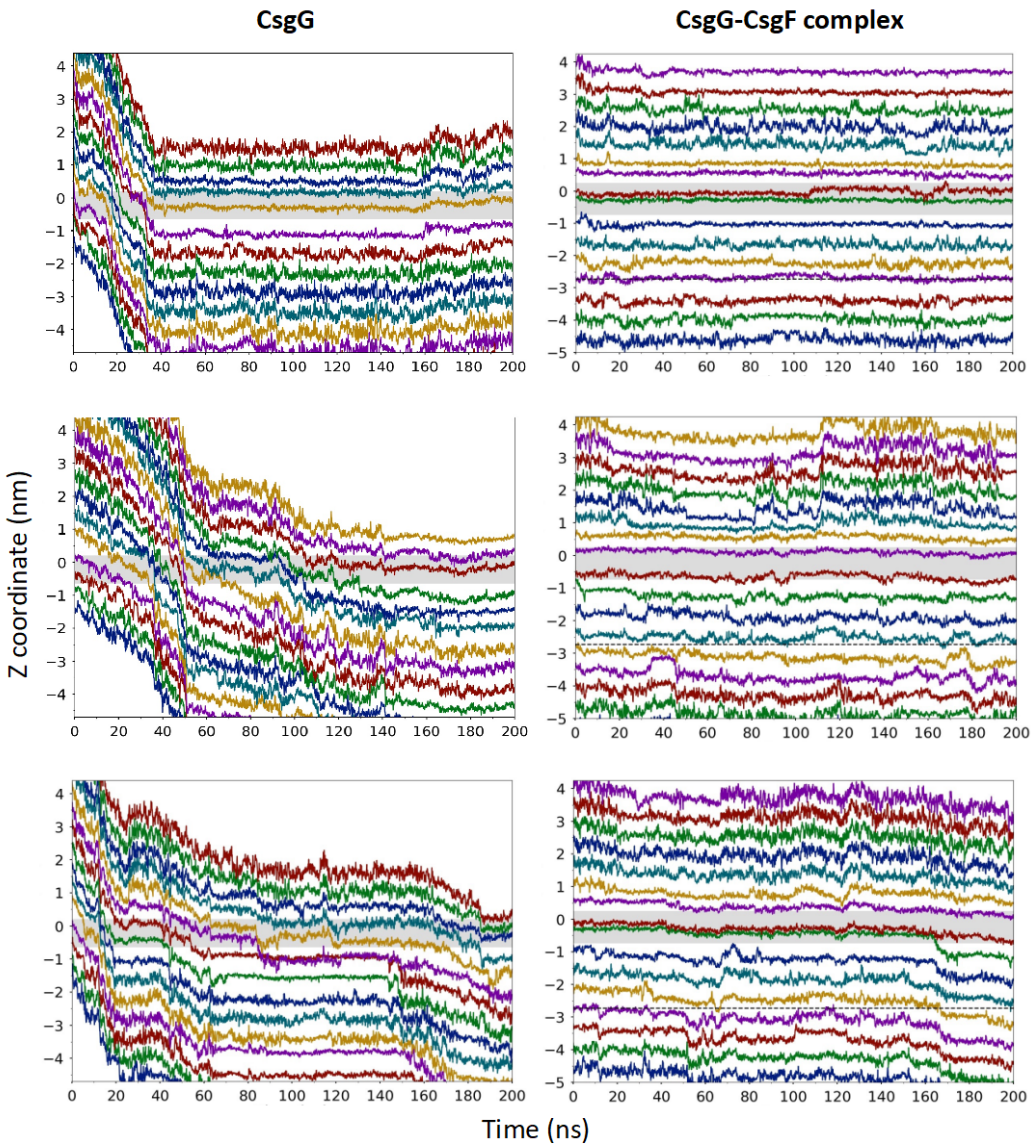

Figure 5.9: The translocation of polyC ssDNA through uncomplexed CsgG and the CsgG-CsgF complex is measured as the Z coordinate of the centre of mass of nucleotides over time in three independent simulations. The eyelet loop region is shaded in grey, and a dashed line marks the CsgF constriction.

The conformations of the DNA strand and the DNA-protein interactions were next examined to elucidate the differences in the translocation of polyC ssDNA compared to polyA ssDNA (Figure 5.10 and Figure 5.11). As before, the interactions of the translocating DNA with uncomplexed CsgG and the CsgG-CsgF complex were quantified by calculating the percentage simulation time that the residues were within 0.4 nm of the DNA strand (Figure 5.12 and Figure 5.13).

In simulations of uncomplexed CsgG, polyC ssDNA remained in an extended conformation through the pore by 200 ns, unlike polyA ssDNA, which was observed to move close to one side of

the pore (Figure 5.10a). polyC ssDNA did not interact with residues in the vestibule region as frequently as polyA ssDNA, which is consistent with faster polyC ssDNA translocation compared to polyA ssDNA. In all simulations, polyC ssDNA interacted most frequently with Phe-56, Asn-55, and Tyr-51 residues in the eyelet loop region (Figure 5.12), and frequently interacted with charged and aromatic residues in the β-barrel region near the pore exit (Figure 5.10b and Figure 5.12), as was observed during polyA ssDNA translocation. The frequency of these interactions was similar for both strands. The DNA-protein interactions by 200 ns in simulation 1, in which DNA translocated the least, are shown in Figure 5.10b. In this simulation, nucleotides were observed to only interact Ser-54 and Asn-55 residues in the eyelet loop region by 200 ns; these interactions in the eyelet loop region were sufficient to halt DNA translocation during ~ 40-200 ns.

In simulations of the CsgG-CsgF complex, polyC ssDNA remained halted in the pore and moved near the pore lining by 200 ns (Figure 5.11a). The DNA-protein interactions by 200 ns in simulation 1, in which DNA translocated the least, are shown in Figure 5.11b. In this simulation, polyC ssDNA formed hydrogen bonds in the eyelet loop region (with Tyr-51 and Asn-55 residues) and the CsgF region (Pro-16, Asn-17, Asn-18, and Asn-24). The frequency of these interactions was higher than those formed during polyA ssDNA translocation, especially in the eyelet loop region, which is consistent with substantially slower translocation observed in these simulations (Figure 5.13). Additionally, polyC ssDNA frequently interacted with Phe-48 residue at the entrance of the eyelet loop region, which further hindered its translocation through the pore. Unlike polyA ssDNA, polyC ssDNA also frequently interacted with residues in the vestibule region due to the strand remaining halted inside the pore for longer durations than polyA ssDNA. In simulation 1, nucleotides interacted with charged (Arg-83 and Lys-94) and polar (Gln-91) residues in the vestibule region by 200 ns, which retained the strand near one side of the pore (Figure 5.11b). In CsgF, polyC ssDNA formed interactions with residues in the constriction region, but not as frequently as during polyA ssDNA translocation (Figure 5.13), which suggests that the eyelet loop region had a greater impact on the DNA translocation rate compared to CsgF.

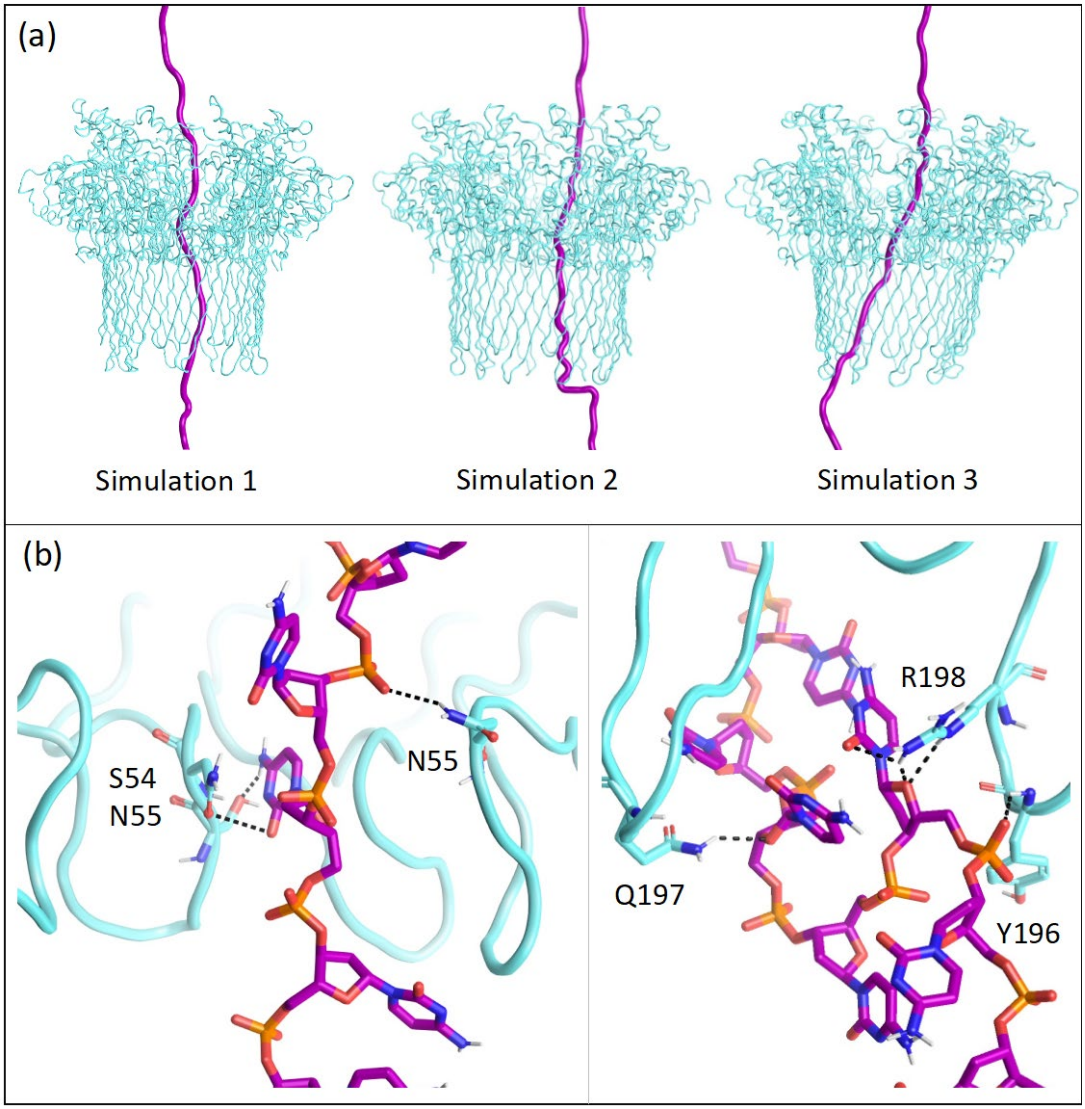

Figure 5.10: (a) The conformation of polyC ssDNA (purple) at 200 ns is shown for three independent simulations of uncomplexed CsgG in 0.57 V. (b) The interactions between polyC ssDNA and the residues in the eyelet loop region in simulation 1 (left), and the β-barrel in simulation 2 (right), at 200 ns are shown. Hydrogen bonds are marked by dashed lines (< 0.32 nm).

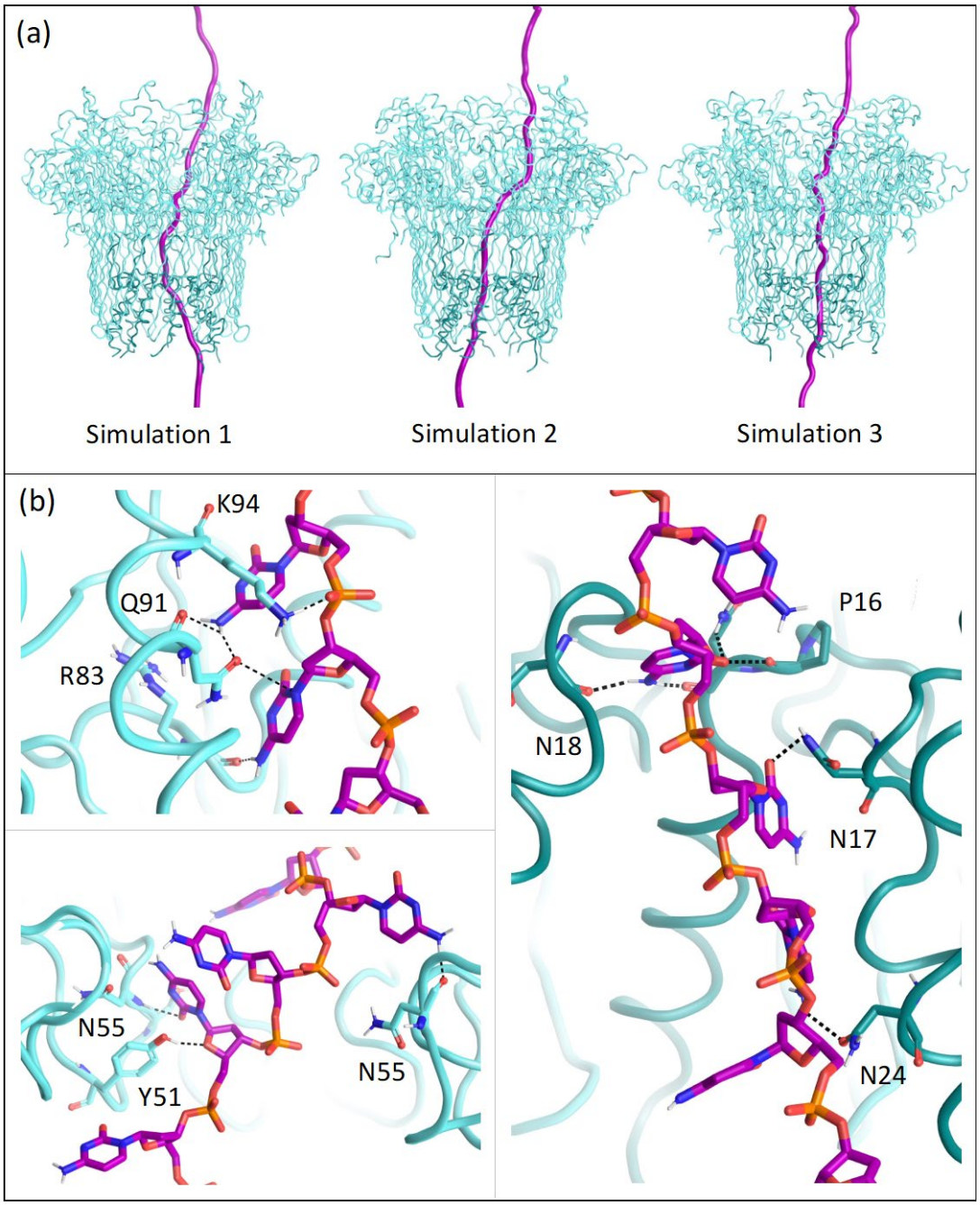

Figure 5.11: (a) The conformation of polyC ssDNA (purple) at 200 ns is shown for three independent simulations of the CsgG-CsgF complex in 0.57 V. (b) The interactions between polyC ssDNA and the residues in the eyelet loop region (left) and CsgF (right) at 200 ns are shown. Hydrogen bonds are marked by dashed lines (< 0.32 nm).

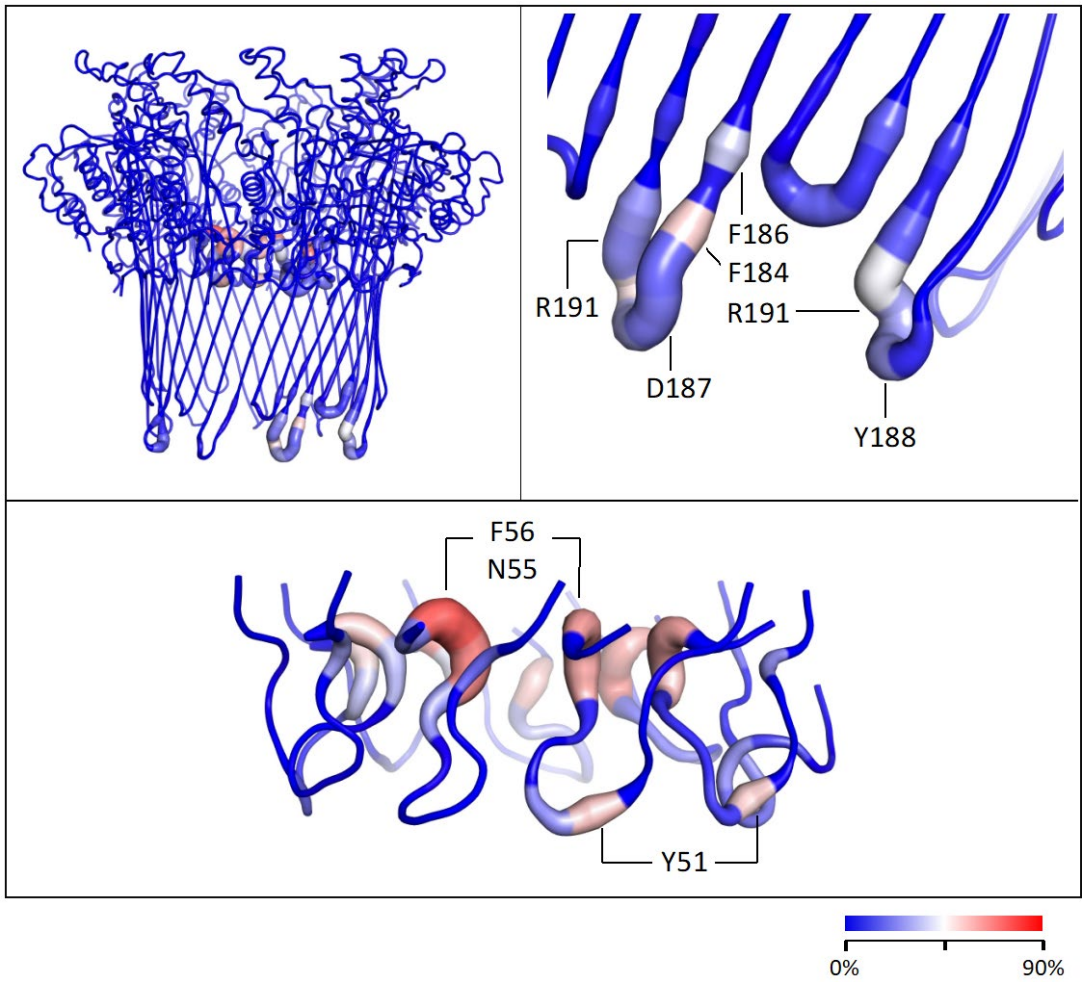

Figure 5.12: polyC ssDNA-protein interactions in simulations of uncomplexed CsgG are coloured by the percentage of simulation time for which the residues interact with the DNA nucleotides in three independent simulations. Interactions are defined as an inter-atomic distance of < 0.4 nm.

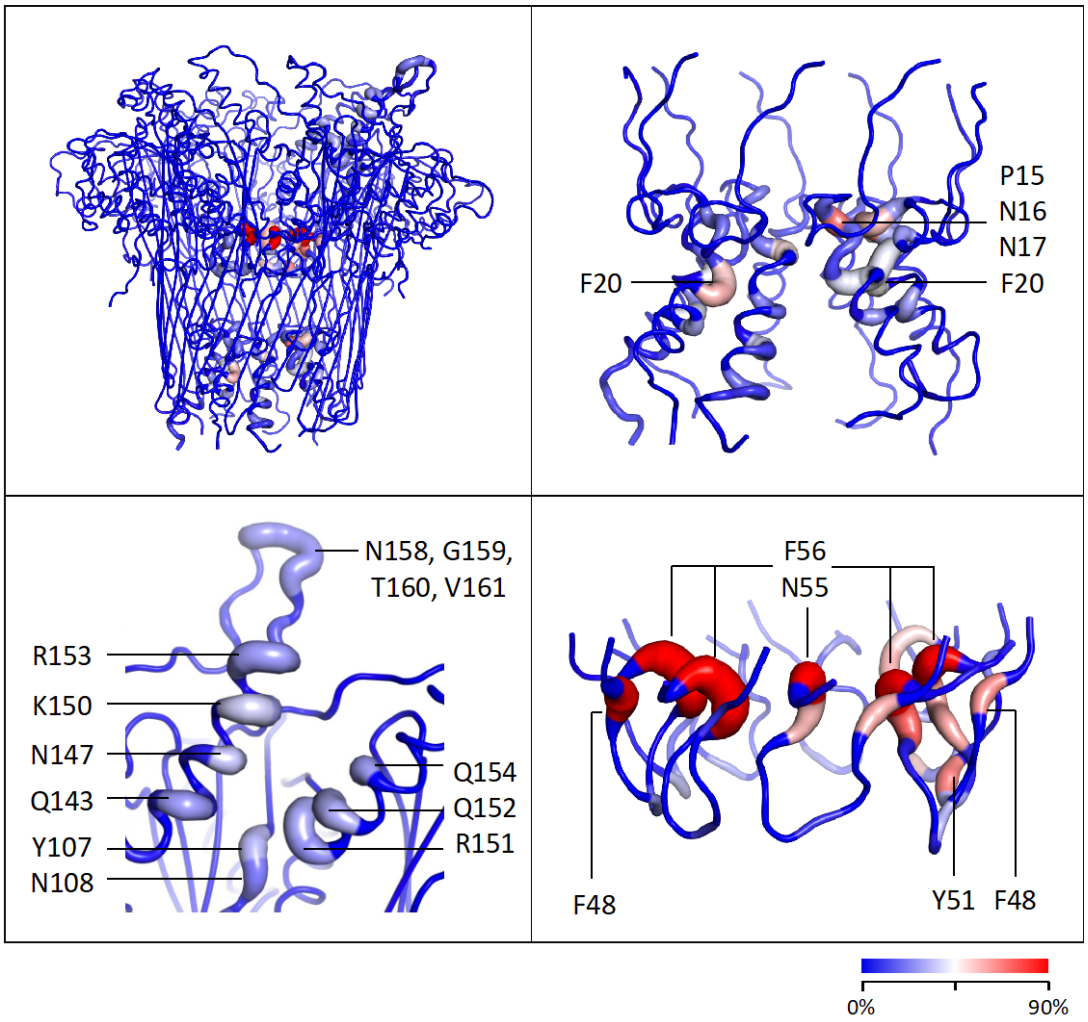

Figure 5.13: polyC ssDNA-protein interactions in simulations of the CsgG-CsgF complex are coloured by the percentage of simulation time for which the residues interact with the DNA nucleotides in three independent simulations. Interactions are defined as an inter-atomic distance of < 0.4 nm.

The conformational drift of the CsgG eyelet loop region from its initial structure was next evaluated to see if its dynamics are impacted by the ssDNA sequence (Figure 5.14). During polyC ssDNA translocation, the eyelet loops were more mobile during polyC ssDNA translocation in uncomplexed CsgG, as indicated by a larger conformational drift exhibited by the eyelet loop region (RMSD ~ 0.30 nm) compared to during polyA ssDNA translocation (RMSD ~ 0.18-0.22 nm). The opposite was observed in simulations of the CsgG-CsgF complex; the eyelet loop region deviated the least from its initial conformation during polyC ssDNA translocation out of all ssDNA simulations (RMSD ~ 0.10 nm). DNA translocation was impacted by the conformational dynamics of the eyelet loop region in both pores. DNA translocation was slower in simulations of polyA

ssDNA through uncomplexed CsgG and polyC ssDNA through the CsgG-CsgF complex, both in which the eyelet loop region was more rigid compared to during translocation of the other strand.

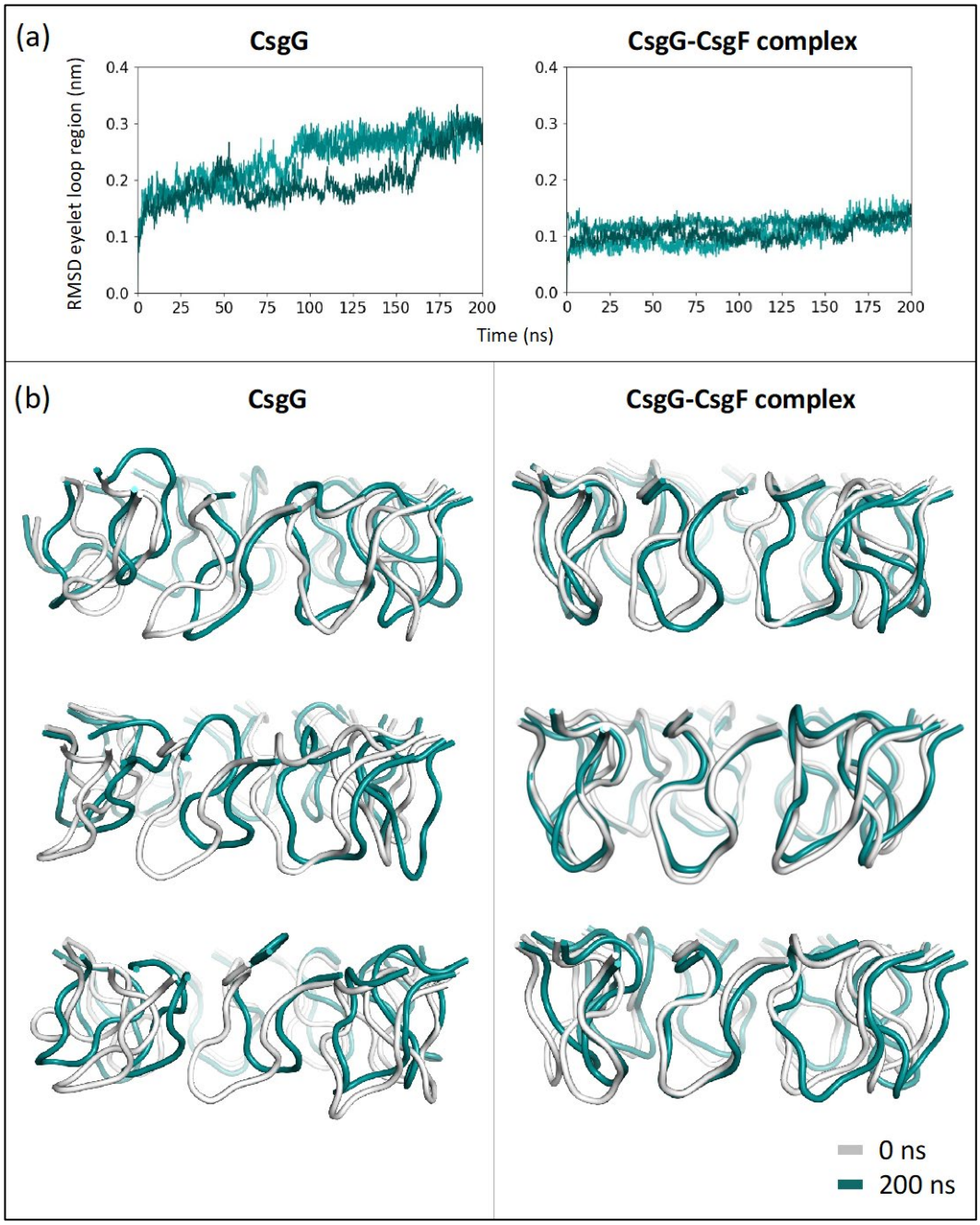

Figure 5.14: (a) RMSD of the eyelet loop region compared to its initial conformation at 0 ns (backbone Cα atoms) is plotted over time for three independent simulations of polyC ssDNA systems. (b) The conformation of the eyelet loop region at 0 ns and 200 ns is shown.

The ionic currents through the pores during polyC ssDNA are reported in Table 5.6. The current through uncomplexed CsgG was on average ~ 2.8 times higher than through the CsgG-CsgF complex. The average current through both pores was higher during polyC ssDNA translocation compared to polyA ssDNA translocation, which is expected due to the smaller size of cytosine in polyC ssDNA occluding the pore to a lesser degree compared to adenine bases in polyA ssDNA. The cumulative current over time is reported in Figure 5.15; a stationary current is indicated by a linear increase of the cumulative current with time. For uncomplexed CsgG, the ionic current was the highest in simulation 1 due to the DNA remaining close to the pore lining during ~ 40-200 ns (Figure 5.10). A steady current flowed through the pore, as indicated by the linear increase in cumulative current over time, due to DNA remaining conformationally restricted once halted inside the pore. In contrast, the current fluctuated over time in two simulations in which DNA translocation was observed. The ionic current during DNA translocation varied substantially amongst three independent simulations; this occurred to a greater degree than during polyA ssDNA translocation (standard deviation: polyC ssDNA = 110, polyA ssDNA = 73) due to the eyelet loop region exhibiting a larger conformational drift from the initial structure in polyC ssDNA simulations compared to polyA ssDNA simulations. Thus, the ionic current varies amongst independent simulations largely due to the conformational dynamics of the translocating ssDNA compared to the dynamics of the CsgG eyelet loop region.

Table 5.6. The ionic current through uncomplexed CsgG and the CsgG-CsgF complex during polyC ssDNA translocation in 0.57 V, calculated for 200 ns in three independent simulations.

System	DNA		I_{total} (pA)	I_K (pA)	I_{Cl} (pA)
CsgG	polyC	Simulation 1	748 ± 52	661 ± 51	87 ± 15
		Simulation 2	529 ± 169	429 ± 142	100 ± 33
		Simulation 3	471 ± 73	434 ± 69	37 ± 13
		Average ± SD	**582 ± 110**	**508 ± 96**	**75 ± 22**
CsgG-CsgF complex	polyC	Simulation 1	143 ± 14	133.3 ± 10.9	10 ± 6
		Simulation 2	277 ± 21	267 ± 20	11 ± 11
		Simulation 3	199 ± 35	186 ± 33	13 ± 9
		Average ± SD	**207 ± 25**	**195 ± 23**	**11 ± 9**

The current through the CsgG-CsgF complex was ~ 2.8 times higher during polyC ssDNA translocation compared to polyA ssDNA translocation. The current through the pore during polyC ssDNA translocation was steadier compared to polyA ssDNA translocation, as indicated by the

linear increase in cumulative current over time. This can be attributed to two factors; polyC ssDNA remained halted and conformationally restricted in the CsgG-CsgF complex in all simulations, and the eyelet loop region did not undergo as many conformational changes as during polyA ssDNA translocation.

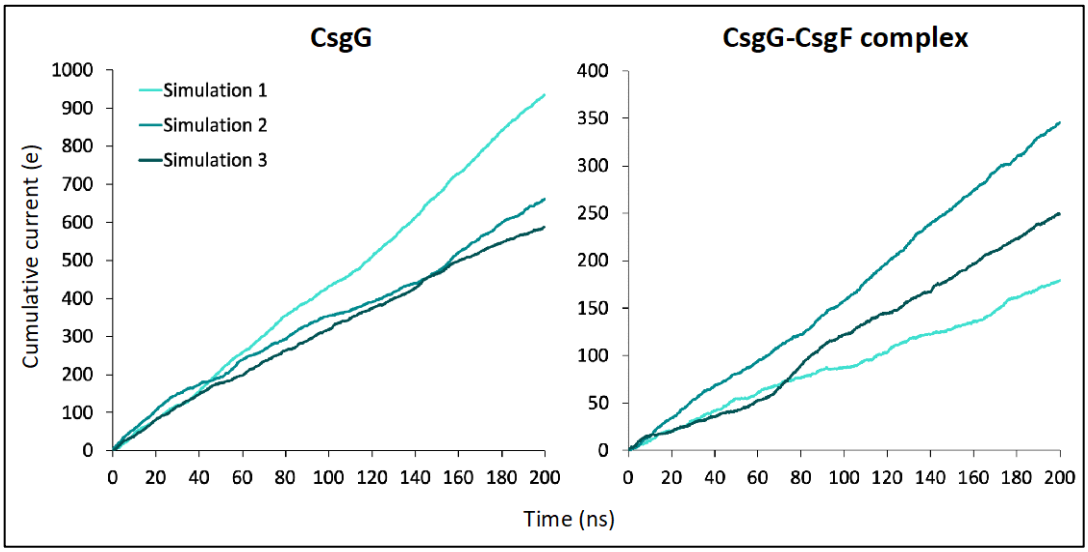

Figure 5.15: The cumulative current is plotted as a function of time in three independent simulations of polyC translocation through uncomplexed CsgG and the CsgG-CsgF complex in 0.57 V. A linear increase of the cumulative currents with time indicates stationary currents; a linear regression fit to these curves gives the average currents in Table 5.6. The cumulative currents are shown in the units of the unitary charge (e = 1.6 × 10^{-19} C).

To obtain ionic currents characteristic to the conformations of polyC ssDNA inside the pore, cluster analysis was performed as was done for simulations of polyA ssDNA to group similar conformations adopted by the DNA inside the constriction region(s) in three independent simulations. For uncomplexed CsgG, the ionic currents for eight cluster populations ranged between ~ 352-1050 pA (Table 5.7), which is higher than polyA ssDNA. However, polyC ssDNA was more coiled inside the pore and the eyelet loop region compared to polyA ssDNA, as indicated by smaller DNA end-to-end distances (DNA in the pore: polyC ssDNA = ~ 4.07-4.53 nm, polyA ssDNA ~ 4.70-5.60 nm; DNA in the eyelet loop region: polyC ssDNA = ~ 1.51-1.78 nm, polyA ssDNA ~ 1.77-1.87 nm). The conformations of eyelet loops in CsgG constriction region were accounted for by calculating SASA of the DNA segments. This showed that the SASA of DNA in the eyelet loop region was higher for polyC (~ 5.6-7.5 nm^2) compared to polyA (~ 5.0-7.2 nm^2), which enabled a

greater flow of ions through the pore during polyC ssDNA translocation. Additionally, the SASA of the eyelet loop region was higher during polyC ssDNA translocation (~ 29.5-33.6 nm^2) compared to polyA ssDNA translocation (~ 24.7-28.2 nm^2). As during polyA ssDNA, the end-to-end distances and the SASA of DNA in the pore and the CsgG eyelet loop region did not correlate with the ionic currents measured for the cluster populations.

Table 5.7. The ionic current is calculated for conformations prominently adopted by polyC ssDNA during translocation through uncomplexed CsgG, obtained from cluster analysis. The DNA end-to-end distance and the SASA of DNA segments inside the pore, and the SASA of the CsgG eyelet loop region, are calculated as an average for each cluster population.

			DNA in CsgG eyelet loop region (4 nucleotides)		**DNA in the pore (9 nucleotides)**		
Cluster	**Duration (ns)**	**I_{total} (pA)**	DNA end-to-end distance (nm)	SASA (nm^2)	DNA end-to-end distance (nm)	SASA (nm^2)	**SASA CsgG eyelet loop region (nm^2)**
1	13.8	1050 ± 214	1.20 ± 0.08	5.6 ± 0.43	4.48 ± 0.16	18.8 ± 0.83	30.4 ± 1.09
2	11.9	723 ± 150	1.74 ± 0.06	5.7 ± 0.44	4.47 ± 0.13	17.4 ± 0.78	29.9 ± 1.02
3	30.6	559 ± 112	1.78 ± 0.10	5.7 ± 0.35	4.53 ± 0.09	18.2 ± 0.72	31.2 ± 0.98
4	22.5	540 ± 89	1.67 ± 0.09	6.6 ± 0.72	4.07 ± 0.16	18.1 ± 1.20	29.5 ± 1.05
5	6.7	519 ± 167	1.60 ± 0.06	7.1 ± 0.33	4.48 ± 0.12	20.2 ± 0.83	33.6 ± 1.17
6	8.0	498 ± 103	1.36 ± 0.21	6.8 ± 0.61	4.10 ± 0.14	18.5 ± 1.08	30.6 ± 0.88
7	10.9	417 ± 218	1.51 ± 0.10	7.5 ± 0.68	4.14 ± 0.17	17.8 ± 1.30	30.2 ± 1.16
8	7.6	352 ± 88	1.62 ± 0.11	6.3 ± 0.93	4.07 ± 0.21	17.5 ± 1.23	32.1 ± 0.99

For the CsgG-CsgF complex, the ionic currents for ten cluster populations ranged between ~ 11-141 pA (Table 5.8), which are lower than polyA ssDNA cluster populations (and opposite to what was observed in uncomplexed CsgG). Although polyC and polyA ssDNA coiled in the eyelet loop region to a similar degree (DNA end-to-end distances: polyA = ~ 1.39-1.80 nm, polyC = ~ 1.38-1.85 nm), the SASA of nucleotides in the eyelet loop region was lower during polyC ssDNA translocation (~ 4.7-6.6 nm^2) compared to polyA ssDNA translocation (~ 4.9-8.5 nm^2), consistent with the lower currents measured through the pore. The differences in the ionic currents through the CsgG-CsgF complex can therefore be attributed to the differences in the conformational dynamics of the eyelet loop region during the DNA translocation. In contrast, SASA of nucleotides in both CsgG and CsgF constriction regions was higher during polyC ssDNA translocation (~ 14.7-

17.5 nm^2) compared to polyA ssDNA translocation (~ 12.8-16.2 nm^2), despite the lower currents measured for polyC ssDNA cluster populations.

As during polyA ssDNA translocation, the DNA end-to-end distances and the SASA of polyC ssDNA in the pore and the CsgG eyelet loop region did not correlate with the ionic currents measured for the cluster populations. This was despite the eyelet loop region undergoing minimal conformational changes during translocation.

Table 5.8. The ionic current is calculated for conformations prominently adopted by polyC ssDNA during translocation through the CsgG-CsgF complex, obtained from cluster analysis. The DNA end-to-end distance and the SASA of DNA segments inside the pore, and the SASA of the CsgG eyelet loop region, are calculated as an average for each cluster population.

			DNA in CsgG eyelet loop region (4 nucleotides)		**DNA in the pore (9 nucleotides)**		
Cluster	**Duration (ns)**	**I_{total} (pA)**	DNA end-to-end distance (nm)	SASA (nm^2)	DNA end-to-end distance (nm)	SASA (nm^2)	**SASA CsgG eyelet loop region (nm^2)**
1	256.1	141 ± 19	1.53 ± 0.08	4.7 ± 0.60	4.50 ± 0.12	14.7 ± 0.84	24.9 ± 0.85
2	72.9	105 ± 48	1.82 ± 0.07	6.61 ± 0.65	4.78 ± 0.13	16.4 ± 0.91	23.6 ± 0.92
3	60.1	78 ± 60	1.82 ± 0.07	5.93 ± 0.67	4.72 ± 0.16	16.3 ± 1.10	24.0 ± 0.91
4	23.5	28 ± 43	1.55 ± 0.08	5.12 ± 0.47	4.57 ± 0.13	16.4 ± 0.89	23.9 ± 1.10
5	14.9	27 ± 81	1.47 ± 0.08	4.75 ± 0.31	4.35 ± 0.11	15.8 ± 1.08	24.6 ± 0.85
6	18.9	26 ± 108	1.72 ± 0.07	6.55 ± 0.53	4.70 ± 0.13	16.6 ± 0.88	23.8 ± 1.02
7	20.9	23 ± 64	1.79 ± 0.06	6.00 ± 0.62	4.80 ± 0.10	16.7 ± 1.05	24.4 ± 0.94
8	14.5	20 ± 118	1.85 ± 0.05	5.63 ± 0.43	4.64 ± 0.09	15.8 ± 1.01	23.7 ± 0.96
9	12.6	17 ± 93	1.38 ± 0.09	5.13 ± 0.51	4.31 ± 0.13	15.4 ± 0.76	24.3 ± 1.15
10	11.6	11 ± 42	1.81 ± 0.06	7.14 ± 0.53	4.60 ± 0.11	17.5 ± 0.84	23.3 ± 0.71

5.4 Conclusions

In conclusion, the translocation of long ssDNA was found to be slower through the CsgG-CsgF complex compared to uncomplexed CsgG. DNA translocation is slowed down in the eyelet loop region as it primarily interacted with Phe-56, Asn-55, and Tyr-51 residues in both protein pores. These interactions occurred more frequently in simulations of the CsgG-CsgF complex compared to uncomplexed CsgG, which is concurrent with the slower DNA translocation observed through the CsgG-CsgF complex. DNA is also slowed by interactions formed by residues in the CsgF

constriction region in the CsgG-CsgF complex, however, these interactions did not occur as frequently as in the eyelet loop region. Therefore, the translocation of long ssDNA is influenced principally by the CsgG eyelet loop region, with the CsgF constriction playing a minor role. This agrees with what was observed during the translocation of short ssDNA in chapter 4.

During polyA ssDNA translocation, there was large variability in the motion of the eyelet loops forming the CsgG constriction region in both uncomplexed CsgG and the CsgG-CsgF complex. The mobility of eyelet loops perturbed the geometry of the CsgG constriction region, which resulted in large fluctuations in the ionic current during DNA translocation. Additionally, the stochastic nature of eyelet loop dynamics resulted in substantial differences in the channel conductance amongst independent simulations. This was observed to a greater degree in simulations of the CsgG-CsgF complex, in which an eyelet loop was observed to 'flip' upwards into the vestibule region in three independent simulations.
The ionic current through the pores did not correlate with the conformations adopted by the polyA ssDNA during translocation. Simulations of the CsgG-CsgF complex without DNA revealed that the conductance did not correlate with the pore width. Thus, the ionic current during polyA ssDNA translocation is influenced by a complex interplay between the dynamics of the DNA and the eyelet loop region.

The translocation of polyC ssDNA through uncomplexed CsgG and the CsgG-CsgF complex was primarily controlled by the CsgG eyelet loop region, which indicates that the influence of the eyelet loop region on DNA movement is independent of the nucleotides in the strand. The translocation of polyC ssDNA was faster compared to polyA ssDNA through uncomplexed CsgG, which is concurrent with the smaller size of cytosine in polyC ssDNA compared to adenine in polyA ssDNA. The eyelet loop region underwent larger conformational changes during polyC ssDNA translocation compared to polyA ssDNA translocation in uncomplexed CsgG. However, polyC ssDNA translocation was slower through the CsgG-CsgF complex. polyC ssDNA remained halted inside the pore as it interacted with residues in the CsgG and CsgF constriction regions more frequently than polyA ssDNA. In contrast to simulations of polyA ssDNA, the geometry of the eyelet loop region remained largely unchanged during polyC ssDNA translocation through the CsgG-CsgF complex, which resulted in reduced fluctuations in the ionic current during the simulations. Despite this, the channel conductance did not correlate with the polyC ssDNA conformations inside the pore.

One of the limitations of this study is that the channel conductance associated with ssDNA conformations in the pores could not be calculated reliably as some conformations were observed for short durations (~ 10 ns) in the simulations. Therefore, it would be of interest to simulate

systems in which the ssDNA and the eyelet loop region are immobilised so that the ionic current can be reliably measured for conformations observed during DNA translocation. The translocation of ssDNA through the pores could be simulated with immobilised eyelet loops to characterise the channel conductance when the eyelet loops remain stationary during translocation. Furthermore, the simulations could be extended for a longer duration (> 500 ns) to evaluate the impact of the protein dynamics on ssDNA translocation and the channel conductance over longer timescales. Lastly, it would be of interest to simulate the translocation of ssDNA that are retained under tension, as done in chapter 3, to ascertain the influence of DNA conformation on the channel conductance during translocation.

In summary, simulations of long ssDNA translocation through uncomplexed CsgG and the CsgG-CsgF complex under an applied electric field provided insights that build upon the findings of chapter 4. More importantly, these simulations revealed the stochastic nature of the eyelet loops of the CsgG constriction region during DNA translocation through uncomplexed CsgG and the CsgG-CsgF. The source of the mobility of the eyelet loops is unclear; although the eyelet loops during polyC ssDNA translocation were more mobile compared to polyA ssDNA translocation in uncomplexed CsgG, the opposite was observed in the CsgG-CsgF complex. Additionally, the conformation of the eyelet loops was found to be stable in the CsgG-CsgF complex in chapter 4, in simulations with and without a short ssDNA immobilised inside the pore.
The mobility of the eyelet loops gave rise to noise in the ionic current during ssDNA translocation through uncomplexed CsgG and the CsgG-CsgF complex. Therefore, significantly reducing the mobility of the eyelet loops in the CsgG constriction region would be an important step in improving the sensing ability of the protein during DNA sequencing.

Chapter 6 Markov State Models for characterising the dynamics of the CsgG eyelet loops

6.1 Introduction

The investigation of the conformational dynamics of CsgG and the translocation of DNA through the protein in chapters 4 and 5 have shown that the eyelet loops forming the CsgG constriction region exhibit large variations in their mobility. The movement, or 'flipping', of the eyelet loops upwards into the vestibule, alters the geometry of the constriction region and results in fluctuations in the ionic current through the protein pore. It is imperative that the nanopores used for DNA sequencing have a stable geometry to ensure a steady ionic current through the pore so that the changes in current during translocation can be associated with the DNA only. The stochastic nature of the behaviour of the eyelet loops is likely to interfere with the application of CsgG as a nanopore for DNA sequencing [204]. It is of great interest to eliminate or reduce the intrinsic activity of the CsgG eyelet loops for optimising the protein pore for DNA sequencing. To achieve this, an extensive understanding of the conformational dynamics of the CsgG eyelet loops is beneficial.

Although MD simulations are proven to be successful in studying the complex dynamics of proteins, they are restricted by sampling limitations. Consequently, large regions of the energy landscape may remain unexplored, which can hinder the understanding of functionally important slow motions. To elucidate the dynamics of the CsgG eyelet loops, a comprehensive model of their conformational ensemble and the underlying free energy landscape is required; these can be obtained from a Markov State Model (MSM) [205-214]. The aim of this chapter is to provide atomistic insights into the mechanism governing the flipping of the CsgG eyelet loops. MD simulations of CsgG, under an applied electric field equivalent to 0.9 V, were used to construct an MSM for the process.

6.2 Methods

6.2.1 MD simulations

The details of the systems and the MD simulation protocol are provided in chapter 4 (section 4.2). The initial structures of CsgG were taken from: CsgG-1 - CsgG crystal structure (PDB 4UV3, 3.59 Å), and CsgG-2 - CsgG taken from the electron cryo-EM structure of the CsgG-CsgF complex with CsgAN6 peptide (PDB 6L7C, 3.34 Å). The simulations of CsgG systems under an applied electric

field equivalent to 0.9 V across the membrane, discussed in chapter 4, were extended to > 200 ns for this study.

Analyses were performed using GROMACS utilities and locally written code. The inter-monomer contact analyses were performed using CONAN [215]. The molecular graphics images were generated using VMD [165] and PyMOL [156].

6.2.2 Markov State Models

MSMs are frameworks parameterised from observations that can be used to predict key thermodynamic and kinetic properties of the system. MSMs simplify the dynamics of the systems by identifying identify long-lived conformations, also known as metastable states, and describing the dynamics as a matrix of conditional transition probabilities between these states. The dynamical evolution of the system is modelled as a Markov chain by sampling the system's state using a timestep (known as the lag time, τ) that is long enough for the transition between states to be Markovian, i.e., the probability of transition from one state to another, after the next increment of lag time, is independent of previous transitions. The MSM is an *n* x *n* square matrix, where *n* is the number of states that divide the conformational space of the system, in which the probability of transition between row-indexed state to column-indexed state after the next increment of lag time is recorded. Multiple short MD simulations, each exploring different regions of the energy landscape of the system, can be combined to build an MSM which can provide insights into long-timescale events due to their Markovian property [216-219].

MSMs have been successfully used to unravel the thermodynamics, kinetics, principal motions, and transition pathways (and their probabilities) of complex processes such as protein folding [220-226], protein-ligand binding [227-232], enzyme kinetics [228, 233-236], and the conformational changes experienced by proteins [214, 237-240], and the statistical uncertainties for all observations. In this chapter, the MD simulation trajectories were processed and used to construct a Markov state model using the software package PyEMMA [241].

6.2.2.1 Description of the method

The construction of MSMs involves reducing the high-dimensional dynamics in MD simulations, followed by partitioning the conformational space of the system into discrete states. A transition counts matrix is subsequently generated and is used to construct an MSM, which describes the conditional probabilities associated with transitions between discrete states at a fixed time τ later.

6.2.2.1.1 Dimensionality reduction

The high-dimensional conformational space is formed by multiple properties, or 'features', describing the dynamical system. For example, the dynamics of proteins can be described by the cartesian coordinates of protein backbone (Cα) atoms, inter-residue distances, and protein backbone or side chain torsion angles. As many of these features contain redundant information, the conformational space is reduced by firstly transforming the coordinates in each frame of the MD trajectories into a feature vector that describes the relevant dynamics of the system. The relevant features can be systematically chosen using the Variational Approach to Markov Processes (VAMP), which enables direct and quantitative comparison of multiple features and their ability to capture the slow long-term dynamics of the system [242, 243]. Specifically, the VAMP-2 score quantifies the kinetic variance contained in the features [244, 245]. The VAMP-2 score is calculated for the relevant features for a range of lag times to ensure robust feature selection.

Next, a linear coordinate transformation is conducted to remove redundant information within the feature space. Time-lagged independent component analysis (TICA) [246] is popularly used to reduce the dimensionality of the feature space $(X(t))$ to the eigenvectors of an autocovariance matrix $(X(t)X^T(t+\tau))$, where $X(t)$ is the data vector at a time t and τ is the lag time. The slowest collective coordinates describing the important dynamics, which correspond to eigenvectors with the largest eigenvalues, are retained [246-248]. It is imperative to select an optimal number of TICA eigenvectors, as an MSM built using too many eigenvectors will contain microstates with low statistical significance due to the finite sampling error [249].

In this chapter, the features were selected using a procedure described in ref. [238], which involves iteratively performing TICA and retaining features that are correlated with the first four eigenvectors (Pearson correlation value of > 0.4). The final features were used to featurise the MD trajectories and perform TICA to reduce the dimensionality of the feature space.

6.2.2.1.2 Discretisation

The reduced conformational space of the transformed data is discretised into multiple microstates using clustering methods. The typically used method is the *k*-means algorithm [250, 251]. Firstly, the number of clusters to be used is calculated as $\sqrt{N}$, where N is the total number of available samples in the input trajectories [241]. The *k*-means algorithm assigns each frame of the MD trajectory to the nearest one of the *k* cluster centres on the state space, the positions of which are initially randomly selected. The positions of the *k* cluster centres are adjusted by computing the mean of all data points in each cluster, and the frames from MD trajectories are

re-assigned to the new *k* cluster centres. This process is repeated to iteratively optimise the *k* cluster centres until their positions remain unchanged. This results in the discretisation of the multi-dimensional conformational space into Voronoi cells, each representing a microstate.

6.2.2.1.3 Markov state model construction

The discretised data is used to construct a transition count matrix (TCM), in which the number of transitions between microstates is recorded. The TCM is symmetrised using a maximum likelihood approach to satisfy detailed balance, meaning that the transition from states j to i is equivalent to the transition in the reverse direction from i to j under equilibrium conditions [216, 221] . This is used to calculate the conditional probabilities of transitions between microstates within the lag time to generate a transition probability matrix (TPM). Information on the thermodynamics and kinetics of the system can be gleaned by decomposing the TPM into its eigenvectors and eigenvalues [216, 252]. Barring the eigenvector with the largest value (equal to 1), which represents the equilibrium population of the system, the eigenvectors are related to the collective transition modes between states and their eigenvalues are related to the relaxation timescale of the transition process [216]. The eigenvectors are ordered according to the speed of dynamics of the system, i.e., the second largest eigenvalue represents the slowest relaxation dynamics, the third largest eigenvalue represents the second slowest dynamics of the system, and so on.

The lag time used for constructing the TPM must ensure Markovian behaviour. To select the optimal lag time, multiple TPMs are generated for various lag times, and the relaxation timescales of the system are extracted using the relation:

$$\tau_i = \frac{\tau'}{\ln\lambda_i} \qquad 6.1$$

where τ' is the lag time used, λ_i is the i^{th} eigenvalue of the transition probability matrix, and the resulting τ_i is called the implied timescale (ITS) [206] corresponding to the i^{th} relaxation mode of the system. If the TPM for a given lag time is Markovian, τ_i is independent of τ'.

6.3 Results and Discussion

6.3.1 Conformational dynamics of CsgG eyelet loops

The conformational dynamics of CsgG eyelet loops were evaluated in six independent simulations of CsgG systems under an applied electric field equivalent to 0.9 V across the membrane. The systems consisted of CsgG embedded in a POPC lipid bilayer and solvated in 1 M KCl. The eyelet loops are formed by residues 47-58, with the amino acid sequence: Gln-Phe-Lys-Pro-Tyr-Pro-Ala-Ser-Asn-Phe-Ser-Thr.

The conformational drift of the eyelet loops during the simulations were evaluated by monitoring the root mean square deviation (RMSD) of the protein backbone (Cα atoms) compared to their initial conformation at 0 ns (Figure 6.1). The RMSD of eyelet loops that moved upwards were observed to reach values as high as ~ 0.80 nm in some simulations, indicating large-scale motions compared to the rest of the eyelet loops for which RMSD ranged between ~ 0.10-0.40 nm. At least two eyelet loops were observed to flip upwards to varying degrees in all simulations, with at most four eyelet loops flipping by 200 ns in one of the simulations. The relative positions of the monomers in which the eyelet loops moved upwards differed amongst independent simulations, which suggests an absence of significant monomer-monomer cooperativity involved in the process (Figure 6.2).

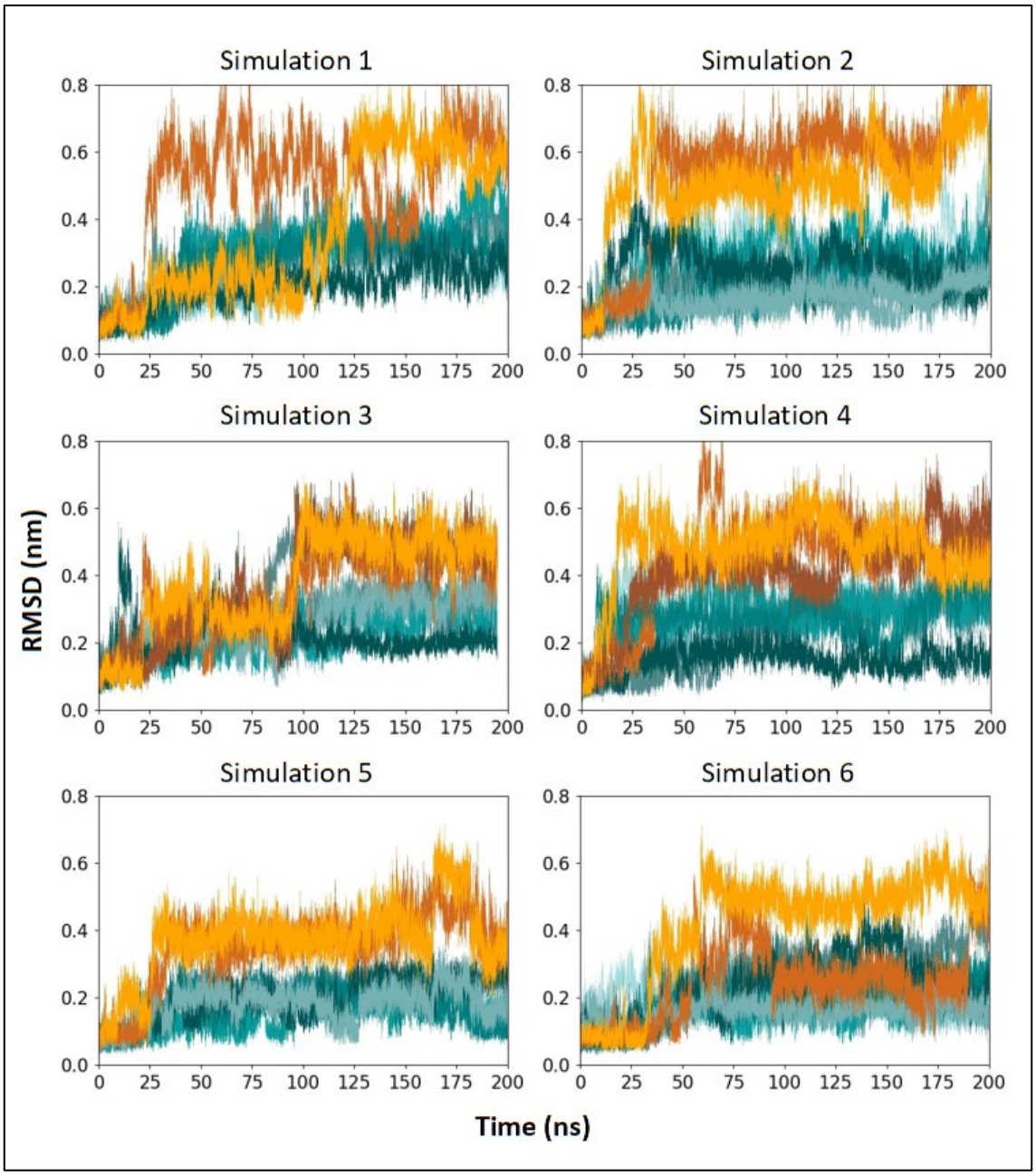

Figure 6.1: RMSD of eyelet loops compared to their initial conformation at 0 ns (backbone Cα atoms) is plotted over time for nine monomers in six independent simulations. RMSD of eyelet loops that were observed to flip up is plotted in orange.

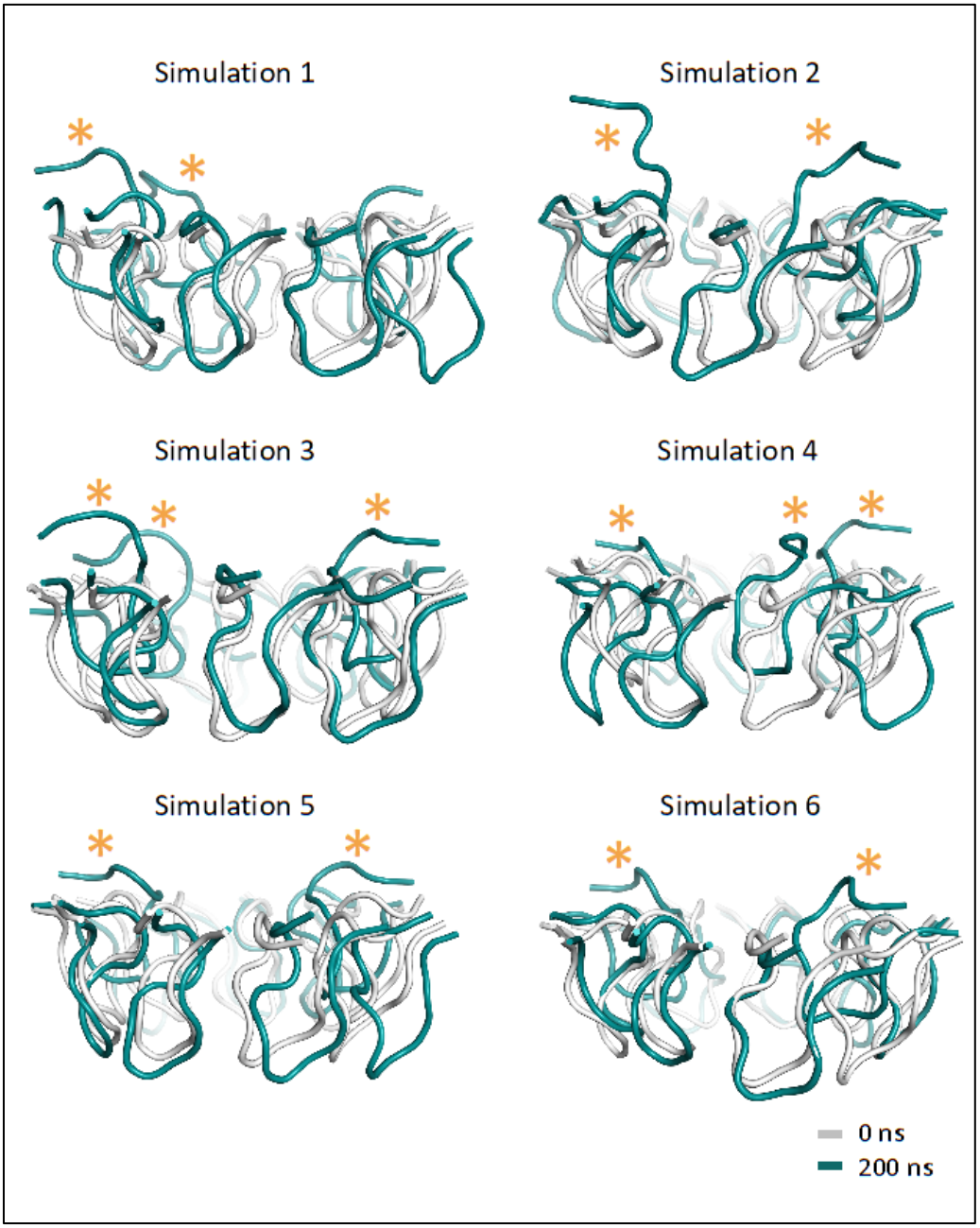

Figure 6.2: The conformation of the eyelet loop region at 0 ns and 200 ns in six independent simulations is shown. The asterisks mark the eyelet loops that flipped up by 200 ns.

Next, to evaluate whether the flipping of the eyelet loop affects the interactions formed between the monomer and adjacent monomers, the inter-monomer interactions were compared pre- and post-eyelet loop flipping (Figure 6.3). This showed that although the inter-monomer interactions are largely unaffected, the flipping of the eyelet loops has varying effects; whilst new inter-monomer interactions were formed following eyelet loop flipping in some monomer pairs, some existing interactions were abolished in other monomer pairs. This demonstrates that the eyelet loop dynamics and their impact on CsgG are complex in nature.

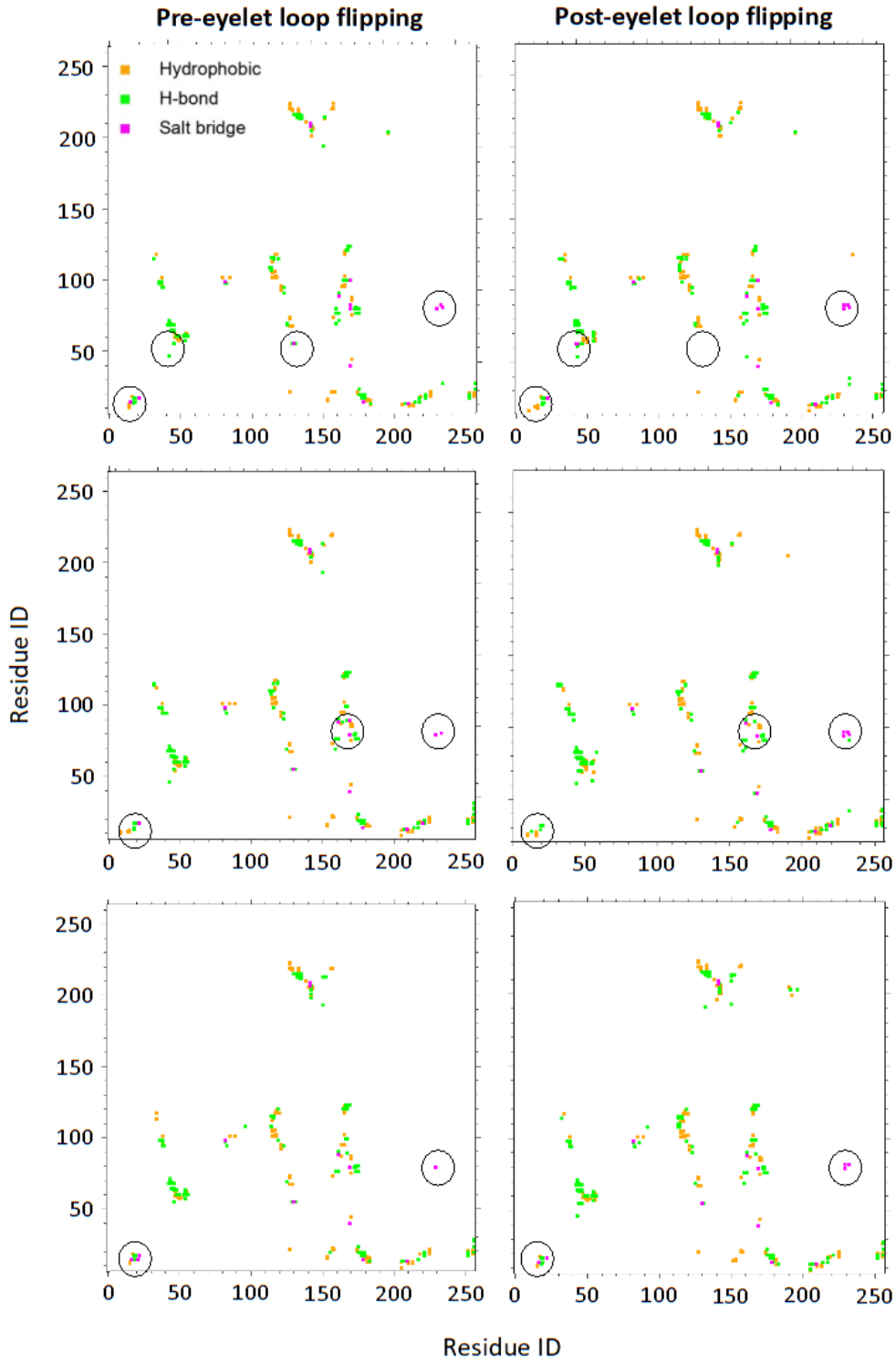

Figure 6.3: The inter-monomer interactions pre- and post-eyelet loop flipping are shown for three CsgG monomer pairs, in which the eyelet loop flipped upwards in one of the monomers. The inter-monomer interactions are coloured according to the type of interaction: hydrophobic interactions, hydrogen bonds, and salt bridge interactions. The differences in inter-monomer interactions pre- and post-eyelet loop flipping are circled. Data is from independent simulations.

6.3.2 Markov State Models of CsgG

6.3.2.1 Markov State Model of whole protein

To identify the key conformational states of the protein and kinetics of state transitions as a result of eyelet loops flipping, an attempt was made to construct an MSM for CsgG. The MD simulation trajectories were featurised using the distances between residue pairs observed to form hydrogen bonds for > 20 % of the total simulation time. This feature was deemed relevant as the flipping of the eyelet loops was followed by changes in the inter-monomer interactions (Figure 6.3). A total of 399 residue pairs were identified, and the minimum distances between any heavy atoms of the residues in a pair were used to create the feature space. To reduce the dimensionality of the feature space, TICA was performed iteratively, and the features with a low correlation (< 0.4) with the first four TICA eigenvectors were eliminated [238]. This method enables the identification of features most relevant to the dynamics of the system in an unbiased way. A final number of 41 residue pairs were identified to describe the slowest collective coordinates using this methodology. These interacting pairs were mainly in the vestibule region, indicating that this region is involved in the slowest and most significant motions of CsgG in these simulations (Figure 6.4). Additionally, many interacting pairs involved at least one residue with a charged sidechain; this is likely due to the change in sidechain orientation caused by force acting on the charged groups in an applied electric field.

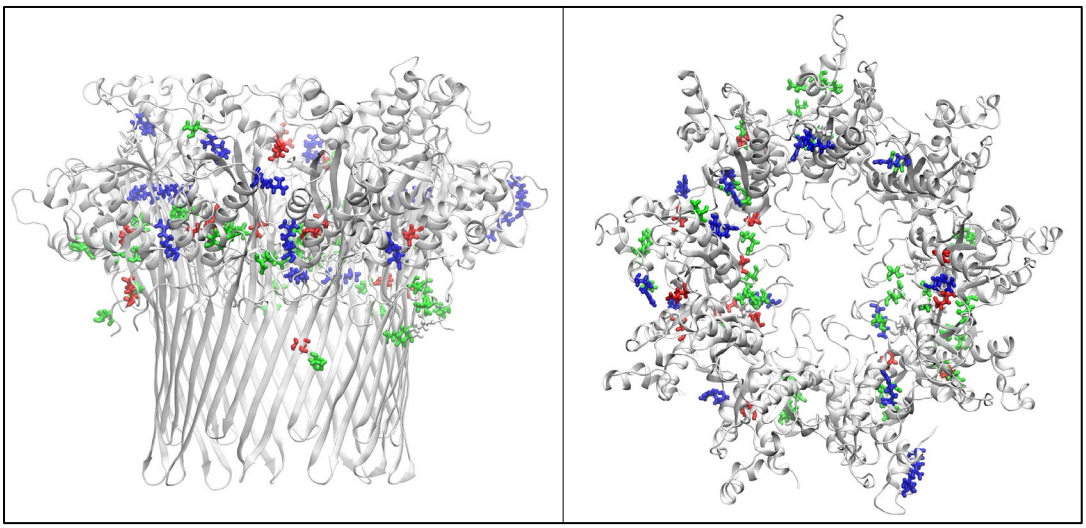

Figure 6.4: Side and periplasmic views of CsgG (grey), with residues forming the 41 residue pairs identified to describe the slowest collective coordinates shown in stick representation. Red: anionic side chains, blue: cationic side chains, green: polar side chains, and white: non-polar side chains.

The MD simulation trajectories and the free energy landscape in the subspace of the first two TICA eigenvectors are shown in Figure 6.5. These showed that the conformations of CsgG did not overlap between the simulations; the three minima in the free energy landscape each correspond to the preferred conformation of CsgG in one of the three simulations only. This is due to the analysis not accounting for the symmetric nature of CsgG in which the nine monomers are identical. Thus, the analysis treats similar conformations observed during all simulations as unique, as illustrated in Figure 6.6.

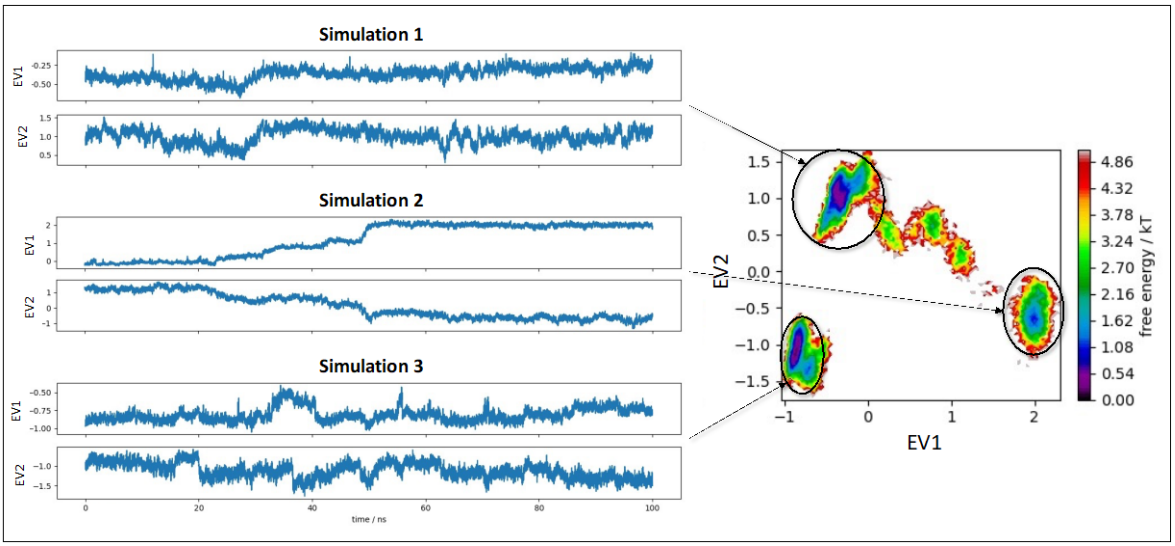

Figure 6.5: The trajectories of three independent simulations of CsgG are plotted over time in the subspace of the first two eigenvectors (left). The free energy landscape projected on the top two TICA eigenvectors is shown (right). The regions with the energy minima are circled, and the corresponding simulations are indicated by arrows.

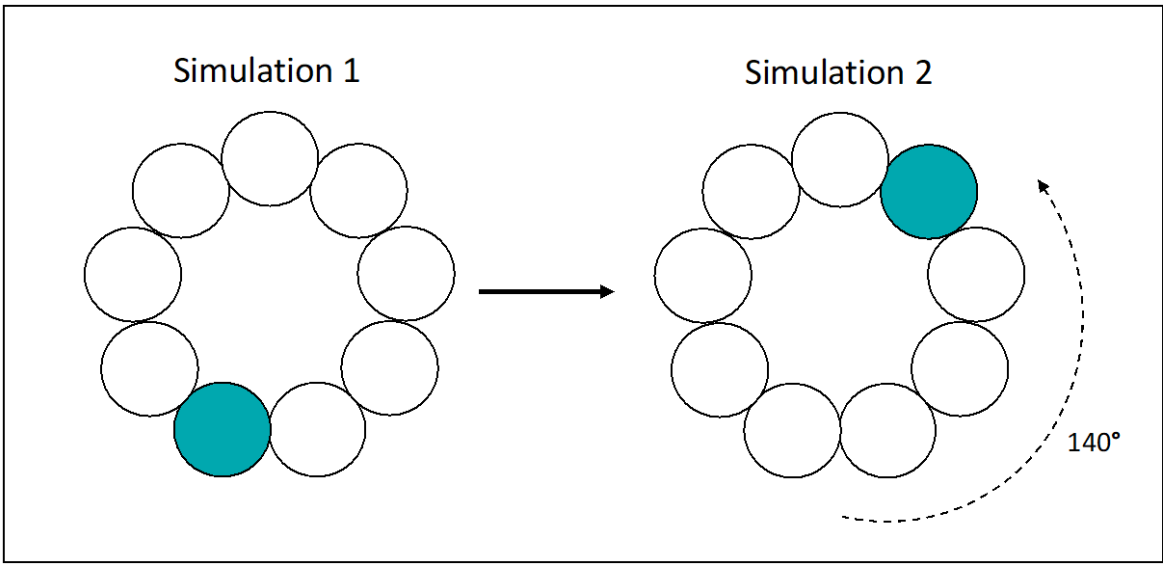

Figure 6.6: Schematic of CsgG, with the nine monomers shown as circles. The conformation of CsgG following the flipping of an eyelet loop in one monomer (cyan) is identical in two simulations in which the position of the monomer differs, as all monomers are identical.

6.3.2.2 Markov State Model analysis of CsgG monomers

As the eyelet loop dynamics were largely uncorrelated between monomers (Figures 6.1 and 6.2), indicating no significant monomer-monomer cooperativity, the trajectory of each CsgG monomer was considered independently to overcome the issues associated with constructing an MSM of the whole protein. An attempt was made to construct an MSM using trajectories of only the monomers in which the eyelet loop was observed to move upwards to characterise the mechanism of the process. An aggregate trajectory data of ~ 3428 ns was used for analyses (14 monomers). Firstly, several features were compared using the VAMP-2 score to identify the feature with the largest kinetic variance (Figure 6.7). This revealed that the highest scoring features across different lag times are the torsion angles (illustrated in Figure 6.8) of the eyelet loop residues compared to the torsion angles of all residues in the monomer and the distances between residues in the eyelet loop region and the rest of monomer that are within ~ 0.5 nm of the eyelet loop region at the beginning of the simulations. Thus, the coordinates in the MD simulation trajectories were transformed, and the extract torsion angles of the eyelet loop residues were extracted for further analysis.

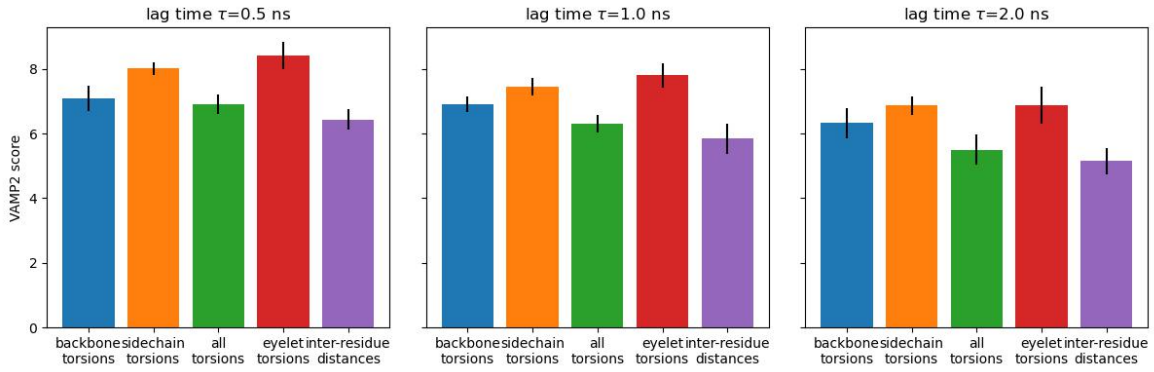

Figure 6.7: VAMP-2 score of features at lag times of 0.5 ns, 1 ns, and 2 ns, with higher scores corresponding to larger kinetic variance described by the feature. Backbone torsions: ϕ and ψ angles in all residues in the monomer, sidechain torsions: χ_1 - χ_3 angles in sidechains of all residues in the monomer, eyelet torsions: all backbone and sidechain torsion angles of residues forming the eyelet loop (residue 47-58), and inter-residue distances: between residues in the eyelet loop region and the rest of monomer that are within ~ 0.5 nm of the eyelet loop region at 0 ns.

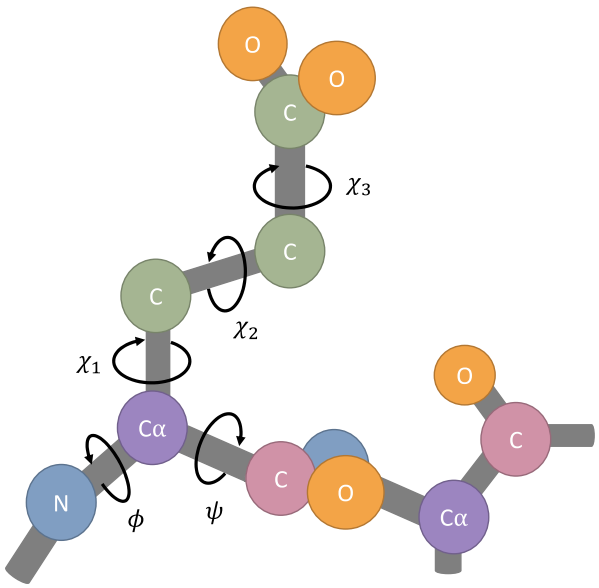

Figure 6.8: The torsion angles in a peptide chain include ϕ and ψ describing the rotation around bonds in the peptide backbone, and χ_1 - χ_3 angles describing the rotations around the bonds in the residue sidechains.

Next, TICA was performed iteratively and the features with a low correlation (< 0.4) with the first four TICA eigenvectors were eliminated. The 92 features were iteratively reduced to 38 features that were involved in the slowest and most significant motions associated with eyelet loop

movement in these simulations. The features included 12 backbone torsion angles of 3 residues that link the eyelet loop to the rest of the monomer, and 26 sidechain torsion angles of 6 residues positioned near both ends of the loop (Figure 6.9). The MD simulation trajectory of one of the 14 monomers and the free energy landscape in the subspace of the first two TICA eigenvectors are shown in Figure 6.10. The slowest eigenvectors clearly capture the conformational change associated with eyelet loop flipping as a distinct change in the values of the first two eigenvectors. The slowest eigenvectors also capture the slow motions associated with smaller conformational changes taking place following eyelet loop flipping. The free energy landscape shows several minima which correspond to stable eyelet loop conformations observed both prior to and following eyelet loop flipping.

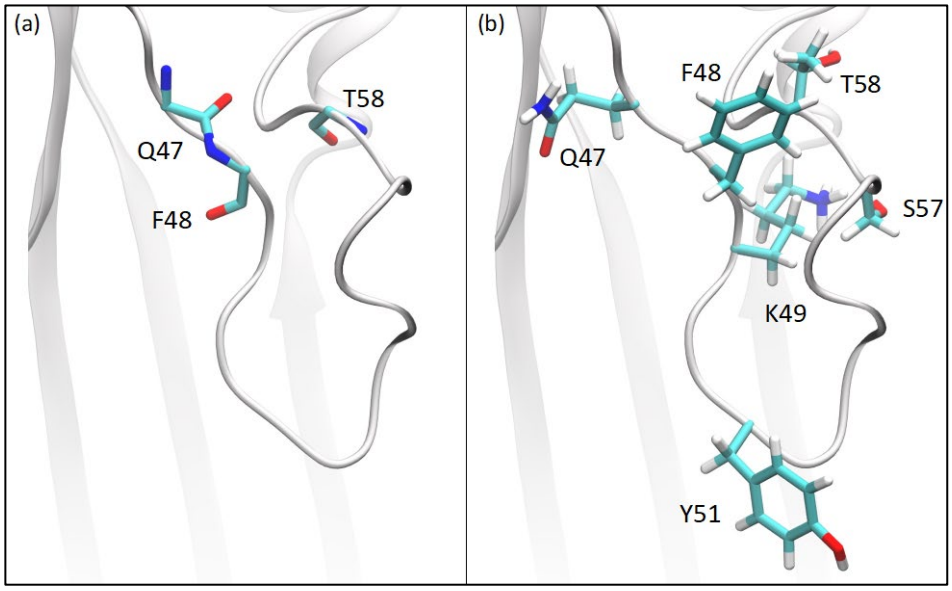

Figure 6.9: Side view of the CsgG eyelet loop (grey), with backbone (a) and sidechains (b) of residues that were identified to describe the slowest collective coordinates shown in stick representation.

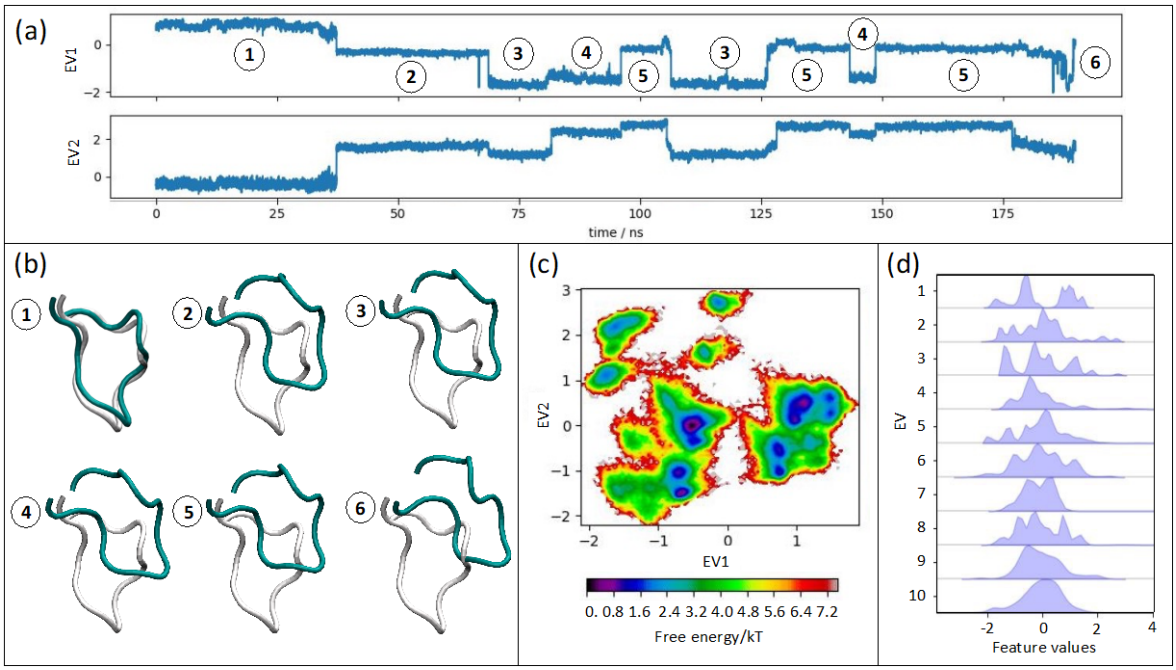

Figure 6.10: (a) A trajectory of one of the CsgG monomers is plotted over time in the subspace of the first two eigenvectors. The values of the eigenvectors corresponding to different eyelet loop conformations are numbered. (b) The conformations of the eyelet loop corresponding to simulation times labelled in (a) are shown (teal). The conformation at 0 ns is shown (grey) for comparison. (c) The free energy landscape projected on the top two TICA eigenvectors is shown. (d) The distributions of the top 10 TICA eigenvectors are shown.

It is imperative to select an optimal number of TICA eigenvectors, as too many eigenvectors can result in microstates with low statistical significance due to a finite sampling error [249]. The first nine TICA eigenvectors were selected for constructing the MSM, as their distribution significantly differed from normal distribution (Figure 6.10d) [237]. Next, the conformational space was discretised into multiple microstates using *k*-means clustering, and multiple transition probability matrices were generated to identify the optimal lag time for Markovian behaviour. The implied timescales (ITS) of the slowest process were not observed to converge at any of the lag times ranging between 1-300 ns, i.e., the ITS of these processes varied with the lag time used. To account for the influence of clustering on the ITS, the number of microstates was varied for discretising the conformational space. The ITS did not converge for any level of discretisation; hence the data was unsuitable for constructing an MSM (Figure 6.11).

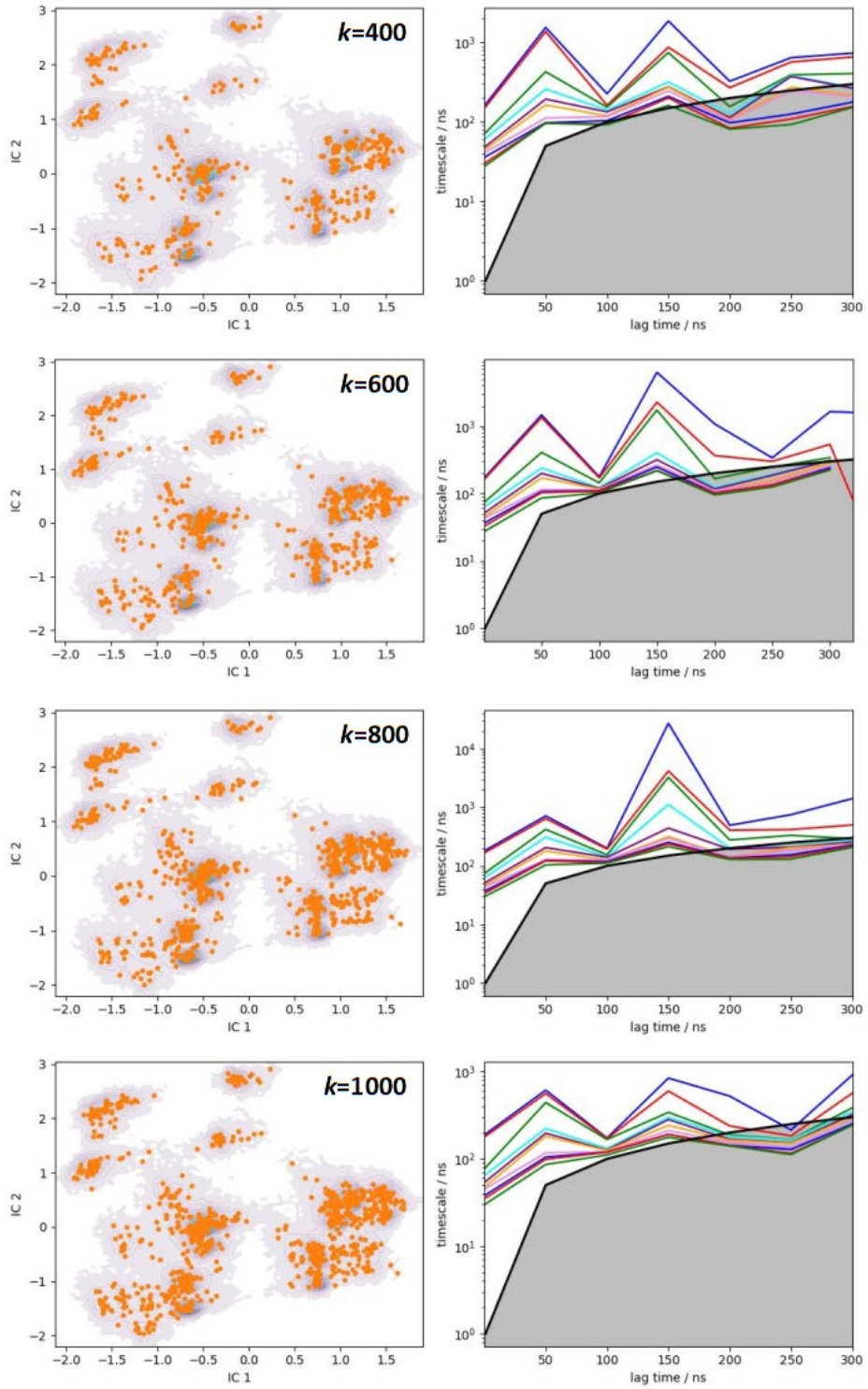

Figure 6.11: The cluster centres (orange) are plotted on the free energy landscape (grey) projected on the top two TICA eigenvectors for different levels of clustering (left). The implied timescales (ITS) of 10 slow processes are plotted at multiple lag times and calculated using the discretised data (right). The grey area indicates timescales that are shorter than the lag time, i.e., processes that are faster than the lag time.

6.4 Conclusions

In conclusion, the dynamics of the eyelet loops forming the CsgG constriction under an applied electric field are complex and vary amongst independent simulations. The relative positions and the number of monomers in which the eyelet loops flipped upwards differed amongst all simulations analysed in this chapter. The sampling limitations associated with MD simulations greatly restrict the investigation of the complex dynamics of the eyelet loops. Hence, an attempt was made to construct a Markov State Model (MSM) to elucidate the CsgG eyelet loops dynamics, as comprehensive insights into the mechanism governing the flipping of the eyelet loops can be obtained from these models. The flipping of the eyelet loops was found to correlate with changes in the torsion angles of the residues near both ends of the loops; hence these residues were used to build an MSM. However, it was not possible to construct an MSM because a Markovian lag time could not be obtained for the MD simulation trajectories. The non-convergence of the implied timescales within 200 ns indicates that longer simulations are required to obtain a Markovian lag time for constructing an MSM.

The selection of lag time for MSM construction raises several challenges for complex systems such as CsgG monomers. Examination of ITS for selecting the lag time is mathematically well-motivated but can result in lag times identified exceeding 100 ns, despite the processes of interest taking place at timescales as short as $\lesssim$10 ns [253]. The validated lag time cannot reliably describe the processes of interest and instead describes slow but uninteresting events [206, 211, 216]. This was also observed in this chapter; the ITS did not converge within ~ 300 ns (Figure 6.11), despite the eyelet loops flipping within ~ 10 ns (Figure 6.1). To circumvent this, longer MD simulation trajectories could be used to construct the MSM [253], but more importantly, the parameters selected at each stage of model construction could be optimised, including the choice of features, the number of TICA eigenvectors retained for the discretisation of the conformational space, and the number of microstates used for clustering [216, 241]. Additionally, it is possible to construct MSMs using a shorter lag time at which the ITS of the slowest process initially converges, even if it diverges later on so that the lag time reflects the timescales associated with the process of interest [254, 255]. Unfortunately, these approaches could not be implemented due to time constraints.

Another strategy to investigate the flipping of CsgG eyelet loops is to employ metadynamics [256, 257]. Metadynamics is a widely used method in MD simulations, which accelerates the sampling of processes of interest so that the free energy can be estimated for complex systems [258, 259]. It can provide insights into the free energy associated with the conformational transitions that take place during processes such as eyelet loop flipping, which can be used to determine whether

the process is favourable to occur in experiments. Moreover, previous studies have used metadynamics for constructing MSMs of systems, as it samples the free energy landscape to a greater degree than unbiased MD simulations [260-263].

In summary, the findings from this chapter and chapters 4 and 5 emphasise the need for a model which describes the dynamics of the eyelet loops simply and quantitatively. Such a model can provide new insights and aid in pinpointing the source of the seemingly spontaneous process. Therefore, guided by the MSM, CsgG can be engineered with the aim of reducing or eliminating the flipping of the eyelet loops to maintain a stable pore geometry for DNA sequencing.

Chapter 7 Conclusions

7.1 Summary

In this thesis, MD was applied to study DNA translocation through nanopores at a molecular level to elucidate the design principles for optimising nanopores for DNA sequencing. Both the dynamics and conformations of the DNA during translocation and its interactions with nanopores were the focus of this work. Comparisons were drawn between the translocation of short and longer ssDNA through simplified model nanopores and the protein nanopores CsgG and the CsgG-CsgF complex.

In chapter 3, MD simulations were performed of the translocation of short flexible ssDNA and a longer ssDNA retained under tension through a series of protein-inspired dual-constriction hydrophobic nanopores. This chapter aimed to investigate the contribution of nanopore geometry and chemistry to conformational behaviour and the rate of DNA translocation. Aromatic residues slowed down the translocation of both short and long ssDNA in the constriction regions of the nanopores. The effect of the interactions with aromatic residues was that the short ssDNA deviated from a linear conformation; however, reducing the width of the nanopore resulted in the DNA retaining a largely extended conformation during translocation. The translocation rate varied greatly for longer tensioned ssDNA, as it depended on whether the strand moved near the aromatic residues in the constriction region due to its inability to coil like the short ssDNA. Lastly, although the pore width correlated to ssDNA entry in the nanopores, it was not found to impact the rate of DNA translocation.

In chapter 4, the proteins CsgG and the CsgG-CsgF complex and the translocation of short ssDNA through them were extensively characterised. The conformational behaviour of CsgG when uncomplexed and in the CsgG-CsgF complex was found to differ under an applied electric field. CsgF stabilised the conformation of both the eyelet loops and the CsgG β-barrel region. In the absence of CsgF, the eyelet loops forming the CsgG constriction region were observed to ‘flip’ upwards into the vestibule to varying degrees both in the absence and presence of ssDNA, which perturbed the pore geometry and resulted in considerable variability in the CsgG channel conductance. In higher electric field strengths, whilst uncomplexed CsgG was unstable, CsgF formed an extensive network of hydrogen bonds and electrostatic interactions with the CsgG β-barrel, resulting in the CsgG-CsgF complex remaining stable under the conditions.
Steered MD simulations, in which short polyA and polyC ssDNA were pulled through CsgG and the CsgG-CsgF complex, revealed that DNA translocation was slowed down due to the strand interacting with residues in the CsgG eyelet loop region and CsgF constriction region. DNA formed

hydrogen bonds with Asn-55 and pi-stacking interactions with Phe-56 and Tyr-51 residues in the CsgG eyelet loop region, and transient hydrogen bonds with Asn-17 residues in the CsgF constriction, which resulted in the progressive movement of the strand through the pores. However, DNA interacted with residues in the eyelet loop region more frequently than in the CsgF constriction. Consequently, the rate of DNA translocation was primarily influenced by the CsgG eyelet loop region in both uncomplexed CsgG and the CsgG-CsgF complex. Although the second constriction formed by CsgF had a minor effect on the DNA translocation rate, it provided several notable advantages for DNA sequencing. Firstly, DNA was retained in a more linear conformation during translocation through the CsgG-CsgF complex compared to uncomplexed CsgG. CsgF slotted inside the CsgG β-barrel to form a hydrophobic channel with a dual-constriction geometry, like the nanopores studied in chapter 3 that also maintained the extended conformation of DNA during translocation. Secondly, the change in channel conductance upon threading ssDNA through the pore was more distinct through the CsgG-CsgF complex than uncomplexed CsgG. Lastly, the CsgG-CsgF complex channel conductance was sensitive to the size of the bases adenine (A) and cytosine (C).

In chapter 5, a more in-depth study of DNA translocation through CsgG and the CsgG-CsgF complex was carried out. To emulate experimental conditions, the translocation of longer polyA and polyC ssDNA was investigated under an applied electric field. DNA translocation was primarily influenced by the CsgG eyelet loop region under an applied electric field, as was observed during the translocation of short ssDNA in chapter 4. However, unlike in steered MD simulations, CsgG eyelet loops were mobile, and there was considerable variability in their motions in both uncomplexed CsgG and the CsgG-CsgF complex. The source of the eyelet loop mobility was unclear; although the eyelet loops during polyC ssDNA translocation were more mobile compared to polyA ssDNA translocation in uncomplexed CsgG, the opposite was observed in the CsgG-CsgF complex in which the eyelet loops remained immobile during polyC ssDNA translocation. The mobility of the eyelet loops gave rise to large fluctuations in the channel conductance during DNA translocation independent of the conformation of the strand inside the pore.

The last chapter focussed on elucidating the mechanism governing the flipping of the CsgG eyelet loops *via* Markov State Model (MSM) analysis. The flipping of the eyelet loops was found to correlate with changes in the torsion angles of the residues near both ends of the loops; hence these residues were used to build an MSM. However, it was not possible to construct an MSM because a Markovian lag time could not be obtained for the MD simulation trajectories. The non-convergence of the implied timescales indicates that longer simulations are required to obtain a Markovian lag time for constructing an MSM.

7.2 Future directions

During the course of this thesis, MD simulations were used to investigate the effects of nanopore geometry and chemistry on the conformational behaviour and rate of DNA translocation. Studies of protein nanopores CsgG and the CsgG-CsgF complex have revealed considerable variability in the dynamics of the CsgG eyelet loops under an applied electric field, both in the presence and absence of DNA. Future work concerns a deeper investigation into the underlying cause of this seemingly spontaneous process *via* MSM analysis, which can be used to identify the key conformational states associated with eyelet loop flipping and compute the probabilities and rates of transitions between states. Longer simulations could be used to generate a model of the dynamics of the eyelet loops in uncomplexed CsgG under an applied electric field. Once a protocol is established, MSMs could be constructed for DNA translocation simulations to assess differences in the dynamics of the eyelet loops in the absence and the presence of DNA. The insights obtained from such MSMs can guide the engineering of CsgG to reduce or eliminate the flipping of the eyelet loops to maintain a stable geometry for DNA sequencing. Furthermore, mutagenesis of the β-barrel region could be undertaken to improve CsgG stability under high electric field strengths.

In nanopore sequencing devices, processive motor enzymes such as DNA helicases [108-110] ratchet ssDNA through the nanopore. Consequently, DNA is under tension inside the nanopore due to the electrophoretic force of the applied voltage pulling the strand downwards whilst it is held by the enzyme. Therefore, further work could include investigating the translocation of long tensioned ssDNA, as done in chapter 3, through uncomplexed CsgG and the CsgG-CsgF complex and comparing DNA behaviour to simulations of non-tensioned ssDNA translocation presented in chapter 5. Further expansion of this study is discussed in chapter 5 (section 5.4).

Regarding DNA translocation, the systems discussed in this thesis could be simulated using the recently developed DES-Amber force field, which has greatly improved descriptions of proteins and ssDNA structures [138]. It is known that the force field employed can influence the properties of the MD simulations, as was observed in chapter 3, in which the water dynamics inside 14LLx2 nanopore differed in simulations using two different force fields; hence simulations using DES-Amber force field would be beneficial for validating the findings presented in this thesis.
As mentioned earlier, another area of further investigation concerns the water dynamics inside the protein-inspired dual-constriction hydrophobic nanopores in chapter 3. Although water dynamics inside 14LLx2 differ according to the force field used, this was not the case for 14Fx2 despite both pores being similar in diameter. Future work could focus on simulating the pore systems in other ionic solutions, in varying applied electric field strengths, and using various water

models [183, 184] to closely examine the impact of pore geometry on water and ion dynamics inside nanometer-sized pores, which is relevant for simulations of narrow protein channels [179].

List of References

1. Kasianowicz, J.J., et al., *Characterization of individual polynucleotide molecules using a membrane channel.* Proceedings of the National Academy of Sciences, 1996. **93**(24): p. 13770-13773.
2. Meller, A., et al., *Rapid nanopore discrimination between single polynucleotide molecules.* Proceedings of the National Academy of Sciences, 2000. **97**(3): p. 1079-1084.
3. 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. 349-353.
4. Stoddart, D., et al., *Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore.* Proceedings of the National Academy of Sciences, 2009. **106**(19): p. 7702-7707.
5. Mohammad, et al., *Controlling a Single Protein in a Nanopore through Electrostatic Traps.* Journal of the American Chemical Society, 2008. **130**(12): p. 4081-4088.
6. Kleefen, A., et al., *Multiplexed Parallel Single Transport Recordings on Nanopore Arrays.* Nano Letters, 2010. **10**(12): p. 5080-5087.
7. Rodriguez-Larrea, D. and H. Bayley, *Multistep protein unfolding during nanopore translocation.* Nature Nanotechnology, 2013. **8**(4): p. 288-295.
8. Ouldali, H., et al., *Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore.* Nature Biotechnology, 2020. **38**(2): p. 176-181.
9. Wang, S., et al., *Engineering of protein nanopores for sequencing, chemical or protein sensing and disease diagnosis.* Current Opinion in Biotechnology, 2018. **51**: p. 80-89.
10. Robertson, J.W.F., et al., *Single-molecule mass spectrometry in solution using a solitary nanopore.* Proceedings of the National Academy of Sciences, 2007. **104**(20): p. 8207-8211.
11. Baaken, G., et al., *Nanopore-based single-molecule mass spectrometry on a lipid membrane microarray.* ACS Nano, 2011. **5**(10): p. 8080-8.
12. Bezrukov, S.M. and J.J. Kasianowicz, *Current noise reveals protonation kinetics and number of ionizable sites in an open protein ion channel.* Physical Review Letters, 1993. **70**(15): p. 2352-2355.
13. Kasianowicz, J.J., et al., *Genetically Engineered Metal Ion Binding Sites on the Outside of a Channel's Transmembrane β-Barrel.* Biophysical Journal, 1999. **76**(2): p. 837-845.
14. Ali, M., et al., *Metal ion affinity-based biomolecular recognition and conjugation inside synthetic polymer nanopores modified with iron-terpyridine complexes.* J Am Chem Soc, 2011. **133**(43): p. 17307-14.
15. Astier, Y., O. Uzun, and F. Stellacci, *Electrophysiological Study of Single Gold Nanoparticle/α-Hemolysin Complex Formation: A Nanotool to Slow Down ssDNA Through the α-Hemolysin Nanopore.* Small, 2009. **5**(11): p. 1273-1278.
16. Angevine, C.E., et al., *Enhanced Single Molecule Mass Spectrometry via Charged Metallic Clusters.* Analytical Chemistry, 2014. **86**(22): p. 11077-11085.
17. Campos, E.J. and J. Yates, *Single molecule characterisation of metal nanoparticles using nanopore-based stochastic detection methods.* Sensors and Actuators B: Chemical, 2018. **255**: p. 2032-2049.
18. Deamer, D., M. Akeson, and D. Branton, *Three decades of nanopore sequencing.* Nature Biotechnology, 2016. **34**(5): p. 518-524.
19. Wanunu, M., et al., *Electrostatic focusing of unlabelled DNA into nanoscale pores using a salt gradient.* Nature nanotechnology, 2010. **5**(2): p. 160-165.
20. Oxford Nanopore Technologies. *Accelerating cancer research through comprehensive genomic analysis*. 2022 [cited 2022 1 September]; Available from: https://nanoporetech.com/sites/default/files/s3/white-papers/Cancer-research-white-paper.pdf.
21. Hu, T., et al., *Detection of Structural Variations and Fusion Genes in Breast Cancer Samples Using Third-Generation Sequencing.* Frontiers in Cell and Developmental Biology, 2022. **10**.

22. Carson, S. and M. Wanunu, *Challenges in DNA motion control and sequence readout using nanopore devices.* Nanotechnology, 2015. **26**(7): p. 074004.
23. Gu, L.-Q., S. Cheley, and H. Bayley, *Electroosmotic enhancement of the binding of a neutral molecule to a transmembrane pore.* Proceedings of the National Academy of Sciences, 2003. **100**(26): p. 15498-15503.
24. Piguet, F., et al., *Electroosmosis through α-hemolysin that depends on alkali cation type.* The Journal of Physical Chemistry Letters, 2014. **5**(24): p. 4362-4367.
25. Talaga, D.S. and J. Li, *Single-molecule protein unfolding in solid state nanopores.* Journal of the American Chemical Society, 2009. **131**(26): p. 9287-9297.
26. Li, M.-Y., et al., *Revisiting the Origin of Nanopore Current Blockage for Volume Difference Sensing at the Atomic Level.* JACS Au, 2021. **1**(7): p. 967-976.
27. Bayley, H. and P.S. Cremer, *Stochastic sensors inspired by biology.* Nature, 2001. **413**(6852): p. 226-30.
28. Gu, L.-Q., et al., *Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adapter.* Nature, 1999. **398**(6729): p. 686-690.
29. Heerema, S.J. and C. Dekker, *Graphene nanodevices for DNA sequencing.* Nature Nanotechnology, 2016. **11**(2): p. 127-136.
30. Dekker, C., *Solid-state nanopores*, in *Nanoscience and Technology*. p. 60-66.
31. He, Y., et al., *Solid-state nanopore systems: from materials to applications.* NPG Asia Materials, 2021. **13**(1): p. 48.
32. Xue, L., et al., *Solid-state nanopore sensors.* Nature Reviews Materials, 2020. **5**(12): p. 931-951.
33. Cao, C., et al., *Single-molecule sensing of peptides and nucleic acids by engineered aerolysin nanopores.* Nature Communications, 2019. **10**(1): p. 4918.
34. Fragasso, A., S. Schmid, and C. Dekker, *Comparing Current Noise in Biological and Solid-State Nanopores.* ACS Nano, 2020. **14**(2): p. 1338-1349.
35. Thomas, P.D. and A. Kejariwal, *Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects.* Proceedings of the National Academy of Sciences, 2004. **101**(43): p. 15398-15403.
36. Katsman, E., et al., *Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing.* Genome Biology, 2022. **23**(1): p. 158.
37. Cumbo, C., et al., *Genomic BCR-ABL1 breakpoint characterization by a multi-strategy approach for "personalized monitoring" of residual disease in chronic myeloid leukemia patients.* Oncotarget, 2018. **9**(13).
38. Hamburg, M.A. and F.S. Collins, *The Path to Personalized Medicine.* New England Journal of Medicine, 2010. **363**(4): p. 301-304.
39. Liu, N., et al., *Two-Way Nanopore Sensing of Sequence-Specific Oligonucleotides and Small-Molecule Targets in Complex Matrices Using Integrated DNA Supersandwich Structures.* Angewandte Chemie International Edition, 2013. **52**(7): p. 2007-2011.
40. Lu, Y., et al., *Simultaneous single-molecule discrimination of cysteine and homocysteine with a protein nanopore.* Chemical Communications, 2019. **55**(63): p. 9311-9314.
41. Vercoutere, W., et al., *Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel.* Nature Biotechnology, 2001. **19**(3): p. 248-252.
42. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads.* Nature Biotechnology, 2018. **36**(4): p. 338-345.
43. Brown, C.G. and J. Clarke, *Nanopore development at Oxford Nanopore.* Nat Biotechnol, 2016. **34**(8): p. 810-1.
44. Jain, M., et al., *The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.* Genome Biology, 2016. **17**(1): p. 239.
45. Kono, N. and K. Arakawa, *Nanopore sequencing: Review of potential applications in functional genomics.* Development, Growth & Differentiation, 2019. **61**(5): p. 316-326.
46. Song, L., et al., *Structure of Staphylococcal α-Hemolysin, a Heptameric Transmembrane Pore.* Science, 1996. **274**(5294): p. 1859-1865.

47. Stoddart, D., et al., *Nucleobase Recognition in ssDNA at the Central Constriction of the α-Hemolysin Pore.* Nano Letters, 2010. **10**(9): p. 3633-3637.
48. Nakane, J., M. Wiggin, and A. Marziali, *A nanosensor for transmembrane capture and identification of single nucleic acid molecules.* Biophysical Journal, 2004. **87**(1): p. 615-621.
49. Bates, M., M. Burns, and A. Meller, *Dynamics of DNA molecules in a membrane channel probed by active control techniques.* Biophysical Journal, 2003. **84**(4): p. 2366-2372.
50. 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.
51. 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.* Biophys J, 1999. **77**(6): p. 3227-33.
52. Ayub, M. and H. Bayley, *Individual RNA base recognition in immobilized oligonucleotides using a protein nanopore.* Nano letters, 2012. **12**(11): p. 5637-5643.
53. Cracknell, J.A., D. Japrung, and H. Bayley, *Translocating kilobase RNA through the staphylococcal α-hemolysin nanopore.* Nano letters, 2013. **13**(6): p. 2500-2505.
54. Vercoutere, W.A., et al., *Discrimination among individual Watson–Crick base pairs at the termini of single DNA hairpin molecules.* Nucleic acids research, 2003. **31**(4): p. 1311-1318.
55. Sutherland, T.C., et al., *Structure of peptides investigated by nanopore analysis.* Nano letters, 2004. **4**(7): p. 1273-1277.
56. Stefureac, R., et al., *Transport of α-helical peptides through α-hemolysin and aerolysin pores.* Biochemistry, 2006. **45**(30): p. 9172-9179.
57. Nivala, J., D.B. Marks, and M. Akeson, *Unfoldase-mediated protein translocation through an α-hemolysin nanopore.* Nature biotechnology, 2013. **31**(3): p. 247-250.
58. Meller, A. and D. Branton, *Single molecule measurements of DNA transport through a nanopore.* Electrophoresis, 2002. **23**(16): p. 2583-91.
59. Faller, M., M. Niederweis, and G.E. Schulz, *The structure of a mycobacterial outer-membrane channel.* Science, 2004. **303**(5661): p. 1189-1192.
60. Butler, T.Z., et al., *Single-molecule DNA detection with an engineered MspA protein nanopore.* Proceedings of the National Academy of Sciences, 2008. **105**(52): p. 20647-20652.
61. Derrington, I.M., et al., *Nanopore DNA sequencing with MspA.* Proceedings of the National Academy of Sciences, 2010. **107**(37): p. 16060-16065.
62. Iacovache, I., et al., *Cryo-EM structure of aerolysin variants reveals a novel protein fold and the pore-formation process.* Nature Communications, 2016. **7**(1): p. 12062.
63. Degiacomi, M.T., et al., *Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism.* Nature chemical biology, 2013. **9**(10): p. 623-629.
64. Cao, C., et al., *Mapping the sensing spots of aerolysin for single oligonucleotides analysis.* Nature communications, 2018. **9**(1): p. 1-9.
65. Iacovache, I., et al., *Cryo-EM structure of aerolysin variants reveals a novel protein fold and the pore-formation process.* Nature communications, 2016. **7**(1): p. 1-8.
66. Cao, C., et al., *Driven translocation of polynucleotides through an aerolysin nanopore.* Analytical chemistry, 2016. **88**(10): p. 5046-5049.
67. Cao, C., et al., *Discrimination of oligonucleotides of different lengths with a wild-type aerolysin nanopore.* Nature nanotechnology, 2016. **11**(8): p. 713-718.
68. Piguet, F., et al., *Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore.* Nature communications, 2018. **9**(1): p. 1-13.
69. Pastoriza-Gallego, M., et al., *Dynamics of unfolded protein transport through an aerolysin pore.* Journal of the American Chemical Society, 2011. **133**(9): p. 2923-2931.
70. Cao, C., et al., *Mapping the sensing spots of aerolysin for single oligonucleotides analysis.* Nature Communications, 2018. **9**(1): p. 2823.
71. 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.

72. Ayub, M., D. Stoddart, and H. Bayley, *Nucleobase recognition by truncated α-hemolysin pores.* ACS nano, 2015. **9**(8): p. 7895-7903.
73. Loferer, H., M. Hammar, and S. Normark, *Availability of the fibre subunit CsgA and the nucleator protein CsgB during assembly of fibronectin-binding curli is limited by the intracellular concentration of the novel lipoprotein CsgG.* Mol Microbiol, 1997. **26**(1): p. 11-23.
74. Goyal, P., et al., *Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG.* Nature, 2014. **516**(7530): p. 250-3.
75. Cao, B., et al., *Structure of the nonameric bacterial amyloid secretion channel.* Proceedings of the National Academy of Sciences, 2014. **111**(50): p. E5439-E5444.
76. Meller, A., L. Nivon, and D. Branton, *Voltage-driven DNA translocations through a nanopore.* Physical Review Letters, 2001. **86**(15): p. 3435.
77. Zhou, W., et al., *Molecular insights into distinct detection properties of α-hemolysin, MspA, CsgG, and aerolysin nanopore sensors.* The Journal of Physical Chemistry B, 2020. **124**(9): p. 1611-1618.
78. Heng, J.B., et al., *Sizing DNA Using a Nanometer-Diameter Pore.* Biophysical Journal, 2004. **87**(4): p. 2905-2911.
79. Aksimentiev, A., et al., *Microscopic Kinetics of DNA Translocation through Synthetic Nanopores.* Biophysical Journal, 2004. **87**(3): p. 2086-2097.
80. Aksimentiev, A. and K. Schulten, *Imaging α-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map.* Biophysical Journal, 2005. **88**(6): p. 3745-3761.
81. Guy, Andrew T., Thomas 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.
82. Brinkerhoff, H., et al., *Multiple rereads of single proteins at single–amino acid resolution using nanopores.* Science. **0**(0): p. eabl4381.
83. Gumbart, J., et al., *Constant electric field simulations of the membrane potential illustrated with simple systems.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2012. **1818**(2): p. 294-302.
84. Sotomayor, M., et al., *Ion Conduction through MscS as Determined by Electrophysiology and Simulation.* Biophysical Journal, 2007. **92**(3): p. 886-902.
85. Khalili-Araghi, F., E. Tajkhorshid, and K. Schulten, *Dynamics of K+ Ion Conduction through Kv1.2.* Biophysical Journal, 2006. **91**(6): p. L72-L74.
86. Pezeshki, S., et al., *Understanding Ion Conductance on a Molecular Level: An All-Atom Modeling of the Bacterial Porin OmpF.* Biophysical Journal, 2009. **97**(7): p. 1898-1906.
87. Chimerel, C., et al., *Transport at the nanoscale: temperature dependence of ion conductance.* European Biophysics Journal, 2008. **38**(1): p. 121-125.
88. 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.
89. 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.
90. Manara, R.M.A., S. Tomasio, and S. Khalid, *The Nucleotide Capture Region of Alpha Hemolysin: Insights into Nanopore Design for DNA Sequencing from Molecular Dynamics Simulations.* Nanomaterials, 2015. **5**: p. 144-153.
91. Mathé, J., et al., *Orientation discrimination of single-stranded DNA inside the α-hemolysin membrane channel.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(35): p. 12377.
92. Yu, M., et al., *Unveiling the Microscopic Mechanism of Current Variation in the Sensing Region of the MspA Nanopore for DNA Sequencing.* The Journal of Physical Chemistry Letters, 2021. **12**(37): p. 9132-9141.

93. Wells, D.B., V. Abramkina, and A. Aksimentiev, *Exploring transmembrane transport through α-hemolysin with grid-steered molecular dynamics.* The Journal of Chemical Physics, 2007. **127**(12): p. 125101.
94. Martin, H.S.C., et al., *Determination of Free Energy Profiles for the Translocation of Polynucleotides through α-Hemolysin Nanopores using Non-Equilibrium Molecular Dynamics Simulations.* Journal of Chemical Theory and Computation, 2009. **5**(8): p. 2135-2148.
95. De Biase, P.M., S. Markosyan, and S. Noskov, *Microsecond simulations of DNA and ion transport in nanopores with novel ion–ion and ion–nucleotides effective potentials.* Journal of Computational Chemistry, 2014. **35**(9): p. 711-721.
96. Matysiak, S., et al., *Dynamics of Polymer Translocation through Nanopores: Theory Meets Experiment.* Physical Review Letters, 2006. **96**(11): p. 118103.
97. Ramachandran, A., et al., *Coarse-Grained Molecular Dynamics Simulation of DNA Translocation in Chemically Modified Nanopores.* The Journal of Physical Chemistry B, 2011. **115**(19): p. 6138-6148.
98. Meller, A., L. Nivon, and D. Branton, *Voltage-Driven DNA Translocations through a Nanopore.* Physical Review Letters, 2001. **86**(15): p. 3435-3438.
99. Kawano, R., et al., *Controlling the Translocation of Single-Stranded DNA through α-Hemolysin Ion Channels Using Viscosity.* Langmuir, 2009. **25**(2): p. 1233-1237.
100. de Zoysa, R.S.S., et al., *Slowing DNA Translocation through Nanopores Using a Solution Containing Organic Salts.* The Journal of Physical Chemistry B, 2009. **113**(40): p. 13332-13336.
101. Howorka, S., S. Cheley, and H. Bayley, *Sequence-specific detection of individual DNA strands using engineered nanopores.* Nat Biotechnol, 2001. **19**(7): p. 636-9.
102. Clarke, J., et al., *Continuous base identification for single-molecule nanopore DNA sequencing.* Nat Nanotechnol, 2009. **4**(4): p. 265-70.
103. Noakes, M.T., et al., *Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage.* Nature biotechnology, 2019. **37**(6): p. 651-656.
104. Nova, I.C., et al., *Investigating asymmetric salt profiles for nanopore DNA sequencing with biological porin MspA.* PLoS One, 2017. **12**(7): p. e0181599.
105. Lieberman, K.R., et al., *Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase.* Journal of the American Chemical Society, 2010. **132**(50): p. 17961-17972.
106. Gyarfas, B., et al., *Mapping the position of DNA polymerase-bound DNA templates in a nanopore at 5 Å resolution.* ACS nano, 2009. **3**(6): p. 1457-1466.
107. Cherf, G.M., et al., *Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision.* Nature biotechnology, 2012. **30**(4): p. 344-348.
108. Derrington, I.M., et al., *Subangstrom single-molecule measurements of motor proteins using a nanopore.* Nature Biotechnology, 2015. **33**(10): p. 1073-1075.
109. Noakes, M.T., et al., *Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage.* Nature Biotechnology, 2019. **37**(6): p. 651-656.
110. Craig, J.M., et al., *Determining the effects of DNA sequence on Hel308 helicase translocation along single-stranded DNA using nanopore tweezers.* Nucleic Acids Research, 2019. **47**(5): p. 2506-2513.
111. Ibarra, B., et al., *Proofreading dynamics of a processive DNA polymerase.* The EMBO journal, 2009. **28**(18): p. 2794-2802.
112. Saecker, R.M. and M.T. Record Jr, *Protein surface salt bridges and paths for DNA wrapping.* Current opinion in structural biology, 2002. **12**(3): p. 311-319.
113. Caldwell, C.C. and M. Spies, *Helicase SPRNTing through the nanopore.* Proceedings of the National Academy of Sciences, 2017. **114**(45): p. 11809-11811.
114. Branton, D. and D.W. Deamer, *Nanopore Sequencing: An Introduction*. 2019: World Scientific Publishing Company.
115. Cressiot, B., et al., *Aerolysin, a powerful protein sensor for fundamental studies and development of upcoming applications.* ACS sensors, 2019. **4**(3): p. 530-548.

116. Reiner, J.E., et al., *Theory for polymer analysis using nanopore-based single-molecule mass spectrometry.* Proceedings of the National Academy of Sciences, 2010. **107**(27): p. 12080-12085.
117. Kim, S.C., et al., *Geometric dependence of the conductance drop in a nanopore due to a particle.* Physical Review E, 2014. **89**(4): p. 042702.
118. Si, W. and A. Aksimentiev, *Nanopore Sensing of Protein Folding.* ACS Nano, 2017. **11**(7): p. 7091-7100.
119. Wilson, J., et al., *Rapid and Accurate Determination of Nanopore Ionic Current Using a Steric Exclusion Model.* ACS Sensors, 2019. **4**(3): p. 634-644.
120. Chinappi, M. and F. Cecconi, *Protein sequencing via nanopore based devices: a nanofluidics perspective.* Journal of Physics: Condensed Matter, 2018. **30**(20): p. 204002.
121. Maglia, G., et al., *Enhanced translocation of single DNA molecules through α-hemolysin nanopores by manipulation of internal charge.* Proceedings of the National Academy of Sciences, 2008. **105**(50): p. 19720-19725.
122. Haynes, T., et al., *Electric-Field-Driven Translocation of ssDNA through Hydrophobic Nanopores.* ACS Nano, 2018. **12**(8): p. 8208-8213.
123. Van der Verren, S.E., et al., *A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity.* Nature Biotechnology, 2020. **38**(12): p. 1415-1420.
124. Yan, Z., et al., *Assembly and substrate recognition of curli biogenesis system.* Nature Communications, 2020. **11**(1): p. 241.
125. Zhang, M., et al., *Cryo-EM structure of the nonameric CsgG-CsgF complex and its implications for controlling curli biogenesis in Enterobacteriaceae.* PLOS Biology, 2020. **18**(6): p. e3000748.
126. Karplus, M., *Molecular Dynamics Simulations of Biomolecules.* Accounts of Chemical Research, 2002. **35**(6): p. 321-323.
127. Andersen, H.C., *Molecular dynamics simulations at constant pressure and/or temperature.* The Journal of Chemical Physics, 1980. **72**(4): p. 2384-2393.
128. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
129. Varadi, M., et al., *AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.* Nucleic Acids Research, 2021. **50**(D1): p. D439-D444.
130. Lindahl, A., Hess, van der Spoel, *GROMACS 2021 Manual.* Zenedo, 2021.
131. Hockney, R.W., S.P. Goel, and J.W. Eastwood, *Quiet high-resolution computer models of a plasma.* Journal of Computational Physics, 1974. **14**(2): p. 148-158.
132. Leach, A.R., *Molecular Modelling: Principles and Applications*. 2001.
133. Ryckaert, J.-P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.
134. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry, 1997. **18**(12): p. 1463-1472.
135. Glass, W.G., et al., *Coarse-Grained Molecular Dynamics Simulations of Membrane Proteins: A Practical Guide*, in *Structure and Function of Membrane Proteins*, I. Schmidt-Krey and J.C. Gumbart, Editors. 2021, Springer US: New York, NY. p. 253-273.
136. 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.
137. Huang, J., et al., *CHARMM36m: an improved force field for folded and intrinsically disordered proteins.* Nature Methods, 2017. **14**(1): p. 71-73.
138. Tucker, M.R., et al., *Development of Force Field Parameters for the Simulation of Single- and Double-Stranded DNA Molecules and DNA–Protein Complexes.* The Journal of Physical Chemistry B, 2022. **126**(24): p. 4442-4457.

139. Jones, J.E., *On the determination of molecular fields.—II. From the equation of state of a gas.* Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 1924. **106**(738): p. 463-477.
140. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
141. Essmann, U., et al., *A smooth particle mesh Ewald method.* The Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.
142. Ashley, S., *Ergodic theory plays a key role in multiple fields.* Proceedings of the National Academy of Sciences, 2015. **112**(7): p. 1914-1914.
143. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath.* The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
144. Bussi, G., D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling.* The Journal of Chemical Physics, 2007. **126**(1): p. 014101.
145. 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.
146. Jarzynski, C., *Nonequilibrium Equality for Free Energy Differences.* Physical Review Letters, 1997. **78**(14): p. 2690-2693.
147. Lu, H. and K. Schulten, *Steered molecular dynamics simulations of force-induced protein domain unfolding.* Proteins: Structure, Function, and Bioinformatics, 1999. **35**(4): p. 453-463.
148. Bavi, N., et al., *Nanomechanical properties of MscL α helices: A steered molecular dynamics study.* Channels, 2017. **11**(3): p. 209-223.
149. Guzmán, D.L., et al., *Using steered molecular dynamics simulations and single-molecule force spectroscopy to guide the rational design of biomimetic modular polymeric materials.* Polymer, 2008. **49**(18): p. 3892-3901.
150. Maglia, G., et al., *Enhanced translocation of single DNA molecules through α-hemolysin nanopores by manipulation of internal charge.* Proceedings of the National Academy of Sciences, 2008. **105**(50): p. 19720.
151. Stoddart, D., et al., *Multiple base-recognition sites in a biological nanopore: two heads are better than one.* Angew Chem Int Ed Engl, 2010. **49**(3): p. 556-9.
152. Bentin, J., S. Balme, and F. Picaud, *Polynucleotide differentiation using hybrid solid-state nanopore functionalizing with α-hemolysin.* Soft Matter, 2020. **16**(4): p. 1002-1010.
153. Rattu, P., et al., *Translocation of flexible and tensioned ssDNA through in silico designed hydrophobic nanopores with two constrictions.* Nanoscale, 2021. **13**(3): p. 1673-1679.
154. Trick, J.L., et al., *Designing a Hydrophobic Barrier within Biomimetic Nanopores.* ACS Nano, 2014. **8**(11): p. 11268-11279.
155. Sánchez, R. and A. Sali, *Comparative protein structure modeling. Introduction and practical examples with modeller.* Methods Mol Biol, 2000. **143**: p. 97-129.
156. Schrödinger, L., *PyMOL Molecular Graphics System, version 2.3*. 2019, Schrödinger, LLC.
157. Lu, X.J. and W.K. Olson, *3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.* Nucleic Acids Research, 2003. **31**(17): p. 5108-5121.
158. Berweger, C.D., W.F. van Gunsteren, and F. Müller-Plathe, *Force field parametrization by weak coupling. Re-engineering SPC water.* Chemical Physics Letters, 1995. **232**(5): p. 429-436.
159. Páll, S., et al. *Tackling exascale software challenges in molecular dynamics simulations with GROMACS*. in *International conference on exascale applications and software*. 2014. Springer.
160. Oostenbrink, C., et al., *Validation of the 53A6 GROMOS force field.* European Biophysics Journal, 2005. **34**(4): p. 273-284.
161. Berger, O.E., O; Jahnig F, *Molecular Dynamics Simulations of a Fluid Bilayer of Dipalmitoylphosphatidylcholine at Full Hydration, Constant Pressure, and Constant Temperature.* Biophysical Journal, 2022. **72**: p. 2002-2013.

162. Manara, R.M.A., E. Jayne Wallace, and S. Khalid, *DNA sequencing with MspA: Molecular Dynamics simulations reveal free-energy differences between sequencing and non-sequencing mutants.* Scientific Reports, 2015. **5**(1): p. 12783.
163. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
164. Smart, O.S., et al., *HOLE: a program for the analysis of the pore dimensions of ion channel structural models.* J Mol Graph, 1996. **14**(6): p. 354-60, 376.
165. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-38.
166. Skylaris, C.-K., et al., *Introducing ONETEP: Linear-scaling density functional simulations on parallel computers.* The Journal of Chemical Physics, 2005. **122**(8): p. 084119.
167. Jo, S., et al., *CHARMM-GUI: a web-based graphical user interface for CHARMM.* J Comput Chem, 2008. **29**(11): p. 1859-65.
168. Jo, S., et al., *Chapter Eight - CHARMM-GUI PDB Manipulator for Advanced Modeling and Simulations of Proteins Containing Nonstandard Residues*, in *Advances in Protein Chemistry and Structural Biology*, T. Karabencheva-Christova, Editor. 2014, Academic Press. p. 235-265.
169. Prentice, J.C.A., et al., *The ONETEP linear-scaling density functional theory program.* The Journal of Chemical Physics, 2020. **152**(17): p. 174111.
170. Perdew, J.P., K. Burke, and M. Ernzerhof, *Generalized Gradient Approximation Made Simple.* Physical Review Letters, 1996. **77**(18): p. 3865-3868.
171. Antony, J. and S. Grimme, *Density functional theory including dispersion corrections for intermolecular interactions in a large benchmark set of biologically relevant molecules.* Physical Chemistry Chemical Physics, 2006. **8**(45): p. 5287-5293.
172. Skylaris, C.-K., et al., *Nonorthogonal generalized Wannier function pseudopotential plane-wave method.* Physical Review B, 2002. **66**(3): p. 035119.
173. Bell, N.A.W., et al., *Asymmetric dynamics of DNA entering and exiting a strongly confining nanopore.* Nature Communications, 2017. **8**(1): p. 380.
174. Beckstein, O., P.C. Biggin, and M.S.P. Sansom, *A Hydrophobic Gating Mechanism for Nanopores.* The Journal of Physical Chemistry B, 2001. **105**(51): p. 12902-12905.
175. Allen, R., S. Melchionna, and J.-P. Hansen, *Permeation of nanopores by water: the effects of channel polarization.* Journal of Physics: Condensed Matter, 2002. **15**(1): p. S297-S302.
176. Beckstein, O. and M.S.P. Sansom, *The influence of geometry, surface character, and flexibility on the permeation of ions and water through biological pores.* Physical Biology, 2004. **1**(1): p. 42-52.
177. Hummer, G., J.C. Rasaiah, and J.P. Noworyta, *Water conduction through the hydrophobic channel of a carbon nanotube.* Nature, 2001. **414**(6860): p. 188-190.
178. Rao, S., et al., *Molecular Simulations of Hydrophobic Gating of Pentameric Ligand Gated Ion Channels: Insights into Water and Ions.* The Journal of Physical Chemistry B, 2021. **125**(4): p. 981-994.
179. Klesse, G., et al., *Induced Polarization in Molecular Dynamics Simulations of the 5-HT3 Receptor Channel.* Journal of the American Chemical Society, 2020. **142**(20): p. 9415-9427.
180. Beckstein, O. and M.S.P. Sansom, *Liquid-vapor oscillations of water in hydrophobic nanopores.* Proceedings of the National Academy of Sciences, 2003. **100**(12): p. 7063-7068.
181. Trick, J.L., et al., *Voltage Gating of a Biomimetic Nanopore: Electrowetting of a Hydrophobic Barrier.* ACS Nano, 2017. **11**(2): p. 1840-1847.
182. Powell, M.R., et al., *Electric-field-induced wetting and dewetting in single hydrophobic nanopores.* Nature Nanotechnology, 2011. **6**(12): p. 798-802.
183. K, V.P., et al., *Water desalination using graphene nanopores: influence of the water models used in simulations.* Physical Chemistry Chemical Physics, 2018. **20**(23): p. 16005-16011.
184. Acharya, A., J.D. Prajapati, and U. Kleinekathöfer, *Atomistic Simulation of Molecules Interacting with Biological Nanopores: From Current Understanding to Future Directions.* The Journal of Physical Chemistry B, 2022. **126**(22): p. 3995-4008.

185. Chapman, M.R., et al., *Role of Escherichia coli Curli Operons in Directing Amyloid Fiber Formation.* Science, 2002. **295**(5556): p. 851-855.
186. Nenninger, A.A., L.S. Robinson, and S.J. Hultgren, *Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF.* Proceedings of the National Academy of Sciences, 2009. **106**(3): p. 900-905.
187. Nenninger, A.A., et al., *CsgE is a curli secretion specificity factor that prevents amyloid fibre aggregation.* Mol Microbiol, 2011. **81**(2): p. 486-99.
188. Rattu, P., et al., *Atomistic level characterisation of ssDNA translocation through the E. coli proteins CsgG and CsgF for nanopore sequencing.* Computational and Structural Biotechnology Journal, 2021. **19**: p. 6417-6430.
189. Emsley, P., et al., *Features and development of Coot.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 486-501.
190. Wu, E.L., et al., *CHARMM-GUI Membrane Builder toward realistic biological membrane simulations.* J Comput Chem, 2014. **35**(27): p. 1997-2004.
191. Jo, S., T. Kim, and W. Im, *Automated Builder and Database of Protein/Membrane Complexes for Molecular Dynamics Simulations.* PLOS ONE, 2007. **2**(9): p. e880.
192. Lee, J., et al., *CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field.* Journal of chemical theory and computation, 2016. **12**(1): p. 405-413.
193. Martin, H.S.C., S. Jha, and P.V. Coveney, *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.
194. Huang, C., et al., *Molecular insights into geometric and electrophoretic effects on DNA translocation speed through graphene nanoslit sensor.* Carbon, 2022. **191**: p. 415-423.
195. Si, W., et al., *Controllable and reversible DNA translocation through a single-layer molybdenum disulfide nanopore.* Nanoscale, 2018. **10**(41): p. 19450-19458.
196. Huang, J., et al., *CHARMM36m: an improved force field for folded and intrinsically disordered proteins.* Nat Methods, 2017. **14**(1): p. 71-73.
197. Wimley, W.C. and S.H. White, *Experimentally determined hydrophobicity scale for proteins at membrane interfaces.* Nature Structural Biology, 1996. **3**(10): p. 842-848.
198. Purnell, R.F., K.K. Mehta, and J.J. Schmidt, *Nucleotide Identification and Orientation Discrimination of DNA Homopolymers Immobilized in a Protein Nanopore.* Nano Letters, 2008. **8**(9): p. 3029-3034.
199. Stoddart, D., et al., *Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore.* Nano Lett, 2010. **10**(9): p. 3633-7.
200. Picaud, F., et al., *Discrimination of Polynucleotide Transport through a Highly Hydrophobic Uncharged Nanopore.* The Journal of Physical Chemistry C, 2017. **121**(13): p. 7525-7532.
201. Manrao, E.A., et al., *Nucleotide Discrimination with DNA Immobilized in the MspA Nanopore.* PLOS ONE, 2011. **6**(10): p. e25723.
202. English, N.J. and C.J. Waldron, *Perspectives on external electric fields in molecular simulation: progress, prospects and challenges.* Physical Chemistry Chemical Physics, 2015. **17**(19): p. 12407-12440.
203. Liu, Y. and S.M. Iqbal, *A mesoscale model of DNA interaction with functionalized nanopore.* Applied Physics Letters, 2009. **95**(22): p. 223701.
204. Chen, M., et al., *Outer membrane protein G: Engineering a quiet pore for biosensing.* Proceedings of the National Academy of Sciences, 2008. **105**(17): p. 6272-6277.
205. Schütte, C., et al., *A direct approach to conformational dynamics based on hybrid Monte Carlo.* Journal of Computational Physics, 1999. **151**(1): p. 146-168.
206. Swope, W.C., J.W. Pitera, and F. Suits, *Describing protein folding kinetics by molecular dynamics simulations. 1. Theory.* The Journal of Physical Chemistry B, 2004. **108**(21): p. 6571-6581.
207. Chekmarev, D.S., T. Ishida, and R.M. Levy, *Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models.* The Journal of Physical Chemistry B, 2004. **108**(50): p. 19487-19495.

208. Singhal, N. and V.S. Pande, *Error analysis and efficient sampling in Markovian state models for molecular dynamics.* The Journal of chemical physics, 2005. **123**(20): p. 204909.
209. Sriraman, S., I.G. Kevrekidis, and G. Hummer, *Coarse master equation from Bayesian analysis of replica molecular dynamics simulations.* The Journal of Physical Chemistry B, 2005. **109**(14): p. 6479-6484.
210. Noé, F., et al., *Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states.* The Journal of chemical physics, 2007. **126**(15): p. 04B617.
211. Chodera, J.D., et al., *Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics.* The Journal of chemical physics, 2007. **126**(15): p. 04B616.
212. Noé, F., *Probability distributions of molecular observables computed from Markov models.* The Journal of chemical physics, 2008. **128**(24): p. 244103.
213. Buchete, N.-V. and G. Hummer, *Coarse master equations for peptide folding dynamics.* The Journal of Physical Chemistry B, 2008. **112**(19): p. 6057-6069.
214. Pan, A.C. and B. Roux, *Building Markov state models along pathways to determine free energies and rates of transitions.* The Journal of chemical physics, 2008. **129**(6): p. 064107.
215. Mercadante, D., F. Gräter, and C. Daday, *CONAN: A Tool to Decode Dynamical Information from Molecular Interaction Maps.* Biophysical Journal, 2018. **114**(6): p. 1267-1273.
216. Prinz, J.-H., et al., *Markov models of molecular kinetics: Generation and validation.* The Journal of Chemical Physics, 2011. **134**(17): p. 174105.
217. Ferrario, V., et al., *Modelling of substrate access and substrate binding to cephalosporin acylases.* Scientific reports, 2019. **9**(1): p. 1-13.
218. Noé, F. and S. Fischer, *Transition networks for modeling the kinetics of conformational change in macromolecules.* Current opinion in structural biology, 2008. **18**(2): p. 154-162.
219. Pande, V.S., K. Beauchamp, and G.R. Bowman, *Everything you wanted to know about Markov State Models but were afraid to ask.* Methods, 2010. **52**(1): p. 99-105.
220. Noé, F., et al., *Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations.* Proceedings of the National Academy of Sciences, 2009. **106**(45): p. 19011-19016.
221. Bowman, G.R., et al., *Progress and challenges in the automated construction of Markov state models for full protein systems.* The Journal of chemical physics, 2009. **131**(12): p. 124101.
222. Voelz, V.A., et al., *Molecular simulation of ab initio protein folding for a millisecond folder NTL9 (1− 39).* Journal of the American Chemical Society, 2010. **132**(5): p. 1526-1528.
223. Bowman, G.R. and V.S. Pande, *Protein folded states are kinetic hubs.* Proceedings of the National Academy of Sciences, 2010. **107**(24): p. 10890-10895.
224. Voelz, V.A., et al., *Markov state models of millisecond folder ACBP reveals new views of the folding reaction.* Biophysical Journal, 2011. **100**(3): p. 515a.
225. Beauchamp, K.A., et al., *Simple few-state models reveal hidden complexity in protein folding.* Proceedings of the National Academy of Sciences, 2012. **109**(44): p. 17807-17813.
226. Sborgi, L., et al., *Interaction networks in protein folding via atomic-resolution experiments and long-time-scale molecular dynamics simulations.* Journal of the American Chemical Society, 2015. **137**(20): p. 6506-6516.
227. Held, M., et al., *Mechanisms of protein-ligand association and its modulation by protein mutations.* Biophysical journal, 2011. **100**(3): p. 701-710.
228. Silva, D.-A., et al., *A role for both conformational selection and induced fit in ligand binding by the LAO protein.* PLoS computational biology, 2011. **7**(5): p. e1002054.
229. Buch, I., T. Giorgino, and G. De Fabritiis, *Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations.* Proceedings of the National Academy of Sciences, 2011. **108**(25): p. 10184-10189.
230. Gu, S., et al., *Quantitatively characterizing the ligand binding mechanisms of choline binding protein using Markov state model analysis.* PLoS computational biology, 2014. **10**(8): p. e1003767.

231. Plattner, N. and F. Noé, *Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models.* Nature communications, 2015. **6**(1): p. 1-10.
232. Jiang, H., et al., *Markov state models reveal a two-step mechanism of miRNA loading into the human argonaute protein: selective binding followed by structural re-arrangement.* PLoS computational biology, 2015. **11**(7): p. e1004404.
233. Peng, S., et al., *Target search and recognition mechanisms of glycosylase AlkD revealed by scanning FRET-FCS and Markov state models.* Proceedings of the National Academy of Sciences, 2020. **117**(36): p. 21889-21895.
234. Da, L.-T. and J. Yu, *Base-flipping dynamics from an intrahelical to an extrahelical state exerted by thymine DNA glycosylase during DNA repair process.* Nucleic acids research, 2018. **46**(11): p. 5410-5425.
235. Bringas, M., et al., *Ligand binding rate constants in heme proteins using markov state models and molecular dynamics simulations.* ChemPhysChem, 2019. **20**(19): p. 2451-2460.
236. Gordon, S.E., et al., *Dynamic Modelling Reveals 'Hotspots' on the Pathway to Enzyme-Substrate Complex Formation.* PLoS computational biology, 2016. **12**(3): p. e1004811.
237. Vasan, A.K., et al., *Role of internal loop dynamics in antibiotic permeability of outer membrane porins.* Proceedings of the National Academy of Sciences, 2022. **119**(8): p. e2117009119.
238. Barros, E.P., et al., *Markov state models and NMR uncover an overlooked allosteric loop in p53.* Chemical Science, 2021. **12**(5): p. 1891-1900.
239. Banerjee, R., H. Yan, and R.I. Cukier, *Conformational transition in signal transduction: metastable states and transition pathways in the activation of a signaling protein.* The Journal of Physical Chemistry B, 2015. **119**(22): p. 6591-6602.
240. Malmstrom, R.D., et al., *Allostery through the computational microscope: cAMP activation of a canonical signalling domain.* Nature communications, 2015. **6**(1): p. 1-11.
241. Scherer, M.K., et al., *PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models.* Journal of Chemical Theory and Computation, 2015. **11**(11): p. 5525-5542.
242. Noé, F. and F. Nuske, *A variational approach to modeling slow processes in stochastic dynamical systems.* Multiscale Modeling & Simulation, 2013. **11**(2): p. 635-655.
243. Nuske, F., et al., *Variational approach to molecular kinetics.* Journal of chemical theory and computation, 2014. **10**(4): p. 1739-1752.
244. Scherer, M.K., et al., *Variational selection of features for molecular kinetics.* The Journal of chemical physics, 2019. **150**(19): p. 194108.
245. Wu, H. and F. Noé, *Variational approach for learning Markov processes from time series data.* Journal of Nonlinear Science, 2020. **30**(1): p. 23-66.
246. Molgedey, L. and H.G. Schuster, *Separation of a mixture of independent signals using time delayed correlations.* Physical review letters, 1994. **72**(23): p. 3634.
247. Pérez-Hernández, G., et al., *Identification of slow molecular order parameters for Markov model construction.* The Journal of chemical physics, 2013. **139**(1): p. 07B604_1.
248. Schwantes, C.R. and V.S. Pande, *Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9.* Journal of chemical theory and computation, 2013. **9**(4): p. 2000-2009.
249. Lyman, E. and D.M. Zuckerman, *On the structural convergence of biomolecular simulations by determination of the effective sample size.* The journal of physical chemistry B, 2007. **111**(44): p. 12876-12882.
250. Lloyd, S., *Least squares quantization in PCM.* IEEE transactions on information theory, 1982. **28**(2): p. 129-137.
251. Sculley, D. *Web-scale k-means clustering*. in *Proceedings of the 19th international conference on World wide web*. 2010.
252. Noé, F., et al., *Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments.* Proceedings of the National Academy of Sciences, 2011. **108**(12): p. 4822-4827.

253. Suárez, E., et al., *What Markov State Models Can and Cannot Do: Correlation versus Path-Based Observables in Protein-Folding Models.* Journal of Chemical Theory and Computation, 2021. **17**(5): p. 3119-3133.
254. Wehmeyer, C., et al., *Introduction to Markov state modeling with the PyEMMA software [Article v1.0].* Living Journal of Computational Molecular Science, 2019. **1**(1): p. 5965.
255. Husic, B.E. and V.S. Pande, *Note: MSM lag time cannot be used for variational model selection.* The Journal of Chemical Physics, 2017. **147**(17): p. 176101.
256. Laio, A. and M. Parrinello, *Escaping free-energy minima.* Proceedings of the National Academy of Sciences, 2002. **99**(20): p. 12562-12566.
257. Laio, A. and F.L. Gervasio, *Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science.* Reports on Progress in Physics, 2008. **71**(12): p. 126601.
258. Bonomi, M., et al., *The unfolded ensemble and folding mechanism of the C-terminal GB1 β-hairpin.* Journal of the American Chemical Society, 2008. **130**(42): p. 13938-13944.
259. Granata, D., et al., *Characterization of the free-energy landscapes of proteins by NMR-guided metadynamics.* Proceedings of the National Academy of Sciences, 2013. **110**(17): p. 6817-6822.
260. Biswas, M., B. Lickert, and G. Stock, *Metadynamics Enhanced Markov Modeling of Protein Dynamics.* The Journal of Physical Chemistry B, 2018. **122**(21): p. 5508-5514.
261. Marinelli, F., et al., *A Kinetic Model of Trp-Cage Folding from Multiple Biased Molecular Dynamics Simulations.* PLOS Computational Biology, 2009. **5**(8): p. e1000452.
262. Bernetti, M., et al., *An Integrated Markov State Model and Path Metadynamics Approach To Characterize Drug Binding Processes.* Journal of Chemical Theory and Computation, 2019. **15**(10): p. 5689-5702.
263. Warfield, B.M. and P.C. Anderson, *Molecular simulations and Markov state modeling reveal the structural diversity and dynamics of a theophylline-binding RNA aptamer in its unbound state.* PLOS ONE, 2017. **12**(4): p. e0176229.