Advances in Nanopore Sensing for DNA and Protein Analysis

by

Nahid Nizar Jetha

BASc, The University of British Columbia, 2004

A THESIS SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies
(Physics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2014

Abstract

Single molecule nanopore sensing is emerging as a powerful tool for probing the properties of individual biomolecules. This is particularly true for DNA where nanopore technology is being actively developed toward ultra low-cost, high-throughput whole-genome sequencing. Most approaches to nanopore DNA sequencing require that DNA translocate a nanometer-scale pore (nanopore). Understanding and characterizing the physics of DNA translocation through nanopores is critical to the design and optimization of these methods. We sought to investigate DNA translocation dynamics and elucidate the mechanism of DNA transport through nanopores. We show that stochastic, sequence-dependent DNA-pore binding interactions play an important role in translocation and lead to subdiffusive translocation dynamics, which in the case of short DNA strands, is consistent with fractional dynamics. We characterize the sequence-dependent kinetics of DNA translocation and show that nucleotide dwell-time in the pore can potentially be used as a metric to distinguish individual nucleotides in nanopore sequencing, opening up new avenues by which to optimize nanopore sequencing technology.

While development of nanopore DNA sequencing has largely dominated nanopore applications, other applications including nanopore protein analysis are of great interest as a means to explore protein conformational dynamics and structure at the single molecule level. We present methods by which to capture and trap proteins in nanopores (via asymmetric salt concentration) and analyze their complex dynamics (via Hidden Markov Model signal processing), resolving two important challenges associated with nanopore sensing of proteins. We apply these methods to characterize the kinetics and dynamics of the prion

protein in a nanopore (a protein whose conversion into a misfolded isoform is responsible for the pathogenesis of prion diseases in humans and animals) as a first step towards understanding the relationship between prion protein conformational dynamics and conversion in disease. Moreover, we demonstrate the potential of nanopore technology for highly-sensitive, real-time protein and small molecule detection based on single molecule kinetics with potential application in medical diagnostics. Our methods enable studies of the long timescale conformational motions of proteins known to be critically important to protein function, at the single molecule level, making nanopore sensing a new tool for studying protein dynamics.

Preface

Some of the content in this thesis has been published previously. In particular,

- Figure 2.1b and Appendix C (forming an α-hemolysin nanopore for single molecule analysis) are reproduced, with minor changes, from "N. N. Jetha, M. Wiggin, A. Marziali. (2009) Forming an α-HL nanopore for single molecule analysis, in *Micro and Nano Technologies in Bioanalysis* (Lee, J.W., and Foote, R.S., ed.), Humana Press, Totowa, NJ". The methods presented in this book chapter were developed by various alumni of the UBC Applied Biophysics Lab (including myself) and adapted from the original protocol by the Nanopore Group at UCSC. I wrote the book chapter with edits and comments from my co-authors.

- Figure 3-1 is reproduced from "M. Wiggin, C. Tropini, V. Tabard-Cossa, N. N. Jetha, A. Marziali. (2008) Nonexponential kinetics of DNA escape from α-hemolysin nanopores. *Biophys. J.* **95**: 5317-5323". I contributed to the theory in this paper and to the writing of the manuscript.

- Chapter 3, section 3.2.1 (sequence effects on DNA escape dynamics) is reproduced, with changes, from "N. N. Jetha, C. Feehan, M. Wiggin, V. Tabard-Cossa, A. Marziali (2011) Long dwell-time passage of DNA through nanometer-scale pores: Kinetics and sequence dependence of motion. *Biophys. J.* **100**: 2974-2980". This paper represents my original work prepared with support from my co-authors.

- Chapter 4, section 4.3 (nanopore analysis of wild-type and mutant prion protein), Appendix D (PrP^C expression and purification) and Appendix F (PrP^C protein calling) are all reproduced, with minor changes, from "N. N. Jetha, V. Semenchenko, D. S. Wishart, N. R. Cashman, A. Marziali (2013) Nanopore analysis of wild-type and mutant prion protein (PrP^C): Single molecule discrimination and PrP^C kinetics. *PLoS One*. **8(2)**: e54982". This paper represents my original work prepared with support from my co-authors. In particular: Conceived and designed the experiments: NNJ AM NRC. Performed the experiments: NNJ. Analyzed the data: NNJ. Contributed reagents/materials/analysis tools: VS DSW. Wrote the paper: NNJ VS AM.

Table of contents

List of tables

List of figures

List of abbreviations

DNA	Deoxyribonucleic acid
KCL	Potassium chloride
RNA	Ribonucleic acid
α-HL	α-hemolysin
dC	Deoxycytidine
dA	Deoxyadenosine
HMM	Hidden Markov Model
PTFE	Polytetrafluoroethylene
MspA	*Mycobacterium smegmatis* porin A
ssDNA	single stranded DNA
FPT	First passage time
CTRW	Continuous time random walk
GLE	Generalized Langevin equation
fLE	fractional Langevin equation
fGn	Fractional Gaussian noise
HPLC	High-performance liquid chromatography
dA27	27-mer poly-dA
dC27	27-mer poly-dC
dCdmC27	Heteropolymer of cytosine and 5-methylcytosine
dmC	5-methylcytosine
PrP^C	Cellular prion protein

CNS	Central nervous system
GPI	glycosylphosphatidylinositol
NMR	Nuclear magnetic resonance
PrP^{Sc}	PrP-Scrapie
CJD	Creutzfeldt-Jakob disease
sCJD	Sporadic Creutzfeldt-Jakob disease
ShPrP(120-232)	Truncated Syrian hamster prion protein
KKRR-ShPrP(120-232)	KKRR adapted ShPrP(120-232)
KKRR-ShPrP(120-232)-D178N	D178N mutant PrP^{C}
RMS	Root mean square
EM	Expectation maximization

Acknowledgements

This work would not have been possible without the contributions of many people. Foremost I thank Andre Marziali, my research supervisor. Andre has been a constant, and incredible source of support and guidance throughout my PhD and in my time as a research engineer in his lab. Andre is a brilliant physicist with remarkable talent and intuition for research. His work in science and education is inspirational. I am very fortunate to have been a graduate student in his group and to have had the opportunity to learn from him.

I thank committee member, and fellow collaborator Neil Cashman for our many fruitful discussions and his constant encouragement of my work; I have very much enjoyed our collaboration. I thank my committee members Alex MacKay, Eric Lagally and Hongshen Ma for their guidance and for serving on my committee.

I thank Matthew Wiggin, Vincent Tabard-Cossa, Jason Dwyer and Dhruti Trivedi for their many contributions to this work. I thank them also for the many discussions and good times that we've had in and outside of the lab. Working with you was very rewarding. I couldn't have asked for better colleagues. In particular, I thank Matthew Wiggin for training me on the method of forming organic pores which is foundational to this work, and Jason Dwyer for carefully reviewing this thesis and providing critical comments that have led to its improvement.

Thank you to Jonathan Nakane for his foundational work in nanopore sensing on which this work is built.

I thank Christopher Feehan for his important contributions, particularly to the sequence dependent DNA escape study that makes up a part of this thesis.

I thank all of the past members of the UBC Applied Biophysics Lab, particularly Joel Pel, Jason Thompson, Dylan Gunn and David Broemeling, for making the lab an extremely enjoyable and stimulating place to work.

I thank fellow collaborator Will Guest for stimulating discussions. I also thank collaborators Valentyna Semenchenko and David Wishart for their support and contributions to our study of PrP^C dynamics. Valentyna, in particular, expressed all of the PrP^C constructs used in this study. Despite having a full plate she always made our work a priority; I am very thankful to her. Thank you to collaborator Kate Lieberman for providing me with purified preformed heptameric α-hemolysin for our nanopore / PrP study, which accelerated my experiments and made pore-forming a much more pleasant exercise.

I thank my brothers, Arif and Nuran, for their support and for mustering up the energy to go out on Friday and Saturday nights when I needed to blow off steam because of terrible lipid, or hemolysin, or some reagent during the week which meant little to no data.

I thank my parents for their extraordinary love and support. From an early age they always stressed the importance of education, and while it arguably took me some time to heed their advice I'd like to think I eventually caught on; no doubt the promise of a new mountain bike for good grades was a major catalyst.

Lastly I thank my wife, Nashina; I love you deeply. The emotional support, love and constant encouragement that you have given me has touched my heart. You have always encouraged me to pursue my passion. Without you this PhD would have remained only a dream; I am forever grateful to you. It is you that set this pursuit in motion and it is to you, and to our two beautiful children, that I dedicate this thesis. I love you more than words can describe.

Dedication

To Isaac, Naya, and Nashina

My world, my life, my beloved

Chapter 1

Introduction

The ability to probe the structure and dynamics of individual biomolecules has enabled the study of biological systems at a fundamental level, allowing scientists to explore research questions outside the capability of classical techniques in biochemistry, i.e. techniques that measure aggregate properties of molecular populations. A key advantage of single-molecule methods, in comparison to ensemble measurements, is that they can uncover information that is hidden by population averaging [1,2,3], such as rare and/or transient phenomena or heterogeneity in molecule populations [1,2,3,4,5]. Single-molecule methods have consequently yielded unprecedented insights into biological systems and processes [1,2,3,4,5,6]. Of the diversity of techniques available (e.g. optical and magnetic tweezers [1,3,4,5,6, 7,8], single-molecule fluorescence [1,3,4,9], tethered particle analysis [1,10,11], and others) nanopore sensing is perhaps the youngest and is quickly emerging as an exceptional tool for probing individual biomolecules [12,13,14,15,16].

The principles of nanopore sensing are based on Coulter counting in which two baths filled with electrolyte are connected by a small orifice [17]. Ionic current through the orifice (e.g. generated by an electric potential) becomes a very sensitive measure of the orifice conductance, and thus an excellent metric by which to detect microscopic objects that modulate ionic current as they transit the orifice [17]. Nanopore sensing employs the same principle but with the orifice replaced by a nanometer-scale pore (nanopore) in an insulating

membrane, thus allowing nanoscale detection. At this scale, and operating in an electrolyte solution, the nanopore system is very well suited to biomolecule analysis, with the biomolecules in a near-native environment. By electrophoretically driving an individual biomolecule into a nanopore (figure 1-1) and directly monitoring the ionic current, information about the molecule's parameters, including its dynamics, structural and chemical properties, and parameters related to its motion in the pore can be extracted from the current signature [13,14,15,16,18,19]. The ability to measure these parameters in a single molecule in a near-native environment, and often without modification, is one of the unique strengths of nanopore methods.

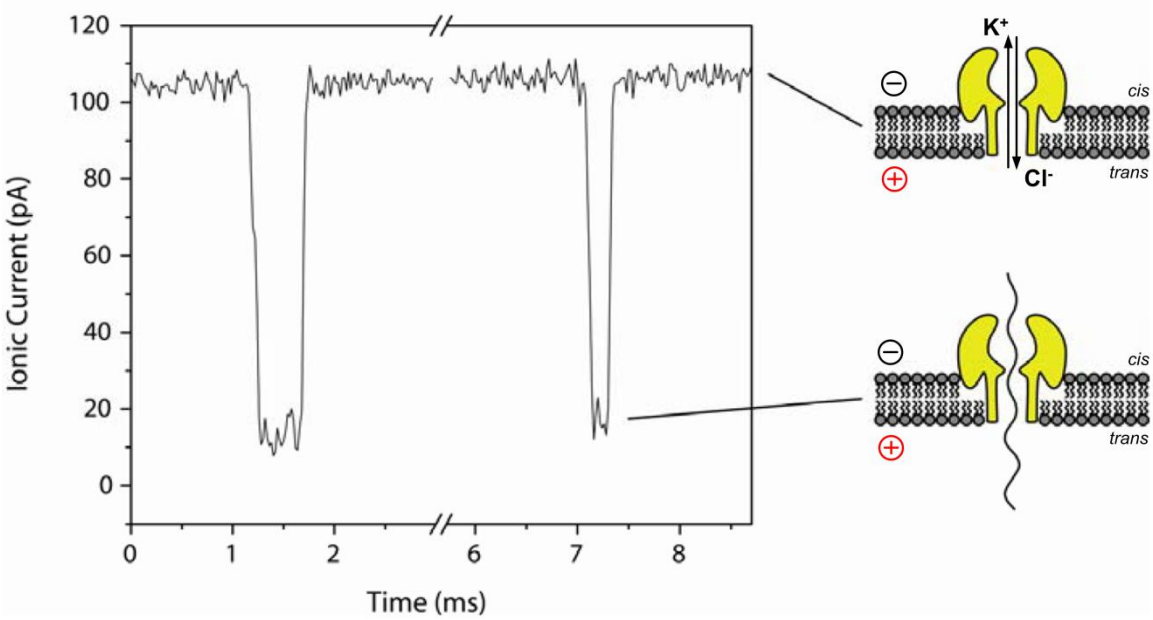


Figure 1-1. Nanopore sensing of DNA. (top right) Cartoon of the organic α-hemolysin nanopore incorporated in a lipid bilayer and immersed in an electrolyte (1M KCl). Application of a 115mV voltage across the pore (voltage polarity given by the plus and minus signs) generates a stable ~115pA current through the pore (left). **(Bottom right)** Single stranded DNA in solution (on the negative side of the pore) is electrophoretically driven into and through the pore thereby causing a measurable reduction in the pore current (direction of DNA motion is *cis*-to-*trans*). Ionic current through the pore is an extremely sensitive metric and rich source of information on the dynamics, structural and chemical properties of the individual biomolecule occupying the pore [14,15,16]. Image source: [20] (adapted, with permission).

Early demonstrations of the utility of nanopores for biomolecule analysis were made by probing the properties of individual strands of DNA and RNA [21,22,23]. These demonstrations showed the influence of molecule sequence on the properties of the current and indicated the potential of nanopore sensing as a pathway to ultra-low-cost, high-throughput DNA sequencing. It is this potential and the discovery that DNA sequence could in principle be detected by pore current that has driven the evolution and rapid growth of the field of nanopore sensing [12]. At the time of this discovery (1999), DNA sequencing was a slow, labor intensive and expensive process [24]. Even by the time the Human Genome was

completed in 2003, the routine sequencing of genomes for clinical diagnosis was far too costly (~100 million dollars per human genome [25]) to be taken seriously. Nanopore sensing was viewed as one possible technology to produce the massive throughput increase and cost reduction in DNA sequencing that would enable whole-genome sequencing to become regular practice in health care allowing, for example, clinicians to tailor strategies for disease prevention, diagnosis, and treatment to specific risk factors found in a patient's personal genome [26,27,28]. Today, while the cost of DNA sequencing has dropped many orders of magnitude to <$10,000 per human genome [25] (see appendix A for additional details on sequencing costs), the appetite for clinical sequencing has grown accordingly. While initially clinical sequencing was imagined as a tool to probe an individual's genome, clinicians are now exploring sequencing of disease genomes, particularly cancer genomes [27]. In this application, an enormous amount of sequencing must be performed for every patient, to collect genomic information independently for different tumors, and even for different locations within a tumor to probe for drug susceptibility. Based on current trends, it is expected that all future directed therapies for cancer (drugs aimed at specific tumor mutations), will require sequencing-based companion diagnostic tests, that must be relatively inexpensive (<$1,000) as they will need to be applied repeatedly through treatment. This scenario will ultimately require that DNA sequence equivalent to a human genome can be determined at costs of $100 or less. Nanopore sequencing remains the most promising technology for achieving this lofty goal, given its potential for very long read lengths (>100kb) [29], and label and amplification free sequencing with minimal sample preparation [29,30]; all features that make rapid, inexpensive DNA sequencing possible [29].

Chapter 1. Introduction

Most approaches to nanopore sequencing require that DNA translocate a nanometer-scale pore. DNA translocation rates however are too fast (microsecond timescale per nucleotide [29,30]) for detection and identification of individual nucleotides in sequencing. Moreover, stochasticity of DNA motion can create uncertainty in the number of bases that have translocated the pore [29], thereby severely limiting the accuracy of base-calling in nanopore sequencing. Indeed, slowing the translocation velocity and controlling DNA motion through pores is a foremost challenge in the development of nanopore sequencing technology [29,30]. Critical to meeting this challenge is obtaining a detailed understanding of the physics of DNA translocation through nanopores. In this regard, early models of translocation characterize the process as 1D normal diffusion [31], predicting exponential decay of the translocation time distribution at long times [32,33]. The appearance of a broader-than-expected distribution of translocation times [18,34], particularly for low velocity translocation of short DNA strands, however, indicates anomalous subdiffusion in DNA translocation. Anomalous diffusion is a generalization of diffusive transport characterized by a mean square displacement of $\left\langle x^2 \right\rangle \propto t^{\alpha}$ with subdiffusion corresponding to $0 < \alpha < 1$ ($\alpha = 1$ is normal diffusion) [35,36]. Two distinct anomalous diffusive models, in particular, have been put forth to explain the origins of subdiffusive DNA translocation dynamics [34,37], however, distinguishing between these models and developing a clear understanding of the underlying physics of subdiffusion in translocation remains an open problem.

Given this, we sought to investigate the effect of sequence on the low velocity translocation of short DNA strands through nanopores, as a means of investigating the origins of subdiffusive DNA translocation dynamics. We show that translocation dynamics

are highly sequence dependent. We develop a model of translocation based on stochastic, sequence dependent binding interactions between DNA and the pore and demonstrate, via Monte Carlo simulation, excellent agreement between our model results and the experimental data thereby showing that anomalous translocation dynamics arise out of DNA-pore binding interactions consistent with an anomalous diffusive model of DNA translocation based on fractional dynamics.

We characterize the translocation kinetics of poly-dA and poly-dC through the α-hemolysin (α-HL) organic nanopore and show an approximately two order of magnitude difference in characteristic translocation time between these two polymers. We present methods for characterizing the sequence dependent interaction energy between DNA and the pore and show that deoxycytidine (dC) binds more strongly to α-HL than deoxyadenosine (dA). Given the highly sequence dependent dynamics of DNA translocation our results point to the possibility of using nucleotide dwell-time in the pore as an orthogonal metric to nucleotide blockage current for distinguishing between individual nucleotides in nanopore sequencing, a possibility that opens new avenues for implementation of nanopore sequencing methods.

Lastly, we explore differences in the effect on DNA translocation dynamics between dC and methyl-dC (a critically important epigenetic nucleotide variant responsible for activation of some cancers) and show that these two nucleotides can be distinguished in wild-type α-HL based on blockage current alone. This demonstrates that in principle nanopore sequencing approaches based on wild-type α-HL can also potentially be used for methylation sequencing, a result which points to the potential of nanopore technology in cancer diagnostics and monitoring.

Our results, in combination with theory, experimental, and simulation work by others indicate the presence of important regimes in translocation in which translocation dynamics is governed by different processes depending on DNA length and translocation velocity.

Our work provides deeper insight into the process by which DNA moves through a nanopore, the understanding and characterization of which is critical to the optimal design and development of next-generation nanopore DNA sequencing and analysis technologies.

While initial interest in nanopore sensing arose out of its potential for low-cost sequencing [29,30] other applications including nanopore protein analysis have become of increasing interest as a means by which to probe the biophysical properties of proteins at the single molecule level. Nanopores offer unique advantages in this regard. In particular, unlike other single molecule methods nanopore sensing allows for label-free detection with minimal sample preparation [29,30] thereby offering the unique opportunity to probe the properties of unmodified proteins in a near-native setting. This is particularly useful in the case of medical diagnostics, where the combination of minimal sample preparation and high sensitivity (i.e. molecule detection at low concentrations [29,30]) makes nanopore analysis an attractive candidate technology for highly sensitive protein and small molecule detection. Given that proteins are an important class of biological marker in medical diagnostics [38,39] a simple method for protein analysis or counting could therefore be extremely useful. Nanopore protein analysis has already shown much promise as a means by which to study the unfolding kinetics of proteins [40,41], characterize protein-pore interactions [42,43,44], and to study the transport properties of proteins through pores [42,45,46]. It is therefore emerging as a

new and powerful method by which to study protein biophysics and form the basis of new biomedical diagnostic technologies.

A particularly exciting potential application of nanopore protein analysis is in the study of protein dynamics, specifically the long timescale conformational motions known to be critically important to protein function [47,48]. A major challenge, however, with nanopore analysis of protein dynamics is that in contrast to uniformly charged biopolymers such as DNA and RNA, the charge distribution of proteins and polypeptides can be highly irregular, positive, negative or neutral, significantly affecting the ability to capture and trap proteins in the pore. Moreover, proteins are folded polymers with a complex free energy landscape encompassing numerous transitions states and conformational sub-states [49]. This complexity coupled with fast transition state kinetics can complicate a detailed kinetic analysis of protein dynamics in the pore.

Given this, we have developed methods to probe the complex dynamics of single proteins with nanopores. In particular, we carry out experiments using a salt concentration gradient across the pore which has been shown to significantly enhance the nanopore capture rate of small molecules in solution (relative to symmetric salt conditions) [50,51]. Moreover, we have developed machine-learning signal processing algorithms, based on Hidden Markov Models (HMMs), to analyze protein conformational dynamics. HMM signal processing is a powerful technique by which to extract and characterize low-level signals buried in background noise [52,53,54]. We apply this to characterize the complex kinetics and dynamics of the prion protein in the pore (a protein whose conversion into a misfolded isoform is responsible for the pathogenesis of prion diseases in humans and animals [55,56,57,58]), which represents a first step in our investigation of the relationship between

prion protein conformational dynamics and the molecular mechanism of prion protein conversion in disease. Moreover, we demonstrate the sensitivity of nanopore analysis in detecting subtle changes in protein structure by showing that a wild-type and mutant prion protein exhibit easily distinguishable current signatures and kinetics inside the pore, and can be discriminated at the single molecule level. Our results point to the possibility of highly sensitive, real-time detection of proteins and small molecules in solution with application in medical diagnostics and demonstrate the ability of nanopore sensing for detailed characterization of single molecule protein dynamics.

In summary, this thesis advances the field of nanopore sensing for DNA and protein analysis. We have focused our efforts on elucidating the physics of DNA translocation through nanopores, understanding that is critically important to the design and optimization of nanopore DNA sequencing and analysis technologies. We have also developed methods by which to probe the dynamics of single proteins with nanopores and applied these in probing the structure and dynamics of the prion protein, demonstrating a potentially clinically significant application of our methods.

This thesis is organized as follows: Chapter 2 is a brief review of the field of nanopore sensing, beginning with a review of the origins and principles of the method followed by a discussion of the seminal work in nanopore sequencing. This discussion includes the principal challenges in developing nanopore sequencing technology along with an overview of the current state of nanopore sequencing including the significant progress made in overcoming these challenges. Chapter 3 deals with the physics of DNA translocation through nanopores, beginning with a brief overview of the early theory of translocation dynamics.

Chapter 1. Introduction

We discuss the experimental work demonstrating deviation from early theory predictions and showing anomalous subdiffusion in DNA translocation. We discuss two anomalous diffusive models put forth to explain the origins of subdiffusive translocation dynamics, namely, fractional dynamics and a model based on the fractional Langevin equation. Following this, we present our work on the sequence dependence of DNA translocation, results of which support the conclusion that subdiffusive dynamics stem from stochastic sequence dependent binding interactions between DNA and the pore. A significant portion of the content in chapter 3 is taken from the following paper:

1. N. N. Jetha, C. Feehan, M. Wiggin, V. Tabard-Cossa, A. Marziali (2011) Long dwell-time passage of DNA through nanometer-scale pores: Kinetics and sequence dependence of motion. Biophys. J. 100: 2974-2980.

Chapter 4 of this thesis presents our work in developing and extending methods of nanopore sensing for protein analysis, and application of these methods for probing the structure and dynamics of the prion protein. This chapter begins with a review of the field of protein analysis with nanopores, followed by a review of the prion protein and prion diseases. Following this, we present details of our work on nanopore analysis of the prion protein. A substantial portion of the content in chapter 4 is from the following paper:

2. N. N. Jetha, V. Semenchenko, D. S. Wishart, N. R. Cashman, A. Marziali (2013) Nanopore analysis of wild-type and mutant prion protein (PrP^C): Single molecule discrimination and PrP^C kinetics. PLoS One. 8(2): e54982.

Chapter 5 concludes and summarizes the thesis.

Chapter 2

Nanopore sensing: Background

2.1 Origins and principles of nanopore sensing

The origins of nanopore sensing can be traced back to the method of Coulter counting invented by Wallace H. Coulter in the 1940s and patented in 1953 under the title “Means for Counting Particles Suspended in a Fluid” [17,59]. In this method a pair of electrodes connected to an electrometer are placed across an orifice filled with, and immersed in, an electrolyte. The solution conductivity is measured as a function of time in the presence of a pressure gradient across the aperture. While the conductivity remains stationary for a pure electrolyte solution, pressure-driven flow of cells or particles, suspended in solution, through the orifice produces discrete resistive spikes in the current stemming from transient occlusions of the orifice by the analyte. The frequency of the pulses is related to the concentration of cells in the sample, while the distribution of pulse amplitudes correlates with the cell size distribution [12]. Originally used for the counting and sizing of blood cells [12,60], Coulter counters are now routinely used in medical laboratories for blood cell, spermatozoa, and platelet counting and have been applied in studies on HIV transmission and in counting cell nuclei in cancer cell lines [12,60,61,62,63]. The size range of the Coulter technique was extended, in the 1970s, to the submicron scale using sub-micrometer track-etched pores [64] but it wasn't until the 1990s that a true revolution of the technique was born in the form of nanopore sensing. With the development of sensitive patch-clamping

electronics in the field of electrophysiology in the 1970s [65] coupled with a leap in the technology for reducing the diameter of the hole through which target analytes translocate to the nanometer-scale, the technique was extended to the analysis of ions, small molecules and individual biomolecules. With the first examples by Sergey Bezrukov *et. al.* [66], George Church (not published) and John Kasianowicz *et. al.* [21], of detection of current blockage by a single molecule in a nanometer scale pore, nanopore analysis was born. Since its inception in the early 1990s the field of nanopore analysis has grown at a considerable rate as a result of large investments into its potential application to DNA sequencing [12].

Nanopore sensing is based on electrical detection of individual biomolecules. A nanometer-scale hole formed in an insulating membrane separates two electrolyte filled compartments. In one configuration, in which a lipid bilayer forms the insulating membrane, the two compartments are connected by an electrolyte-filled U-shaped Teflon® tube (figure 2-1). The compartments are connected to an Axopatch® 200B patch-clamp amplifier, by Ag/AgCl electrodes, which regulates the applied voltage across the pore. Application of a voltage drives the flow of ions through the pore, generating an ionic current.

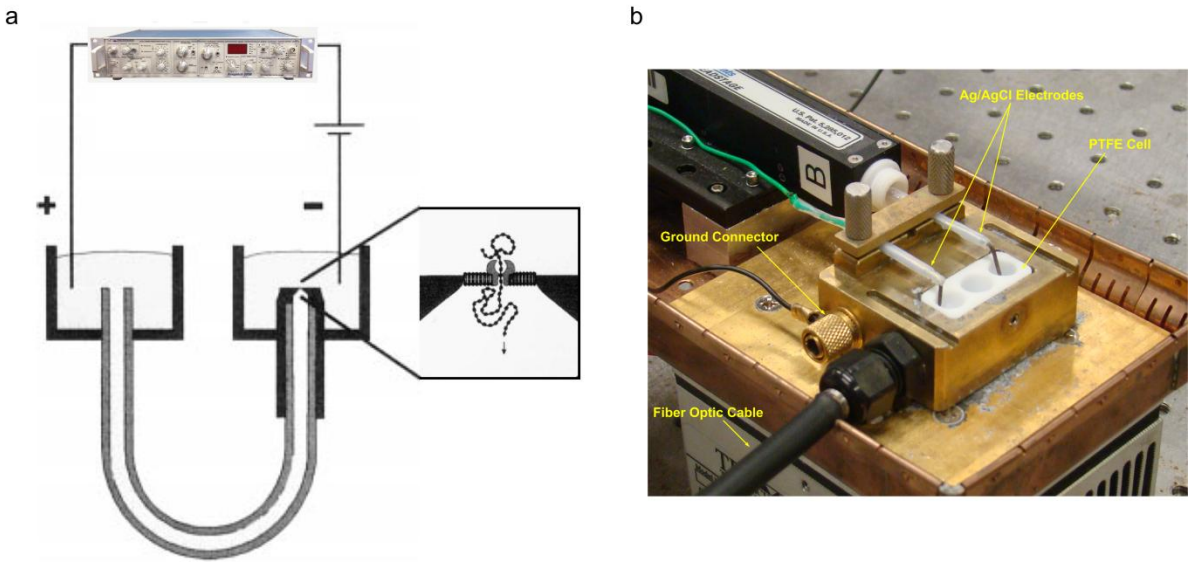


Figure 2-1. Bilayer and experimental apparatus. (a) A U-shaped Teflon® tube, filled with electrolyte, connects two electrolyte filled compartments. The compartments are connected to an Axopatch 200B patch-clamp amplifier by Ag/AgCl electrodes, which regulates the applied voltage and monitors ionic current with pA level resolution. One end of the Teflon® tube has a tip that narrows to a ~25µm diameter aperture (zoom view). A lipid bilayer is formed across the aperture, and one or more organic pores (e.g. α-hemolysin) are inserted into the bilayer (refer to Appendix C for details on the method of bilayer and pore formation). Nucleic acids are driven through the pore by an applied voltage. Image source: [22] (adapted by permission from Elsevier: Biophysical Journal [22], copyright 1999) **(b)** Experimental apparatus with components in place. The Polytetrafluoroethylene (PTFE - i.e. Teflon®) cell makes up the bilayer apparatus. A faraday cage is placed overtop of the system, once a pore has inserted into the bilayer, to attenuate electromagnetic noise on the current (refer to Appendix C for additional details on the experimental apparatus).

Charged molecules in solution, such as DNA, are electrophoretically driven through the pore and can be detected as a change in the ionic current through the pore relative to the open channel state (figure 1-1). Typically, capture and translocation of an individual biomolecule through the pore results in a decrease in the current, however, the magnitude and direction of the change is determined primarily by two factors [20]: Steric exclusion, which reduces the current by decreasing the effective pore-diameter, and ion-polymer interactions which includes co-ion repulsion (resulting in a decrease in the current through the pore) and

counter-ion attraction (which results in an increase) [20,67,68]. The relative effect of these factors largely depends on the salt concentration of the solution, which in 1M KCl (the most common ionic solution in nanopore sensing) steric exclusion and co-ion repulsion dominate, resulting in a ~80% reduction in the ionic current through the α-hemolysin protein pore upon entry and translocation of ssDNA (figure 1-1) [20,21,22,23,69,70]. As with Coulter counting, the frequency of the current spikes is related to the analyte concentration, while the distribution of pulse amplitudes and duration yields information on the properties of the translocating molecule and its interactions with the pore [21,22,23,42,43,69].

Two main classes of pores are used in nanopore sensing: Organic and synthetic pores. Organic nanopores, in general, have several advantageous features:

a) Their structure is highly reproducible, with atomic level precision, upon formation in a lipid bilayer [71].
b) Their structure can be determined with angstrom level precision using high-resolution techniques such as X-ray crystallography [71,72].
c) Organic pores typically have a hydrophilic interior and therefore immediately conduct current in aqueous solution upon application of a voltage [71,73,74].

The α-hemolysin protein nanopore is the most widely used biological pore in nanopore analysis. α-hemolysin (α-HL) is a structure formed in a lipid bilayer by self-assembly of seven identical polypeptides secreted from the bacterium *Staphylococcus aureus* [72]. It is a mushroom-shaped heptameric pore that consists of a transmembrane β-barrel and a cap region which is positioned outside of the membrane (figure 2-2). The pore is 10 nm in length

with an inner diameter that varies between 2.9 nm at the *cis* entrance, 4.1 nm in the internal cavity, 1.4 nm at the inner constriction, and 2.6 nm at the *trans* entrance of the β-barrel [71].

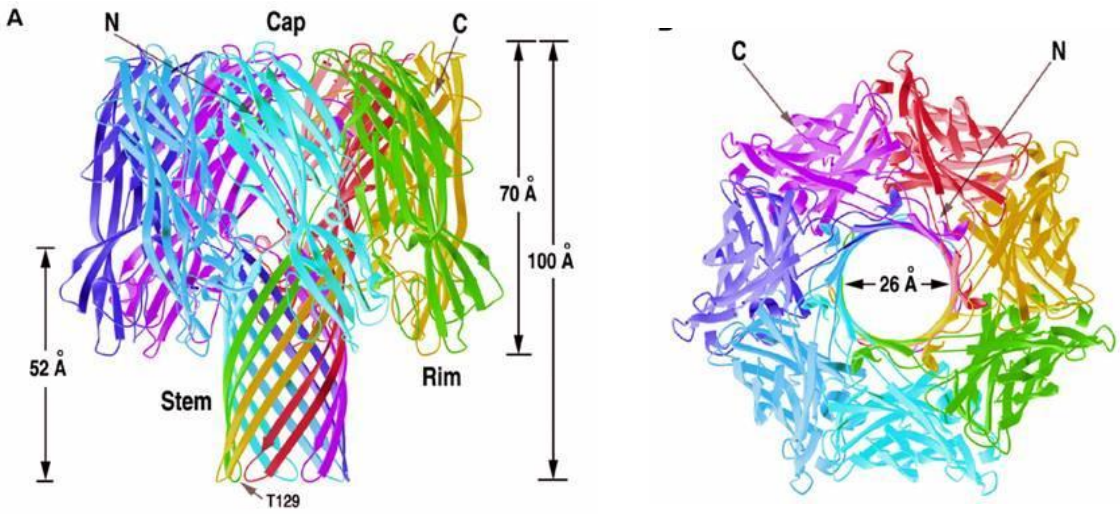


Figure 2-2. Crystal structure of the organic α-hemolysin nanopore (PDB code: 7ahl). α-HL is a heptameric transmembrane pore formed by the self-assembly of monomeric subunits secreted by the bacterium *Staphylococcus aureus* [72]. The cap region of the pore (at the top) which comprises the pore vestibule has an internal diameter that ranges from 2.9 nm at its entrance to 4.1 nm in the internal cavity [71]. The limiting aperture of the pore (i.e. the smallest constriction) is ~1.4 nm in diameter and connects the cap with the stem region of the pore that comprises the pore β-barrel [72]. Image source: [20] (adapted from [72]) (reproduced, with permission).

Several features of α-HL make it very well suited for nanopore sensing:

a) The pore is stable and rarely gates (i.e. it rarely undergoes a spontaneous conformational change, typically closing, which in the case of other pores, for example, can be triggered by voltage, anions or cations). Moreover, α-HL maintains this stability under a variety of solvent conditions (i.e. over a range pHs, salt concentrations, and temperatures) [73,75].

b) The pore is very robust and has a high resistance to denaturation. For example, α-HL can maintain a stable conformation over a range of solvent pHs (pH 3 – 11)

[75,76,77], at very high temperatures (~ 100°C) [78] and in the presence of high concentrations of denaturants (e.g. 2M guanidinium chloride) [79].

c) α-HL is a low noise pore. Under an applied voltage the ionic current conducted by the pore is ~100 fold greater than stochastic current noise. This means that current blockage events comprising reductions of as little as a few percent of the current can be detected.

d) α-HL is of the same size and scale of most biomolecules of interest. The internal dimensions of the pore are similar in size to individual DNA, RNA and protein molecules, which means that in general only a single biomolecule can occupy the pore at any given time, and that current blockages from individual molecules are well above the background noise.

e) α-HL, like most other organic pores, is amenable to protein engineering. Well established techniques in biochemistry and molecular biology, such as site-directed mutagenesis, can be used to modify the structure and chemical properties of the pore in order to optimize its functionality as a biosensor [71]. For example, in order to facilitate capture and translocation of DNA through the pore, negatively charged amino acids at the pore constriction can be substituted for neutral or positively charged amino groups [80].

While α-HL is the most extensively used organic pore in nanopore sensing several other organic pores have also recently been used. Most notably the channel protein *Mycobacterium smegmatis* porin A (MspA). MspA is an octameric, funnel shaped, protein pore with a limiting aperture that is ~1.2 nm in diameter (i.e. slightly smaller than the α-HL constriction)

[30]. In contrast to α-HL, the constriction is located at the *trans* opening of the pore and is a mere ~0.6 nm in length [15]. In the context of DNA sequencing this has the advantage that a smaller number of nucleotides make-up the total pore current during DNA translocation, resulting in a significant improvement in resolution relative to α-HL [81]. In addition to MspA, the Aerolysin organic pore has also been used in nanopore sensing [82,83,84]. Aerolysin however, has a net negative charge which poses a challenge in the context of DNA analysis because the high negativity (-52e [84]) translates into a large energy barrier for capture of DNA into the pore. Nevertheless, Aerolysin (whose internal diameter varies between 1 – 1.7 nm along the length of the pore [84]), has been shown very useful in nanopore protein analysis for probing the properties of proteins and polypeptides [82,83,84].

Despite their many advantages organic pores also present some challenges in the context of nanopore sensing, the most notable being their high fragility while incorporated in a bilayer [71]. Mechanical or electrical perturbations can disrupt the bilayer in which the pore is incorporated thereby terminating any nanopore experiment. Methods to improve lipid bilayer stability, including reducing the diameter of the aperture over which the bilayer is painted [85] (from the micron scale to the nano scale) have been successful for example, in increasing the operating voltage under which typical nanopore experiments may be run [85]. However, bilayer stability is an active field of research in nanopore science particularly given the serious challenge posed by fragile bilayers with respect to incorporating organic pores in devices for commercial applications. Given this, synthetic pores have been developed which can resolve these issues and which have the added advantage of being stable over a greater range of solution / environmental conditions (e.g. pH, voltage) than most organic pores, however they too suffer from problems, mostly in the form of very low signal-to-noise ratios

which can severely affect their utility in nanopore sensing [86]. Further discussion of synthetic pores can be found in Appendix B, as they are not central to the work presented in this thesis.

2.2 Seminal work in nanopore sequencing

Nanopore sensing of DNA and RNA was first reported in 1996 by Kasianowicz *et. al.* who showed that individual downward spikes in the channel current corresponded to the single-file translocation of individual nucleic acids through an α-HL nanopore [21]. The authors further demonstrated that nanopore sensing could in principle be used to characterize the length of the translocating polymer by measurements of the translocation time [21], thereby showing that properties of the molecule can be deduced from statistical analysis of the pore current. Given that individual DNA strands could be detected as they translocated the pore, and that molecule properties can be characterized during its translocation, it was proposed that the sequence of the translocating polymer might also be determined through ionic current measurements [21]. In particular, if the individual nucleotides that make up the translocating strand blocked the channel current in a characteristic manner as they passed single-file through the pore, then DNA sequencing with a nanopore could be performed through straightforward reading of the sequence of current modulations during strand translocation [29]. To test this, Akeson *et. al*. translocated DNA and RNA homopolymers and heteropolymers with different nucleotide compositions through an α-HL nanopore and characterized the sensitivity of the pore current to DNA and RNA sequence [22]. While the pore current (blockage amplitude and translocation time), is indeed sequence dependent, it was found that given the speed with which DNA translocates the pore (1-20μs per nucleotide

at 120mV across the pore [87]), and given that multiple adjacent nucleotides are collectively responsible for blocking the pore current at any given time during translocation, a contiguous sequence of at least 30 identical nucleotides in a given strand was required in order for sequence dependent differences in blockage amplitudes to be detected [22]. These results were also later supported by Meller *et. al.* [23]. Moreover, it was shown by Akeson *et. al.* that pore current is also affected by the structural properties of the translocating molecule, particularly DNA / RNA secondary structure [22]. For example, differences in pore current between poly-dC (DNA) and poly-C (RNA) were shown to be primarily due to differences in secondary structure between these two homopolymers (poly-dC is largely unstructured whereas poly-C tends to form a tight helix of similar diameter to the α-HL pore constriction [22]. Consequently poly-C occludes the ionic current to a greater degree than poly-dC [22]). This complex dependence of the pore current on the properties of DNA was some of the first evidence that the use of nanopores for DNA sequencing would not be simple. Nevertheless, the discovery that nucleotides can be distinguished by nanopore current showed that in principle the method could be made workable for sequencing provided two fundamental issues can be resolved:

1) **Pore spatial resolution and sensitivity.** The sensitivity of the pore would need to be improved such that the pore current at any given time during translocation would sample a smaller number of nucleotides (ideally a single nucleotide).

2) **DNA translocation rate and translocation control.** DNA translocates the pore too quickly to observe individual nucleotides in translocation [29,30]. The translocation

rate would need to be reduced by ~3 – 5 orders of magnitude, to enable longer dwell time of DNA in the pore, enabling longer integrations times to achieve single nucleotide resolution. Moreover, greater control over the translocation process will be required. The appearance of a broad distribution of DNA translocation times increases the uncertainty in the number of bases that have transited the pore [29]. Reducing the stochasticity of DNA motion and improving overall control of translocation (e.g. controlling back-stepping of nucleotides) will be critical for successful sequencing.

Outside of these, other issues remain (e.g. improving lipid bilayer stability in the case of sequencing using a protein pore, dealing with DNA secondary structure, etc...), however it is recognized that the above two problems represent large hurdles with respect to the realization of nanopore sequencing. As discussed below, however, significant progress has been made in addressing both of these issues.

2.2.1 Pore spatial resolution and sensitivity

In the case of biological pores, the group of Hagan Bayley has demonstrated the great utility of protein engineering for improving pore sensitivity [88,89]. In one study the authors showed significantly improved discrimination between individual nucleotides contained in an otherwise poly-dC homopolymer using mutant α-HL (over wild-type) [88]. Site-directed mutagenesis of glutamic acid (E111N) and lysine (K147N) residues in the central constriction of the α-HL pore (a key base recognition site of α-HL), to form the so-called NN mutant, is thought to increase the effective diameter of the constriction in the presence of

ssDNA thereby increasing the current flow through the pore and improving discrimination between nucleotides in the region [88]. In a subsequent study, additional changes to residues in the α-HL constriction, in particular M113Y, led to further improvements in base discrimination over both the NN mutant and wild-type α-HL [89].

While α-HL is the most widely used pore in nanopore sensing, a genetically engineered form of the MspA biological pore, with a decreased negative charge density relative to wild-type MspA to facilitate capture and translocation of DNA, has recently been developed and has shown considerable promise as a candidate pore with excellent spatial resolution [81]. As mentioned previously, MspA is a funnel shaped pore with a ~1.2 nm diameter constriction (slightly smaller than the α-HL limiting aperture) [30]. Its shape and constriction size mean that the current through the pore in the presence of single stranded DNA (ssDNA) is determined by a smaller subset of nucleotides relative to α-HL [81]. Moreover, the separation in nucleotide-specific current through MspA is enhanced by approximately 3.5 times relative to α-HL [81]. The substantially improved sensitivity of the MspA pore for base recognition over wild-type α-HL make it an attractive candidate for nanopore DNA analysis and sequencing applications.

2.2.2 DNA translocation rate and translocation control

Substantial developments have also been made with respect to slowing the translocation velocity of DNA [30]. By controlling a variety of experimental parameters such as temperature, solution viscosity (e.g. through the addition of glycerol), salt concentration, and applied voltage, Fologea *et. al.* demonstrated a ~10 fold reduction in translocation speed relative to standard experimental conditions (e.g. 1M KCl, 20ºC, 120mV) [90]. Use of

organic salts (versus KCl) has also been shown to slow DNA translocation [91]. Others have explored the use of chemically labeled nucleotides on the DNA strand thereby increasing the effective diameter of the labeled nucleotide and slowing its passage (and consequently that of the entire strand) through the α-HL pore [92].

An interesting approach proposed by Derrington *et. al.*, that effectively addresses both issues of translocation speed and control, and which in early demonstrations with MspA has been shown capable of sequencing very short strands of DNA (five nucleotides in length), involves interspacing individual nucleotides in a DNA strand with a small DNA duplex [81]. In this approach, known as duplex-interrupted sequencing [81], analyte DNA is converted into a synthetic DNA strand whereby each nucleotide to be sequenced is converted into the same nucleotide plus a small (~14 base) duplex encoding region. After conversion, DNA strands that are complementary to the duplex encoding region are added to solution hybridizing to appropriate regions on the synthetic strand, thereby forming a DNA molecule with individual nucleotides interspaced by duplexes. The molecule is then electrophoretically driven into an MspA nanopore, translocating the pore until a duplex encounters the pore constriction, at which point the strand is halted and the single nucleotide overhang resident within the constriction is read via the pore current. Continuous application of the voltage eventually results in duplex dissociation, and subsequent strand translocation until the next duplex in the molecule encounters the constriction. In this approach DNA translocation is regulated by the binding energy of the duplex and the voltage-dependent timescale at which duplex dissociation occurs. Free design of the duplex, therefore, gives significant control over the translocation process, particularly the translocation rate, providing sufficient time to capture and read the residual current associated with the individual nucleotide occupying the

limiting aperture of the pore. Given that the pore current through MspA is primarily determined by a set of trinucleotides that occupy its limiting aperture [81], this approach to sequencing has been shown workable in the case where individual nucleotides are represented by a set of trinucleotides (e.g. A -> AAA, C-> CCC, etc…). In this case, as the duplex halts translocation, a trinucleotide set occupies the pore constriction and can be readily identified from observation of the ionic current [81]. While this approach to sequencing could potentially be fast and offer long read-lengths, the ability to convert and read large fragments of DNA with high accuracy via this method has yet to be demonstrated.

A promising approach to slowing DNA translocation to a uniform rate that may enable direct strand sequencing is through the use of enzyme motor proteins (e.g. a DNA polymerase) that bind DNA and regulate its motion through the pore via synthesis of a complementary DNA strand [93]. This approach to translocation rate control has four key advantages: (i) Because the enzyme binds DNA in solution the enzyme-DNA complex can be easily driven into the pore with an applied voltage. (ii) As the enzyme synthesizes the complementary strand the ssDNA that is threaded through the pore experiences a much slower and controlled motion as each nucleotide is incorporated into its complementary strand (e.g. a typical polymerase has turnover times of tens of milliseconds per nucleotide added [94]). (iii) Instead of a continuous motion of DNA through the pore, enzymatic control produces a ratcheting effect in which each nucleotide halts in the pore as the enzyme goes through its synthesis cycle [93], allowing for greater signal averaging of the current produced by each nucleotide. (iv) A significant amount of control over the translocation process can be

obtained by regulating the concentration of divalent cations in solution required for DNA synthesis.

2.2.3 Nanopore sequencing approaches

Using this approach to translocation control coupled with the sensitivity of the MspA nanopore, Manrao *et. al.* recently demonstrated the feasibility of nanopore DNA sequencing [15] (figure 2-3). Using phi29 DNA polymerase to control DNA motion through the pore (an enzyme previously shown capable of reducing the DNA translocation rate from microsecond timescales per nucleotide to tens of milliseconds per nucleotide [93]), the authors were able to resolve differences between sets of four nucleotides occupying the MspA pore constriction during translocation with single nucleotide resolution [15]. Much work remains in order to demonstrate complete sequencing via this method, for example, development of automatic sequence extraction algorithms, handling extended homopolymer regions, and measuring the library of current levels for all possible four nucleotide sequence combinations [15]. Nevertheless, this work represents an important milestone on the path towards nanopore sequencing and clearly demonstrates the potential of this approach to achieve low-cost, high-throughput DNA sequencing.


Cis
Trans
DNA motion
CTATC
GATAG
ACTCTAGACTTTTTTAAATTTGGGTTTCCCCCXXTCA

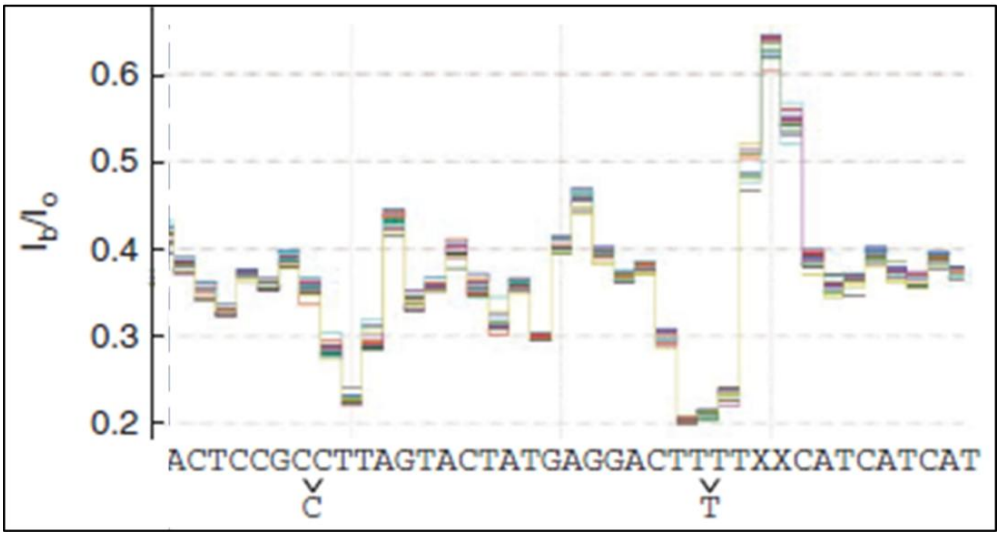

0.6
0.5
0.4
0.3
0.2
I_b/I_o
ACTCCGCCTTAGTACTATGAGGACTTTTXXCATCATCAT
C
T

Figure 2-3. Nanopore sequencing with a DNA polymerase. (top) Cross-sectional view of an MspA nanopore (purple) incorporated in a lipid bilayer. Captured, and docked over top of the pore is a phi29 DNA polymerase - DNA complex (polymerase in green) synthesizing a complementary strand to the single stranded DNA (ssDNA) molecule that is threaded through the pore. As individual nucleotides are added to the complimentary strand ssDNA in the pore moves, in single nucleotide steps towards the *cis* side of the pore. The polymerase regulates the motion of ssDNA through the pore thereby reducing the DNA translocation rate from microsecond timescales per nucleotide to tens of milliseconds per nucleotide enabling single nucleotide resolution [15,93]. **(bottom)** Demonstration of the repeatability of nanopore sequencing with a polymerase and the similarity of the current levels for the bases that makeup the translocating DNA molecule. Each color represents a single event in which the entire complimentary strand (sequence shown below) is synthesized (shown are results for 20 different events). Each current level represents the current through the pore in the presence of a given set of four nucleotides in the pore constriction (current is expressed as the ratio of the blockage current (I_b) to the open-channel current (I_O)). Undetected current levels, due to repeat nucleotides and / or similar currents are indicated with the missed nucleotide shown below the sequence. The excellent repeatability of this approach demonstrates its viability for nanopore sequencing. Image source: [15] (reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology [15], copyright 2012).

Another interesting approach that makes use of a DNA-processing enzyme was proposed by Clarke *et. al.* who use an exonuclease to cleave off individual mononucleotides from a DNA strand which are subsequently electrophoretically driven through a mutant α-HL nanopore [95]. The mutant pore is engineered by a combination of site-directed mutagenesis of residues in the pore constriction, for improved base sensitivity over wild-type α-HL [88,89,95], and by covalent attachment of an aminocyclodextrin adaptor in the β-barrel of α-HL to further constrict the channel [95]. Using this approach the authors show that the individual mononucleotides translocate the pore in millisecond timescales (significantly slower than in direct DNA polymer translocation) and could be detected and identified with greater than 99% accuracy [95]. Moreover, in addition to distinguishing between the four DNA bases, 5'-methylcytosine could also be readily detected and identified via this method, thereby demonstrating the potential of nanopore sensing for methylation sequencing with

application to cancer diagnostics and monitoring. This approach to sequencing known as ‘sequencing by digestion’ or ‘exonuclease sequencing’ could become feasible provided there is a way to chemically attach an exonuclease to the pore and to guarantee that each individual nucleotide, once cleaved, enters and translocates the pore with high probability. Exonuclease sequencing is one of two strategies to nanopore sequencing currently in development by Oxford Nanopore Technologies, a leading nanopore technology company in pursuit of developing the first commercial nanopore sequencer.

In addition to the aforementioned approaches to nanopore sequencing with demonstrated proof-of-concepts, several other approaches that are uniquely suited to synthetic nanopores are also under active investigation, including sequencing by measuring characteristic tunneling currents across individual nucleotides translocating a synthetic pore embedded with electrodes [30,96] (figure 2-4). Experiments with scanning tunneling microscopes [97] indicate the possibility of discriminating between the nucleotides in this way and therefore sequence DNA oligomers [30,98]. This is because the energy gap between the highest occupied and lowest unoccupied molecular orbitals of each nucleotide is unique [30,99]. Nanofabricated metallic gap junctions have also been used to identify individual nucleotides by measuring tunneling currents [30,96]. Neither of these studies involved nanopores, however if a tunneling sensor could be fabricated in a solid-state nanopore, and the DNA translocation velocity reduced, it might be possible to sequence DNA in this way [30].

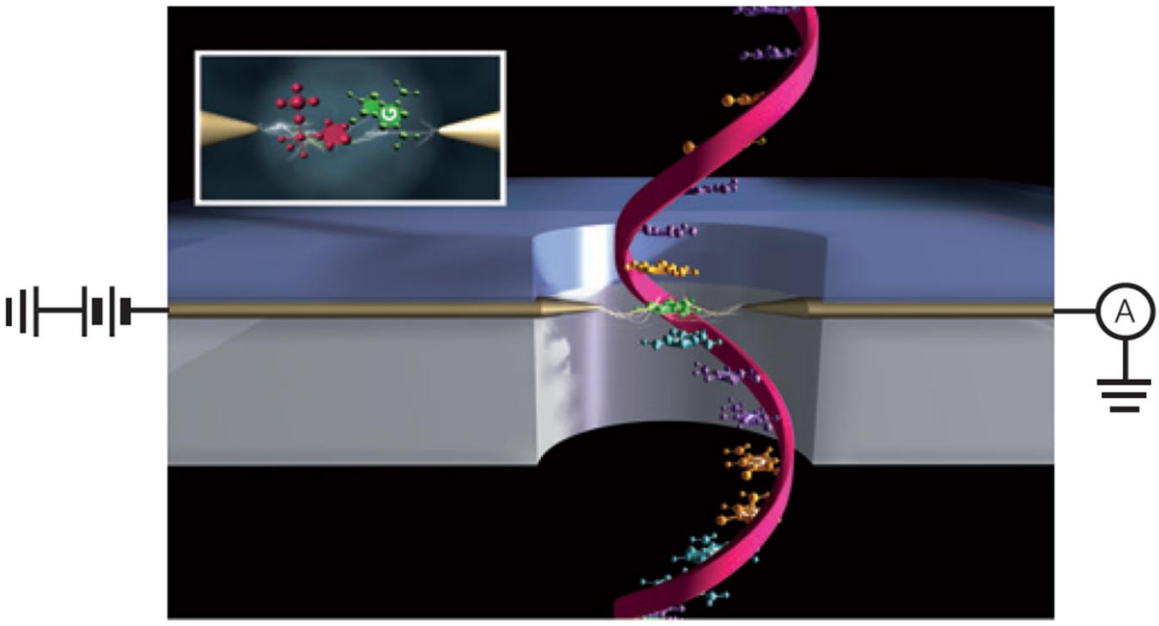


Figure 2-4. Nanopore sequencing via electron tunneling. Single stranded DNA translocating a synthetic nanopore embedded with a set of tunneling electrodes (gold) spaced ~1 nm apart. Changes in tunneling current as individual nucleotides pass the electrodes could be used to identify the bases. **Inset:** A nucleotide positioned between the tunneling electrodes in the pore. Image source: [30] (reprinted by permission from Macmillan Publishers Ltd: Nature Nanotechnology [30], copyright 2011).

Several other approaches to nanopore sequencing are in development, including: Hybridization assisted sequencing [100], optical recognition of converted DNA [101], base-by-base ratcheting of ssDNA using electrostatic traps [102], and graphene-based sequencing [103,104,105] (refer to the following reviews [29,30,106] which provide additional details on these methods). Regardless of the method, however, these approaches to nanopore sequencing (and others) require that DNA translocate a nanometer-scale pore, and therefore require precise control over the dynamics of DNA translocation. Unfortunately, uncontrolled DNA translocation times are highly variable with identical polynucleotides transiting the pore in times that can be orders of magnitude apart [22,23,29]. This can severely limit the accuracy with which bases can be detected and identified in a translocating strand for

sequencing. Critical to improving DNA translocation control is a detailed understanding and characterization of the physics of DNA motion through nanopores. Given this, we sought to investigate DNA translocation dynamics and elucidate the mechanism of DNA transport through nanopores.

Chapter 3

Subdiffusive dynamics of DNA translocation

Understanding and characterizing the physics of DNA translocation through nanopores is critical to the functionality of nanopore sequencing. Translocation is one part of the process of DNA motion from one side of a rigid wall to the other through a nanopore, a process that is also governed by the dynamics of DNA capture. DNA capture is the process by which DNA in a coiled state in solution arrives at the mouth of the pore with one of its ends threaded into the pore. This process has been covered in detail elsewhere [50,107] but briefly involves a combination of processes including drift-diffusion of DNA towards the mouth of the pore and the crossing of an entropic barrier arising (primarily) from the localization of one of the chain ends at the pore entrance to enable single-file threading through the pore. This chapter is concerned with the dynamics of DNA threading through the nanopore and into the receiver compartment upon capture (i.e. DNA translocation). Despite the conceptual simplicity of the translocation process it is governed by a variety of factors which can complicate an analysis of the dynamics.

This chapter begins with a discussion of the theory of DNA translocation dynamics. The presence of a broader-than-expected distribution of translocation times in the translocation of short, single stranded DNA at low velocity indicates subdiffusive dynamics of DNA translocation. We discuss two distinct models for explaining the origins of subdiffusion in

translocation, namely, fractional dynamics and the fractional Langevin equation, both of which are consistent with the data. We propose a method for distinguishing between these models which involves probing the sequence dependence of DNA translocation and we show, in carrying out this investigation, that subdiffusive translocation dynamics stem from stochastic, sequence dependent binding interactions between DNA and the pore in accordance with a model of DNA translocation based on fractional dynamics.

Our results, in combination with experimental, theoretical and simulation work by others indicates the presence of regimes in translocation in which translocation dynamics is governed by different processes depending on DNA length and translocation velocity, which we discuss.

3.1 DNA translocation dynamics: Background

Early models of DNA translocation through a narrow pore, based on the experimental results of Kasianowicz *et. al.* [21], describe the process as 1D diffusive given by a 1D Fokker-Planck equation [31],

$$\frac{\partial P}{\partial t} = \left(-v\frac{\partial}{\partial x} + K\frac{\partial^2}{\partial x^2} \right) P(x,t) \qquad (1)$$

where x is the distance along the polymer chain contour that has translocated the pore, $P(x,t)$ is the probability density function to find the chain with a length x at time t in the receiver compartment, and v and K are the drift and diffusion constants respectively. The first-passage

time distribution $F(x_0, t)$, i.e. the distribution of polymer translocation times, can be determined from the probability density function by [33]

$$F(x_0,t) = -\frac{d}{dt}\int dx P(x,t) \tag{2}$$

Assuming translocation without retraction from the pore with an initial position of the chain end at the pore entrance ($x_0 = 0$), in the presence of an external bias, $F(0,t)$ is given by [32,33],

$$F(0,t) = \frac{L}{\sqrt{4\pi K t^3}} \exp\left(-\frac{(L-vt)^2}{4Kt}\right) \tag{3}$$

where L is the length of the translocating polymer. Equation (3), in particular, predicts exponential decay at long translocation times and a mean first passage time (FPT) of $\int_0^\infty tF(0,t)dt = L/v$ [32], i.e. a mean FPT that corresponds to classical linear motion with a drift velocity v.

The first experimental observation of deviation from the predictions of Markovian dynamics (equation 3) came from Bates *et. al.* [34] in probing the translocation dynamics of short, single stranded DNA molecules (poly-dA, 60 nucleotides in length) through α-HL in the drift-free and low drift regime. In particular the authors showed that translocation times are broad relative to exponential decay, containing non-negligible contributions at long times

indicative of additional mechanisms in translocation that are not accounted for by simple Markovian models based on Brownian motion [32]. In exploring this further, we probed the translocation dynamics of short ssDNA molecules (poly-dA ranging in length from 15 to 65 nucleotides) by observation of the thermally activated escape of avidin-bound-DNA from an α-HL nanopore over an electrostatic energy barrier (figure 3-1a) [18]. In this configuration we obtain excellent control of the translocation process by simply tuning the voltage which adjusts the energy barrier for escape enabling the exploration of translocation dynamics over a very wide range of timescales (milliseconds to seconds) in the low drift regime [18]. We found that the first passage time distribution for DNA escape is slower than exponential, exhibiting a power law tail at long times (figure 3-1b) [18] thereby demonstrating non-Markovian translocation dynamics, specifically anomalous subdiffusion.

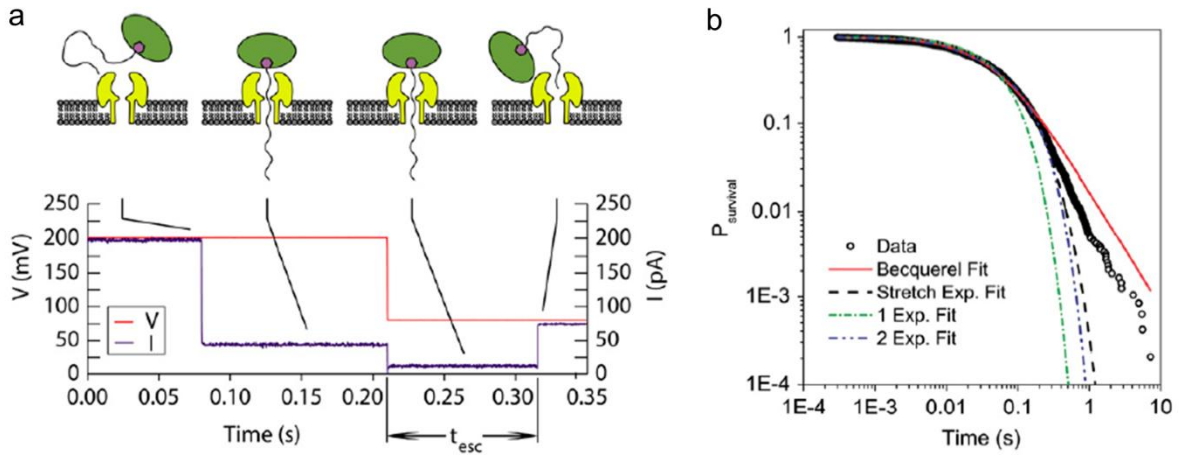


Figure 3-1. Single molecule DNA escape from α-HL. (a) Avidin-bound-ssDNA in solution (avidin in green, ssDNA in black) is captured into an α-HL nanopore (yellow) at 200mV (voltage trace shown in red), decreasing the ionic current (*I*) through the pore. Avidin acts as an anchor preventing the full translocation of ssDNA. The potential is then reduced to the escape potential (80mV shown). The molecule undergoes thermally activated escape from the pore after a time t_{esc} (the escape time). **(b)** Survival probability of poly-dA (27 nucleotides in length) in the pore against the escape potential (80mV). The survival probability is the probability that DNA is in the pore at time *t* upon application of the escape potential and is the integral of the FPT distribution for escape [32,36]. The log-log plot emphasizes the long power law region between 0.1s and 10s. Shown are single exponential, two-term exponential, stretched exponential and Becquerel fits to the data. Only the Becquerel fit (red), which decays as a power law at long times [18], qualitatively reproduces the data.

Unlike normal diffusion, anomalous diffusion is characterized by slowly decaying memory effects which translates into long-tails in the first passage time distribution [32,35,36]. In this case the mean-square displacement is given by $\langle x^2 \rangle \propto t^{\alpha}$, with subdiffusion corresponding to $0 < \alpha < 1$ [32,35,36]. Anomalous diffusion therefore is a generalization of diffusive transport with normal diffusion corresponding to the case of $\alpha = 1$. While our work was an important step towards understanding deviations from Markovian translocation dynamics, an open question remains as to the underlying physical mechanism giving rise to anomalous subdiffusion in DNA translocation. Two distinct models, in

particular, have been put forth in this regard: Anomalous diffusion based on fractional dynamics [32] arising, for example, from binding interactions between DNA and the pore [18], and anomalous diffusion based on the fractional Langevin equation due to the multiple timescale relaxation dynamics of the DNA chain [37]. These models are governed by fundamentally different equations (as discussed below) and therefore determining the appropriate model that accurately characterizes the subdiffusive dynamics of DNA translocation is critically important to understanding the translocation process. Before discussing our method for distinguishing between these models and presenting our results we first provide a brief review of each.

3.1.1 Fractional dynamics

Fractional dynamics is based on the continuous time random walk (CTRW) model of diffusion [35,36]. CTRW is a generalization of a simple random walk whereby the step length and waiting time between steps are distributed random variables given by the jump distribution $\psi(x,t)$ (this is in contrast to a simple random walk in which the step length and waiting time are constant). The distribution of step lengths can be determined from the jump distribution by,

$$\lambda(x) = \int_0^\infty dt\,\psi(x,t) \tag{4}$$

and the distribution of waiting times by,

$$w(t) = \int_{-\infty}^{\infty} dx\, \psi(x,t) \tag{5}$$

A simple random walk is therefore a specific case of a CTRW where $\lambda(x)$ and $w(t)$ are delta function distributions. If the step length and waiting time are independent random variables the jump distribution is decoupled and can be expressed as $\psi(x,t) = \lambda(x)w(t)$.

Fractional dynamics emerges if, for a given CTRW process, the characteristic waiting time T given by,

$$T = \int_{0}^{\infty} dt w(t) t \tag{6}$$

or the step length variance Σ^2,

$$\Sigma^2 = \int_{-\infty}^{\infty} dx \lambda(x) x^2 \tag{7}$$

diverge. Of particular relevance, with respect to DNA translocation, is when, for finite Σ^2, the distribution of waiting times exhibits a power law tail at long times of the form,

$$w(t) \sim \frac{\tau^{\alpha}}{t^{1+\alpha}}, \qquad (0 < \alpha < 1) \tag{8}$$

where τ is an internal timescale. In this case T diverges (along with all other moments of the waiting time distribution) and large waiting times occur with comparatively high probability thereby governing the transport process. It can be shown [35,36,108] that under these conditions, in the presence of an external field, the transport dynamics is subdiffusive and is given by the fractional Fokker-Planck equation,

$$\frac{\partial P}{\partial t} = {}_0D_t^{1-\alpha}\left(-v_\alpha \frac{\partial}{\partial x} + K_\alpha \frac{\partial^2}{\partial x^2}\right)P(x,t) \tag{9}$$

where v_α is a generalized velocity with dimension $[v_\alpha] = ms^{-\alpha}$, K_α is the generalized diffusion coefficient given by $K_\alpha = {}^{Kv_\alpha}\!/_{v}$, and the fractional Riemann-Liouville operator,

$${}_0D_t^{1-\alpha}P(x,t) = \frac{\partial}{\Gamma(\alpha)\partial t}\int_0^t \frac{P(x,t')}{(t-t')^{1-\alpha}}dt' \tag{10}$$

In this case the mean displacement is given by $\langle x \rangle = v_\alpha t^\alpha$, in contrast to normal diffusion where $\alpha = 1$ and the corresponding mean is $\langle x \rangle = vt$. In normal diffusion the fractional Fokker-Planck equation (equation 9) reduces to the Fokker-Planck equation (equation 1).

In the context of DNA translocation, Metzler *et. al.* showed that if the normal diffusive motion of DNA out of the pore is successively stalled with a waiting time for the process to

resume that is distributed according to equation 8 then such a multiple trapping model would lead to fractional dynamics of DNA translocation [32,35]. In this case the governing dynamics is given by the fractional Fokker-Planck equation (equation 9) and the distribution of DNA translocation times (i.e. the FPT distribution) would decay as a power law at long times [32] consistent with our findings (figure 3-1b).

A physical mechanism that could potentially give rise to fractional DNA translocation dynamics is stochastic binding interactions between DNA and the pore. In such a model the normal diffusive motion of DNA out of the pore is successively stalled by random binding events between DNA and the pore. In exploring this hypothesis we developed and simulated a model of DNA translocation involving stochastic binding interactions between DNA and the pore. Before presenting details of our model and results we discuss anomalous translocation dynamics based on the fractional Langevin equation.

3.1.2 Fractional Langevin equation

The Langevin equation approach to characterize the motion of a Brownian particle is based on Newton's second law extended to include random forces. In particular, for standard Brownian motion the random force corresponds to uncorrelated (white) Gaussian noise (i.e. standard Brownian motion is a memoryless process). When the random noise is non-white, however, and exhibits correlations, the resulting motion is described by the generalized Langevin equation (GLE), which in one dimension is given by,

$$m\frac{d^2x(t)}{dt^2} = -\zeta\int_0^t dt' K(t-t')\frac{dx(t')}{dt'} - \frac{dU(x)}{dx} + \theta(t) \qquad (11)$$

where *m* is the mass of the diffusing test particle, ζ is the friction coefficient, *U(x)* is the potential energy, and *K(t)* is the memory function which is related to the statistics of the random variable *θ(t)* by the fluctuation-dissipation theorem, i.e. $\zeta k_B T K(|t-t'|) = \langle \theta(t)\theta(t') \rangle$, with k_B Boltzmann's constant, and *T* the temperature. The fractional Langevin equation (fLE) is a specific case of the GLE where the noise term *θ(t)* is modeled as fractional Gaussian noise (fGn), i.e. long-range, power law correlated Gaussian random variables [109] with zero mean [37]. In this case the memory function becomes [37] $K(|t-t'|) = 2H(2H-1)|t-t'|^{2H-2}$ where *H* is the Hurst exponent, a measure of the temporal correlations in the noise and is defined in the range $1/2 < H < 1$ [37]. The fractional Langevin equation can be transformed into an equivalent equation for the probability density function *P(x,t)* which can be shown is given by a Fokker-Planck equation with time-dependent drift and diffusion coefficients [36,37,110]. This is in contrast to fractional dynamics which is described by the fractional Fokker-Planck equation (equation 9) highlighting a fundamental difference between these two approaches in their characterization of anomalous diffusion. Nevertheless the fractional Langevin equation describes subdiffusive dynamics and can yield FPT distributions that decay as a power law at long times [37], consistent with our findings (figure 3-1b).

In the context of DNA translocation, simulation studies of the translocation of long DNA molecules reveals subdiffusive transport stemming from the dynamics of the DNA chain whose interconnected parts have different rates of relaxation [37,111,112,113]. Diffusion in this case, which translates into temporal correlations in the random fluctuating forces of the dissipative medium [37], is very well described by the fLE where the use of fGn, rather than simple white noise, is intended to mimic the long memory fluctuations expected to be present

in a system with multiple timescales of relaxation [37,110]. In the unbiased and low-drift regime, for example, theory and simulations by Panja *et. al.* show that local strains induced in the polymer as monomers move across the pore, stretching the polymer on one side of the pore and compressing it on the other, can be expressed in terms of the polymer fluctuation modes and corresponding relaxation times [112,113]. In this case, it can be shown that the strain decays as a power law in time resulting in anomalous translocation dynamics [112,114].

While these studies show that the complex relaxation dynamics of DNA leads to anomalous subdiffusion of long DNA molecules through nanopores [111,112], it is unclear the degree to which these same effects effect the translocation dynamics of short polymers [31,115]. Indeed, Lubensky *et. al.* show that for short ssDNA (~< 300 nucleotides in length) the polymer relaxes quickly such that relaxation effects may be ignored in the analysis of translocation dynamics [31], a result that is in part corroborated by Brownian dynamics simulations of short ssDNA translocation through nanopores [115]. Nevertheless, the role of DNA relaxation dynamics in translocation and a corresponding fLE treatment of the translocation process is an important hypothetical interpretation and model of our experimental results on the translocation dynamics of short ssDNA (figure 3-1b).

3.2 Investigating the origins of subdiffusive dynamics of DNA escape

Our results on the escape dynamics of ssDNA from α-HL (figure 3-1b) demonstrate that low velocity motion of short DNA molecules through nanopores is subdiffusive. We have

discussed two possible models of DNA translocation to explain these results, namely: Fractional dynamics arising from stochastic binding interactions between DNA and the pore and a model based on the fractional Langevin equation arising from the complex relaxation dynamics of DNA. Both models are consistent with our data [18,37]. Nevertheless, these models are characterized by fundamentally different phenomenological equations. Determining the appropriate model, therefore, is critically important to understanding and characterizing the translocation process.

In this regard, the relaxation dynamics of DNA (with respect to the fLE model of translocation), to first-order, can be determined from coarse-grained models of polymer dynamics, i.e. the Rouse and Zimm models, in which the individual monomers are identical and sequence effects, in the case of DNA, are of higher-order [107]. Anomalous translocation dynamics arising from the relaxation dynamics of DNA are therefore expected to be sequence independent, to first-order, being determined largely by the dynamics of the polymer chain. In contrast, a fractional dynamics model of translocation based on binding interactions between DNA and the pore is expected to be sequence dependent due to differing association constants between individual nucleotides and the amino acids that decorate the internal surface of the α-HL pore, which, in some cases, can differ by an order of magnitude [116]. Exploring the effect of sequence on the escape dynamics of DNA, therefore, should provide significant insight on the origins of anomalous subdiffusion in DNA translocation. Given this we sought to investigate the effect of sequence on the escape dynamics of DNA.

3.2.1 Sequence effects on DNA escape dynamics

In exploring the sequence effects on DNA escape we characterize the escape dynamics of short, avidin-bound DNA homopolymers composed of adenine and cytosine from α-HL, and show that DNA escape dynamics are highly sequence dependent. In particular we show that the escape timescale of a homopolymer of adenine is ~2 orders of magnitude faster than that of cytosine indicative of sequence dependent interactions between DNA and the pore. We develop a three state model of DNA escape that involves stochastic binding events between DNA and the pore and show, through Monte Carlo simulation, that it accurately reproduces the experimental data. Our results therefore indicate that the origin of subdiffusive dynamics of DNA escape stem from binding interactions between DNA and the pore consistent with a model of DNA translocation based on fractional dynamics.

Through our model we characterize the escape and binding kinetics of poly-dA and poly-dC from α-HL and present methods for characterizing differences in their interaction energy with the pore. Our results show that the interaction strength of adenine with α-HL is substantially lower than that of cytosine. Lastly, we explore the effect on poly-dC escape dynamics upon substitution of a subsection of its bases with 5-methylcytosine. Although we observe no change in dynamics beyond the error bars of our measurements we show that discrimination between 5-methylcytosine and cytosine can be achieved based on differences in ionic current through wild-type α-HL. This differs from previous studies, which discriminated between these two nucleotides using mutant α-HL [95,117], and thereby shows that in principle nanopore sequencing approaches based on wild-type α-HL can also potentially be used for methylation sequencing which points to the potential application of this technology in cancer diagnostics and monitoring.

In addition to providing significant insight on the subdiffusive dynamics of DNA translocation, our results, which demonstrate that low velocity DNA motion through nanopores is highly sequence dependent, point to the possibility of using nucleotide dwell-time in the pore as a second metric by which to distinguish individual nucleotides in nanopore sequencing (in addition to nucleotide blockage current), opening new avenues by which to improve the overall design and optimization of nanopore sequencing methods.

3.2.1.1 Materials and methods

In similarity to our previous work [118] we perform DNA escape experiments with multiple α-hemolysin pores in parallel. By performing synchronous parallel single molecule measurements, we significantly reduce the time to acquire statistically relevant data by an amount proportional to the number of pores, thereby obtaining high quality data with only a few measurements, as compared to hundreds of consecutive single molecule measurements. In contrast to escape experiments with a single pore, whereby the measured current is in one of two states (capture or escape), the signal extracted in a multi-pore experiment rises steadily as an increasing number of DNA molecules escape from the pore.

We electrophoretically insert single stranded DNA (ssDNA) molecules, each coupled to avidin, into α-HL nanopores (one DNA molecule per pore) and subsequently allow them to thermally escape against an applied electric potential (figure 3-2).

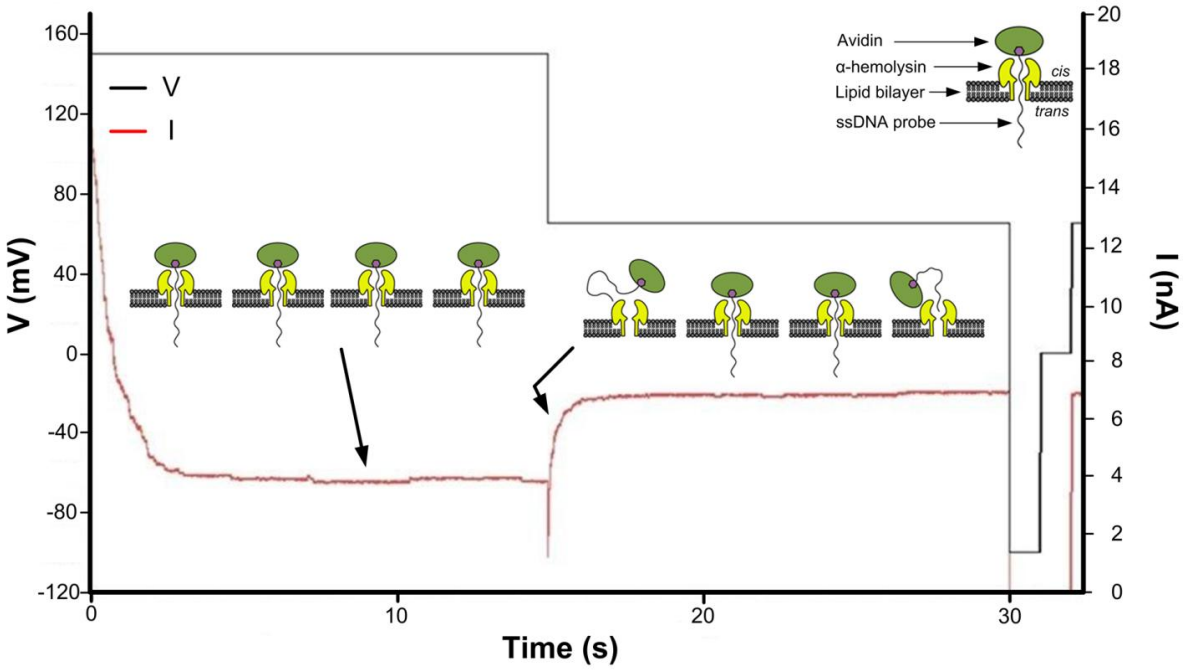


Figure 3-2. Multi-nanopore DNA escape. Application of a capture voltage (+150mV) leads to the capture and trapping of probes, resulting in the gradual blockage of the pores and decrease in the ionic current. After sufficient time to allow for >95% of the pores to be filled, each with a single DNA molecule, the voltage is reduced to the escape potential (+65mV shown), allowing for thermally activated escape of DNA from the pore against the electrostatic energy barrier. After sufficient time such that >99% of probes have escaped from the pore, the voltage is then reduced to a large negative potential (-100mV shown) forcing any remaining probes out of the pore, followed by a return to 0mV for ~1s and a return to the escape voltage (+65mV shown) for ~0.5s to determine I_{final}. The current through the pore as a function of time during escape is used to determine the likelihood that DNA is in the pore at time *t* upon application of the escape potential (i.e. the survival probability of DNA in the pore during escape).

The α-HL pores are formed using a method adapted from that of Akeson *et. al.* [22]. Briefly, a black lipid membrane of 1,2-diphytanoyl-sn-glycero-3-phosphocholine (Avanti, Alabaster AL) and hexadecene (Sigma-Aldrich, St. Louis MO) is formed across a 25μm PTFE aperture connecting two baths filled with 1M KCl, 10mM HEPES, pH 8.0 solution. Details on lipid bilayer formation can be found in Appendix C and that of data collection is described elsewhere [119]. An Axon Axopatch 200B patch clamp amplifier (Molecular

Devices, Sunnyvale CA) is used to measure the ionic current. Data is low-pass filtered at 10kHz by a 4-pole Bessel filter and sampled at 100kHz. Unless otherwise stated, DNA oligomers were synthesized by Integrated DNA Technologies (Coralville, IA) to be 27 nucleotides in length (approximately the full length of α-HL [18]), single stranded, biotinylated at the 5' end and coupled to avidin (all oligomers were HPLC purified). These complexes are for our purposes referred to as probes. Different probes used in this work included a homopolymer of adenine (dA27) and a homopolymer of cytosine (dC27). DNA escape experiments are conducted with 50 – 100 pores incorporated into the bilayer by injection of free subunits into solution which subsequently self-assemble into heptameric, membrane-spanning pores. Confirmation of the number of pores is achieved by applying a +100mV electric potential across the membrane (*trans* side positive) and observing the stepwise increase in the current (~97pA per pore), indicative of single pore incorporations in the bilayer. Probes are driven into the *cis* side of the α-HL nanopores (figure 3-2 – inset) by applying a +150mV electric potential across the bilayer for 15s, at which time >95% of the pores have ssDNA trapped within them (calculated based on the single pore IV-curves for an open and DNA occupied pore [118]). Avidin acts as an anchor preventing the translocation of DNA, thus ensuring that DNA exits on the *cis* side of the pore. Note that at 150mV the likelihood of dissociating the biotin-avidin complex is extremely low [120]. The electric potential is subsequently reduced to the escape potential (+65mV or +75mV) for 15s (which is ~5000 times the characteristic timescale for escape from α-HL for dA27 at +65mV – table 3-1). A typical parallel single-molecule current trace of escape events is shown in figure 3-2. From the raw current trace during thermal escape, the survival probability as function of time, $P_s(t)$, is determined, which is the likelihood that a DNA molecule is still present in the

pore at time *t* upon application of the escape potential (the survival probability is the integral of the FPT distribution for escape [32,36]). Specifically,

$$P_s(t) = 1 - \frac{I(t) - I_0}{I_{final} - I_0} \tag{12}$$

where $I(t)$ is the total ionic current through all pores at time *t* during escape, I_0 is the total current through the pores at the onset of escape ($t = 0$), and I_{final} is the total open-channel current through all pores at the escape voltage. I_0 must be calculated because its value is lost in the capacitive current spike that ensues upon changing the voltage from the capture potential (+150mV) to the escape potential. It can be calculated, however, from the α-HL IV-curves [118], the total number of pores, and the total current through all pores just prior to initiating the escape process. I_{final} is determined empirically at the end of the cycle of voltages (figure 3-2). Refer to [119] for further details on how to calculate I_0. Only the first 95% of probe escapes are analyzed (i.e. to $\log P_s(t) \sim -1.3$) to avoid the significant influence of current noise when the number of occupied pores becomes small[*]. Between 3000 and 10,000 single molecule escape events are recorded to accurately determine the survival probability for any given polynucleotide. Given that the DNA escape experiments are performed with multiple pores in parallel, and that probe molecules are present in solution on the *cis* side of the pores, it is possible that during the escape process, ssDNA in solution can be captured

[*] N. B. We find that the standard deviation in the total open channel current through all pores is ~2% of the total current at 100mV. In the case of 60 pores this amounts to a standard deviation in the current of ~115pA which is equivalent to the current through ~7 blocked pores (or ~1 open pore) at 100mV.

into empty pores (that were previously occupied) under the escape potential. We find that this should have a negligible effect on the calculated survival probability for dA27 at both +65mV and +75mV and for dC27 at +65mV. This is based on capture rate data for dC30 at +65mV (~0.0305/s per pore) and at +75mV (~0.064/s per pore) given in [121], which has been determined under roughly the same conditions (i.e. salt concentration, DNA concentration, temperature, etc…) as reported here. For 100 pores at a +65mV escape potential (for example) this translates into a capture rate of ~3/s. At +65mV the data, for dC27, is fit over an experimental time of ~1.5s (i.e. to $logP_s(t) \sim -1.3$), therefore a crude estimate for the number of captures is ~5 during the experiment time. It should be emphasized however, that the estimate of 5 captures assumes that at $t = 0$ all pores are open, which is not the case (i.e. at $t = 0$ almost all pores, >95%, are DNA occupied). Therefore the true number of expected captures during the relevant portion of the experiment time (~1.5s) is expected to be much lower than 5 (a similar calculation can be carried out for dA27 at +65mV and at +75mV). In any case, an error bar of 5 pores in the current decay does not affect our results, and is comparable with our inherent current noise. For dC27 at +75mV, however, it is possible that captures during the escape process have an observable influence on the calculated $P_s(t)$, and that the +75mV data should be viewed with this in mind; i.e. the dC27 +75mV data is an upper limit on survival probability and likely overestimates the escape time. Given that our conclusions are primarily based on +65mV data, however, this does not alter our conclusions with respect to sequence dependence of escape dynamics.

Detection and differentiation between 5-methylcytosine and cytosine was accomplished by electrophoretically trapping a probe molecule into a pore and recording the absolute ionic

current traversing the blocked pore for 1s. Consequently, these experiments were performed on a single pore, as opposed to multiple pores as described above. Two polynucleotides were used for comparison: a homopolymer of cytosine (dC27) and a heteropolymer consisting of cytosine and 5-methylcytosine (dCdmC27). Specifically nucleotides 10 to 14 (inclusive, measured from the 5' to 3' end) of dCdmC27 were 5-methylcytosine, while the remaining nucleotides were cytosine. Bases 10 – 14 were chosen to be methylated since it is likely, given the internal dimensions of α-hemolysin [72], that these nucleotides span the limiting aperture of α-HL when the molecule is driven into and trapped in the pore. We expect the greatest occlusion of the ionic current to occur at this location. Between 2000 and 3000 single molecule capture events were recorded. Histograms of the ionic current through the pore were then determined for both polynucleotides. All experiments were performed at room temperature.

3.2.1.2 Results and discussion

Figure 3-3 shows the survival probability for escape of dA27 and dC27 probes from α-HL at an escape potential of +65mV. Consistent with previous observations on the escape dynamics of poly-dA [18], the survival probability for both dA27 and dC27 decay as a power law at long times (>~0.008s in the case of dA27 and >~0.5s for dC27), demonstrating subdiffusive dynamics of DNA escape.

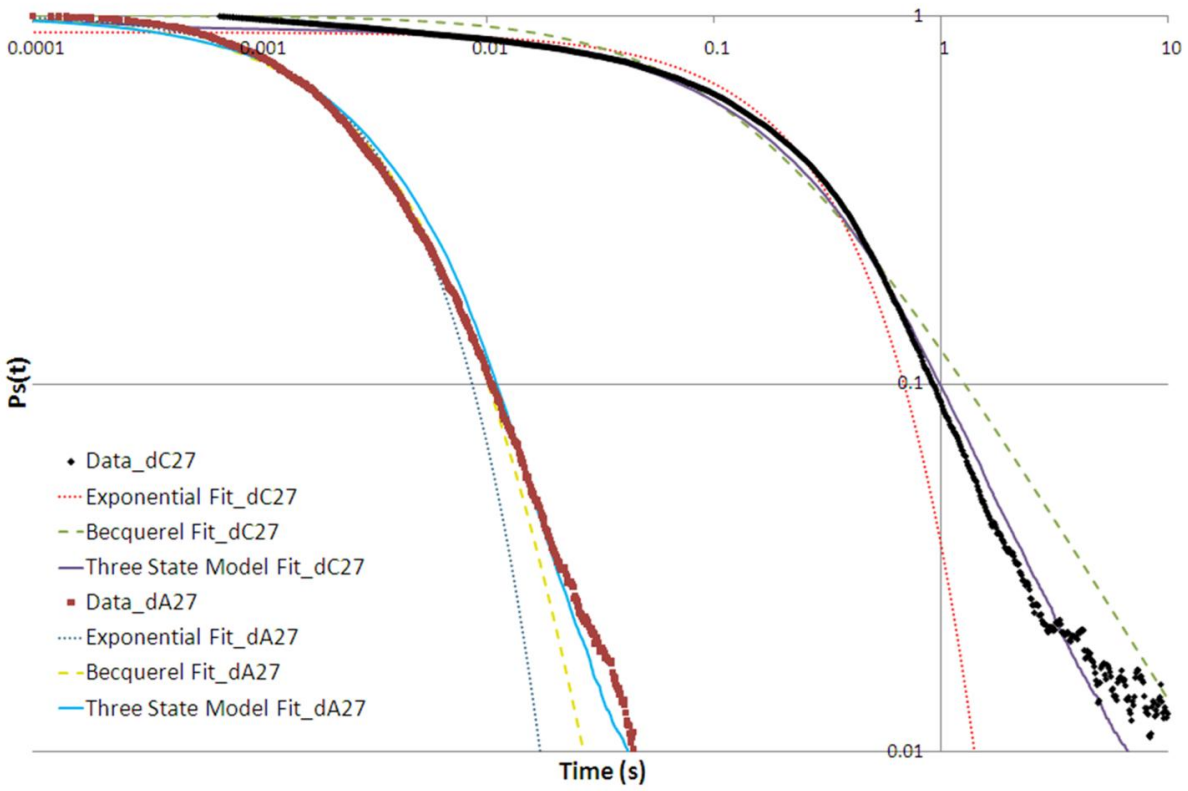


Figure 3-3. Log-log plot of the survival probability vs. time. The survival probability of DNA in the pore (upon application of the escape potential) vs. time, for dA27 (>3000 single molecule events) and dC27 (>5000 single molecule events) at an escape voltage of +65mV. The large number of events comprising each dataset ensures that long escape events are statistically significant. Also shown are the corresponding fits to the data from the Becquerel, single exponential and three state model fits. In both cases (dA27 and dC27) the exponential fit is a poor fitting function to the data. With respect to the Becquerel decay law, although it fits the dA27 data reasonably well, it is a poor fitting function to the dC27 data. In contrast our three state model fit forms an excellent fit, particularly at long times, to both the dA27 and dC27 data sets. All fits require 4 parameters for two data sets (in the case of the exponential fit a prefactor is used which is itself a free parameter).

The timescale for escape from the pore for dC27 is greater by more than an order of magnitude relative to dA27 (table 3-1), highlighting the highly sequence dependent dynamics of DNA escape, indicative of sequence dependent interactions between DNA and the pore. Our results therefore point to fractional dynamics as an appropriate modeling approach of the translocation process whereby subdiffusive transport arises from binding interactions

between DNA and the pore. To explore this hypothesis further we developed a physical model of the DNA escape process that involves binding and sequence dependent unbinding of DNA to (from) the pore. It is a three state model adapted from Bates *et. al.* [34] where DNA in the pore is in either a free state (i.e. not bound to the pore) or a bound state (i.e. bound to the pore) and escapes from the pore from the free state (figure 3-4). In our model both the escape rate and binding rate are constant whereas the unbinding rate is a uniformly distributed random variable. To see this note that state S_B in the model incorporates all possible bound states of DNA to the pore. DNA, however, can bind to the pore in different configurations with corresponding differences in binding strength (or energy barriers for unbinding). Moreover, the inner walls of the α-HL pore are decorated with a variety of amino acids with differing binding affinities to the DNA backbone and its nucleotides. This variability in binding strength can therefore be accounted for by distributing the unbinding rate. With respect to the choice of distribution, it has previously been shown that a uniform distribution of exit rates from a given state yields a dwell time distribution in said state that decays as a power law [122,123]. A uniform distribution of unbinding rates, therefore, translates into a power law distributed waiting time in the bound state, thereby making our three state model consistent with a model of DNA escape based on fractional dynamics.

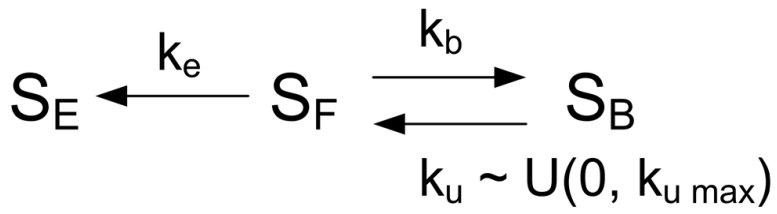


Figure 3-4. Model of the DNA escape process from nanometer-scale pores. Model for the DNA escape process where S_F represents the free state of DNA (i.e. DNA not bound to the pore), S_B the bound state (i.e. when DNA is bound to the pore) and S_E the state in which DNA has escaped from the pore. Our model invokes a constant escape rate k_e, and binding rate k_b, and a uniformly distributed unbinding rate k_u ranging from *0* to $k_{u\,max}$ (i.e. $k_u \sim U(0, k_{u\,max})$), where $k_{u\,max}$ is a sequence dependent rate (as opposed to the binding rate and escape rate which are both sequence independent).

In addition to being distributed, the unbinding rate is also sequence dependent, being largely determined by the binding strength of DNA to the pore, which is a function of the specific nucleotide composition. The escape rate and binding rate, however, are independent of sequence. The electrostatic energy barrier for escape is determined by the effective charge of the DNA backbone in the pore and the electric potential applied across the pore, both of which are sequence independent. With respect to binding, we expect this to be primarily a normal diffusive process, and therefore unaffected by sequence. However it should be emphasized that a sequence independent binding rate is a first-order approximation, which ignores differences in nucleotide size between purines and pyrimidines, and presumes that sequence does not influence (or influences to second-order) the diffusive properties of DNA in a confined volume.

To compare our model predictions with the experimental data we performed a Monte Carlo simulation of the three state model (figure 3-4) and characterized the resulting survival probability of DNA in the pore. In particular, our model of the escape process models the

escape of a single DNA oligomer from the pore. A single escape event begins in the free state S_F and ends upon entering the escape state S_E. The total escape time t_{esc} for a single event is given by the sum of the total dwell time in the free state t_{free} and the total dwell time in bound state t_{bound}, i.e.

$$t_{esc} = t_{free} + t_{bound} \tag{13}$$

both t_{free} and t_{bound} are the sums of all the individual dwell times in the free state and bound state respectively. Hence,

$$\begin{aligned} t_{free} &= \sum_i t_f^i \\ t_{bound} &= \sum_i t_b^i \end{aligned} \tag{14}$$

The individual dwell time in the free state t_f is exponentially distributed (i.e. the survival probability in the free state exhibits first-order kinetics) with a transition rate out of the state given by the sum of the binding rate k_b and escape rate k_e. i.e.

$$f(t_f,\, k_e + k_b) = (k_e + k_b)\exp\left[-(k_e + k_b)t_f\right] \tag{15}$$

where $f(t_f, k_e + k_b)$ is the individual dwell time distribution in the free state. The probability of entering the escape state S_E upon transitioning out of the free state is given by,

$$p_E = k_e / k_e + k_b \quad (16)$$

The individual dwell time in the bound state t_b is also exponentially distributed with a rate however, that is uniformly distributed along the interval [0, $k_{u\,max}$].

Thus, beginning in the free state, t_f is drawn from the distribution given by equation 15. To determine whether the DNA oligomer has entered the escape state or the bound state a uniform random variate along the interval [0, 1] is generated. If this number is less than p_E then the DNA oligomer has escaped from the pore (i.e. has entered the escape state). If however, the converse is true (i.e. the number is greater then p_E) then the oligomer has bound to the pore (i.e. has entered the bound state). In the bound state a uniform random variate representing the unbinding rate for that particular instance is drawn. t_b is then drawn from an exponential distribution with said unbinding rate, thereby transitioning the system back to the free state. The simulation continues until the oligomer transitions to the escape state at which point the total escape time t_{esc} is determined and recorded.

The simulation is run N = 32000 times for both dC27 and dA27, where N represents the total number of DNA escape events. From the generated escape time distribution the survival probability of DNA in the pore is then determined as outlined in [119].

65 mV	k_b (1/s)	k_e (1/s)	$k_{u\ max}$ (1/s)	τ_{escape} (s)
dA27	4570	389	43360	0.0035
dC27			210	0.1405
75 mV	k_b (1/s)	k_e (1/s)	$k_{u\ max}$ (1/s)	τ_{escape} (s)
dA27	1560	127	3005	0.0201
dC27			5.65	4.6065

Table 3-1. Best fit kinetic rates for dA27 and dC27 at +65mV and +75mV escape voltage. Rates are obtained from a Monte Carlo simulation of the DNA escape model presented in figure 3-4. The binding rate k_b and escape rate k_e are both sequence independent quantities whereas the maximum unbinding rate, $k_{u\ max}$, is sequence dependent. The rate of unbinding is ~2 – 3 orders of magnitude higher in the case of dA27 vs. dC27 indicative of a weaker interaction with the α-HL nanopore in the case of adenine relative to cytosine as shown in figure 3-5 (refer to text regarding discussion on the trend of decreasing rates with increasing escape voltage). Rates obtained at +75mV involve fitting to a partial dC27 data set. τ_{escape} is the timescale for escape obtained from the Becquerel fit to the data as defined in [18].

Results from the Monte Carlo simulation are shown in figure 3-3. In order to fit both the dA27 and dC27 data sets simultaneously four parameters are required for the simulation: k_e, k_b, $k_{u\ max,\ dA}$, $k_{u\ max,\ dC}$ (the escape rate and binding rate are presumed the same for both dA27 and dC27). As can be seen from the figure the corresponding fits, obtained from the simulation, are in good agreement with both data sets at short and (more importantly) at long times. These results support the conclusion that the origin of subdiffusive dynamics of DNA escape stem from stochastic binding interactions between DNA and the pore consistent with a model of DNA translocation based on fractional dynamics. Also shown in figure 3-3 are the corresponding fits to the data for both the exponential and Becquerel decay laws. The exponential and Becquerel fits contain the same number of fitting parameters as our model to fit two data sets (in the case of the exponential fit a prefactor is used during fitting which is itself a free parameter). The exponential fit is a poor fitting function to both the dA27 and

dC27 data sets, failing to reproduce the DNA escape data at long times. With respect to the Becquerel fit, although it fits the dA27 data well, it provides a poor fit to the dC27 data. The corresponding rates, determined from our simulation, are shown in table 3-1 for escape at a +65mV and +75mV applied potential. It is interesting to note that all rates decrease with increasing voltage. A decreasing escape rate with increasing voltage is not surprising and is a result of a higher electrostatic energy barrier for escape. We reason that a decreasing binding rate with increasing voltage may be understood by considering that the electric force acting on trapped DNA will tend to stretch the polynucleotide, thus potentially hindering its ability to adopt conformations that allow it to explore the inner volume of the pore. This may translate into a reduced rate of encounter with binding sites in the pore, resulting in an overall decrease in the binding rate. The reason for a decreasing unbinding rate with increasing voltage, however, is not immediately clear and may be a result of the applied potential affecting either the energy barrier or the attempt rate (k_D in the Arrhenius relationship) for unbinding (or both).

Given the sequence dependent DNA escape kinetics (table 3-1) we can characterize the differences in interaction energy between dA27 and dC27 with the pore. In particular a distribution of kinetic rates $H(k)$ yields a distribution of activation energies $g(E_b)$ [122,124]. In our case a distribution of unbinding rates translates into a distribution of activation energies for unbinding. The relationship between $H(k)$ and $g(E_b)$ is,

$$-H(k)dk = g(E_b)dE_b \qquad (17)$$

which is derived elsewhere [124]. The fundamental relationship relating the rate of transitioning from one state to the next with the energy barrier height that separates the two states is the Arrhenius relationship,

$$k = k_D \exp\left(-E_b / K_b T\right) = 1/\tau \tag{18}$$

where k_D is the attempt rate for barrier crossing, and τ is the timescale for state transitioning. Using equations (17) and (18) an expression for $g(E_b)$ can be determined,

$$g(E_b) = \frac{H(k)k}{K_b T} = g\left(K_b T \ln(\tau)\right) \tag{19}$$

where K_b is Boltzmann's constant, and T is the absolute temperature. This can, alternatively be expressed as,

$$g\left(\frac{E_b}{K_b T}\right) = \frac{H(k)}{\tau} = g\left(\ln(\tau)\right) \tag{20}$$

For a uniform distribution of unbinding rates $H(k)$ along the interval [0, $k_{u\,max}$] we have,

$$H(k) = 1/k_{u\max} = \tau_{u\min} \tag{21}$$

where $\tau_{u\ min}$ is the shortest timescale for unbinding. Thus,

$$g(\ln(\tau)) = H(k)\exp[-\ln(\tau)] = \exp[-(\ln(\tau) - \ln(\tau_{u\min}))] \tag{22}$$

Figure 3-5 shows the distribution of energy barrier heights for unbinding $g(E_b/K_bT)$ for dA27 and dC27, at +65mV, expressed as the distribution $g(ln(\tau))$. Without knowledge of the attempt rate, k_D, the absolute energy barrier height cannot be directly determined. However the random variables $ln(\tau)$ and E_b/K_bT are related by a constant, $ln(\tau_d)$ (i.e. $ln(\tau) - ln(\tau_D) = E_b/K_bT$, where $\tau_D = 1/k_D$). Thus subtracting $ln(\tau_D)$ from $ln(\tau)$ simply shifts the location of the distribution on the x-axis, while preserving its shape exactly. Therefore, we can determine relative differences in energy barrier heights from $g(ln(\tau))$, despite being unable to determine absolute activation energies. The distribution $g(ln(\tau))$, and thus $g(E_b/K_bT)$, is a shifted exponential distribution (equation 22), and is shifted by the logarithm of the shortest timescale for unbinding i.e. $ln(\tau_{u\ min}) = ln(1/k_{u\ max})$. Thus, in the case of $g(ln(\tau))$ timescales for unbinding that are shorter then $\tau_{u\ min}$ are not possible. In the case of the distribution of energy barrier heights, $g(E_b/K_bT)$, the shift amount is the minimum energy barrier height for unbinding $E_{b\ min}$ for a given polynucleotide. Thus energy barrier heights for unbinding that are smaller than $E_{b\ min}$ are not possible ($\tau_{u\ min}$ and $E_{b\ min}$ are related through the Arrhenius relationship, $\tau_{u\min} = \tau_D \exp\left(E_{b\min}/K_bT\right)$). The average energy barrier height for unbinding

for dC27 is ~5.33 K_bT higher (at +65mV) relative to dA27 (assuming equal attempt rates, k_D in the Arrhenius relationship, for both dA27 and dC27), which indicates that the binding strength of adenine to the α-HL pore is substantially lower than that of cytosine. It is interesting to note that the association constant of cytosine for lysine at room temperature and neutral pH is ~10 fold greater than that of adenosine [116]. Thus it is possible that a nucleotide specific interaction with the ring of lysines that forms the limiting aperture of α-HL may contribute to the observed difference in unbinding energy / unbinding rate between dA27 and dC27. Further work involving DNA escape experiments with heteropolymers composed of varying proportions and configurations of cytosine and adenine, and / or site-directed mutagenesis of α-HL may elucidate the role that the lysines play in governing the kinetics of DNA escape. At +75mV the difference in the activation energy for unbinding between dC27 and dA27 is ~6.28 K_bT (distributions not shown). The variability in the energy barrier heights for unbinding for both dA27 and dC27 as determined by the standard deviation of $g(ln(\tau))$ is 1 K_bT, for both polynucleotides, independent of voltage and $k_{u\ max}$, which is a direct mathematical consequence of a uniform distribution of unbinding rates.

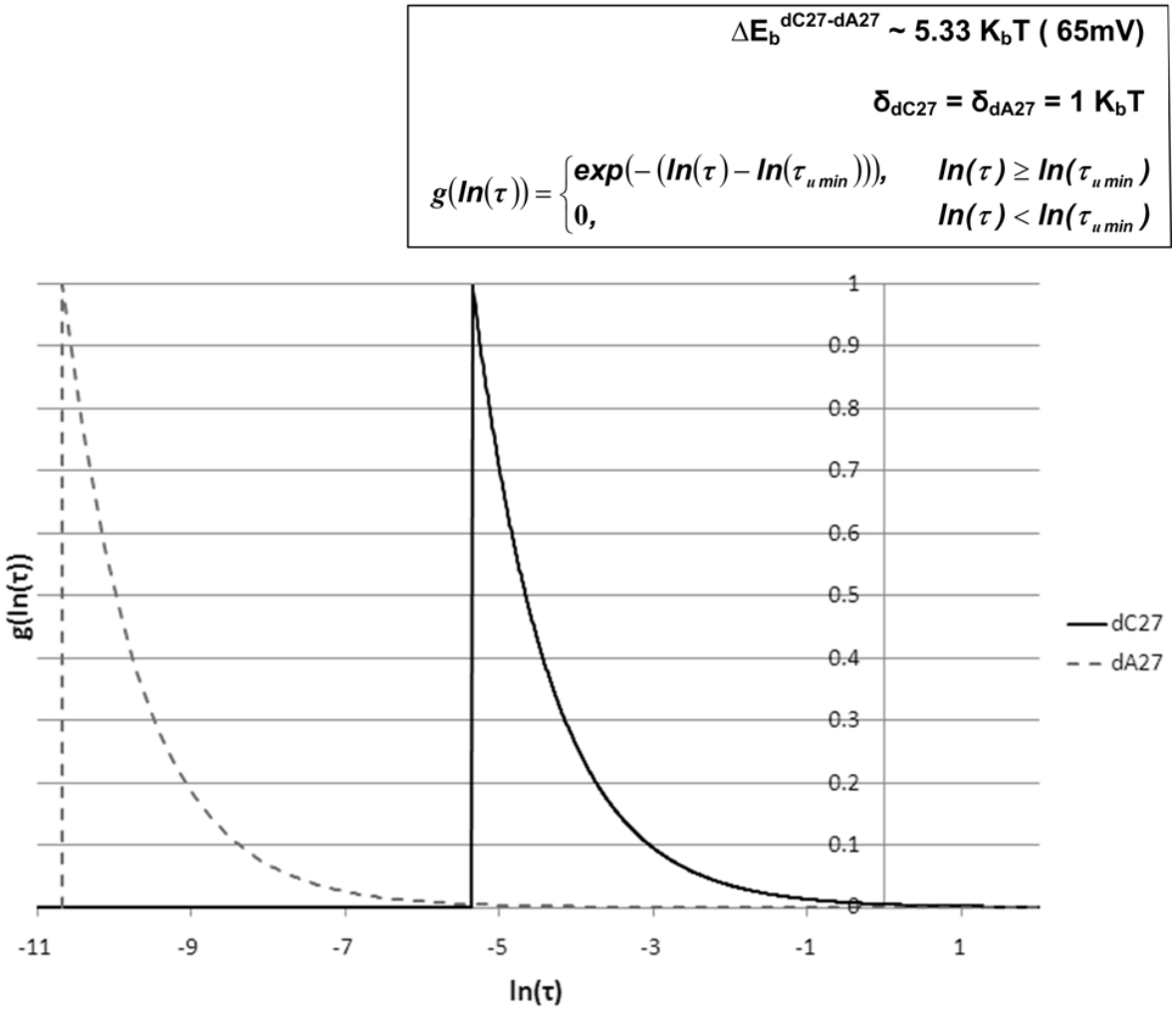


Figure 3-5. Activation energy distribution *g(E$_b$/K$_b$T)* for unbinding from the pore expressed as *g(ln(τ))*. Shown is *g(ln(τ))* for both dA27 (dashed line) and dC27 (solid line) at an escape voltage of +65mV. *g(ln(τ))* and *g(E$_b$/K$_b$T)* differ from each other only by their location along the x-axis – refer to text. The distribution, *g(ln(τ))*, for both dA27 and dC27 is a shifted exponential distribution (inset) shifted by the logarithm of the shortest timescale for unbinding (*ln(τ$_{u\,min}$)* where *τ$_{u\,min}$ = 1/k$_{u\,max}$*). In the case of *g(E$_b$/K$_b$T)* the shift amount is the minimum energy barrier height for unbinding. **Inset:** The average energy barrier height for unbinding is ~5.33 *K$_b$T* higher for dC27 relative to dA27, which indicates a greater binding interaction / strength of cytosine to the α-HL pore relative to adenine. The standard deviation of each distribution is 1 *K$_b$T*.

Interestingly, our observation that poly-dA escapes from α-HL faster than poly-dC is opposite to what might be expected based on studies of high-speed ssDNA translocation through α-HL. In this case it has been shown that a homopolymer of poly-dA translocates

slower than a homopolymer of poly-dC [23,69], i.e. poly-dA moves through the α-HL pore slower than poly-dC. Based on this we would expect that dA27 would escape slower than dC27. This discrepancy highlights an important difference between DNA motion when the motion is fast (microsecond timescale in high-speed translocation) versus when it is slow (millisecond timescale during escape or low-velocity translocation). The binding rate of DNA to the pore is in the millisecond timescale (table 3-1). During high-speed translocation, nucleotides move through the pore in a timescale of 1 – 20μs [22,23,29], which is too fast for a given nucleotide to have a reasonable likelihood of binding to the pore. Thus in uncontrolled high-speed translocation the probability of binding is low and the effects reported here should have a minimal effect on the translocation process. In this case we expect translocation dynamics that are closer to Markovian, specifically for short DNA strands where relaxation effects may be ignored [31] (see section 3.1.2). Indeed, in experimental studies of high speed translocation of short ssDNA molecules through α-HL the translocation time distribution is reported to decay exponentially at long times [23,69], with a characteristic translocation time that scales linearly with polymer length [21], consistent with Markovian dynamics and a 1D normal diffusive model of DNA translocation [107].

When the motion of DNA is slowed, however, from microseconds to milliseconds sequence dependent DNA-pore binding interactions significantly affect translocation dynamics giving rise to anomalous transport of DNA through the pore (i.e. subdiffusive translocation dynamics). Thus we highlight an important difference in the process by which DNA moves through an α-HL nanopore when the motion is fast (microsecond timescale) as compared to when it is slow (millisecond timescale) and dominated by binding interactions with the pore.

In addition to investigating differences in DNA escape dynamics between poly-dA and poly-dC from α-HL, we also studied the effect on poly-dC escape dynamics upon substitution of a subsection of its bases with 5-methylcytosine (dmC), by comparing escape dynamics between dC27 and dCdmC27. Our results (not shown) were inconclusive. Under the given experimental conditions we could not detect a difference in escape dynamics (based on the experimentally derived survival probability curves for each polynucleotide) between dC27 and dCdmC27 greater than the error bars on our measurements. A potential source of error in our measurements is room temperature fluctuations ($\sim 21°C\ ^{+3°C}_{-1°C}$) and the experiment temperature drift over time (~0.023°C/min)[†], both of which may have an observable effect on escape dynamics. Given that methylation is a minor structural modification to dC, particularly in comparison to dC vs. dA, and is therefore expected to have a subtle effect on escape dynamics (if at all) it is possible that with more accurate temperature control some effect of methylation might be observed. Our results indicate that the escape timescale for dCdmC27 lies within the range τ_{escape} = 0.1405s ± 0.06s at an escape potential of +65mV (0.1405s being the timescale for escape of dC27 – table 3-1). Despite our inability to reproducibly detect a difference in escape dynamics between dC27 and dCdmC27, we did however notice a clear difference in the absolute blockage current through the pore between the two polynucleotides (figure 3-6). Consistent with previous observations [95,117], our results show that dmC tends to occlude the ionic current to a greater degree then dC. These results demonstrate that unmodified, wild-type α-HL can distinguish between

[†] Reported temperatures were measured with our current experimental apparatus which includes temperature control to ± 0.1°C.

dmC and dC, based on blockage current alone, whereas previous explorations of this blockage current difference [95,117] were made using modified α-HL.

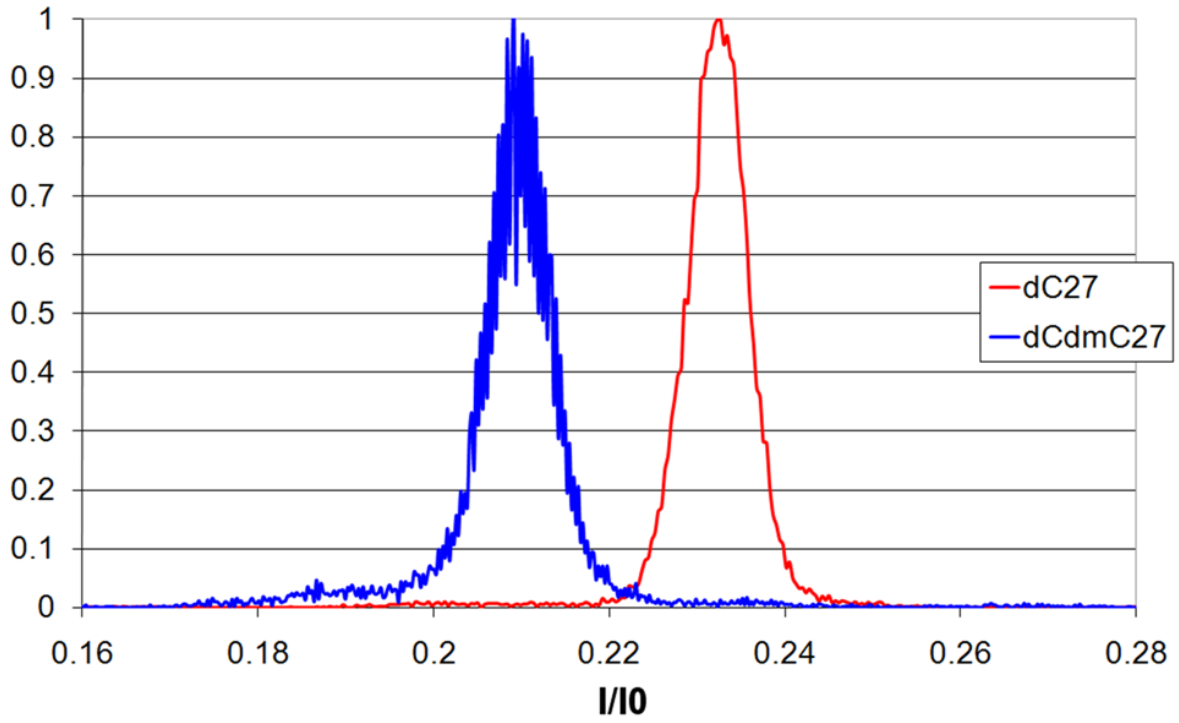


Figure 3-6. dC27 vs. dCdmC27 ionic current histograms. Shown are the ionic current histograms through a single α-HL pore upon capture and trapping of the dC27 vs. dCdmC27 polynucleotides (>2000 single molecule events comprise each histogram). The histograms are expressed as I/I_0, and are normalized to the peak of the distribution, respectively (I_0 is the open pore current - not shown). dC27 and dCdmC27 differ only by the presence of 5-methylcytosine at bases 10 – 14 (inclusive, measured from the 5' to 3' end) in the case of dCdmC27. Consistent with [95,117] these results show that dmC occludes the ionic current to a greater degree than dC. However, these results show that unmodified, wild-type α-HL can distinguish between dmC and dC by measurements of the blockage current alone.

3.2.1.2.1 Translocation regimes

Our work, in combination with theory, simulation and experimental studies by others on single stranded DNA translocation through narrow pores, indicates the presence of important

regimes in translocation where translocation dynamics is governed by different processes depending on DNA length and translocation velocity.

Short polymers, low velocity translocation

Our work shows that DNA motion in this regime is subdiffusive stemming from stochastic, sequence dependent binding interactions between DNA and the pore. In this case the appropriate phenomenological equation describing the translocation process is given by the fractional Fokker-Planck equation (equation 9).

Short polymers, high velocity translocation

Our results indicate that binding between DNA and the pore, in this regime, is unlikely. Translocation dynamics of short DNA molecules, for which relaxation effects may be ignored [31] (see section 3.1.2), should therefore approach Markovian, consistent with experimental findings of high speed, short ssDNA translocation through α-HL [21,23,69] which indicate Markovian translocation dynamics and a corresponding description of the translocation process by a 1D Fokker-Planck equation [107].

Long polymers, low velocity translocation

For long polymers translocating at low velocity our work in combination with theoretical and simulation studies by others indicate that translocation dynamics is likely determined by a complex interplay of multiple processes. In particular, we have shown that at low translocation velocity binding interactions between DNA and the pore play a significant role in translocation dynamics. Moreover, theory and simulations of DNA translocation through

non-interacting pores, in the unbiased and low-drift regime, show that DNA relaxation dynamics for long polymers, e.g. chain tension relaxations [112] (see section 3.1.2), also play an important role in translocation [111,112,113]. Given this, we expect both processes (i.e. binding interactions and relaxation effects) to have a governing effect on translocation dynamics. We speculate that DNA motion in this regime is subdiffusive, given that each process independently gives rise to subdiffusive translocation dynamics, however further experimental and / or simulation work is required to better understand and characterize translocation dynamics in this regime. We note here that these processes are governed by different phenomenological equations (as discussed previously). It is unclear, therefore, of the governing equation that might describe translocation dynamics in this regime where we expect a complex interplay between these processes to play out, however, we note that the theory of correlated CTRWs [125] may offer a potential way forward in this regard.

We note here, that this translocation regime is particularly relevant to nanopore sequencing with respect to achieving the full promise of this technology in terms of sequencing accuracy, cost and throughput (see chapters 1 and 2). In this regard, further studies on translocation dynamics will yield insights that may be important for improving the design and optimization of nanopore sequencing technology.

Long polymers, high velocity translocation

Theoretical analysis of driven translocation by Sakaue shows that translocation dynamics is subdiffusive and stems from the progressive response of the DNA chain to the driving field and consequent tension propagation along the chain backbone [126]. In this case the polymer translocation time is given by the characteristic time for the tensile force to reach the chain

end, which scales as $\tau \sim N^{1+\upsilon}$ [126] (for strong driving forces, where N is the number of monomers and υ is the Flory exponent) in agreement with simulation studies that demonstrate anomalous dynamics of forced translocation [127]. We note that in this regime DNA moves through the pore at high speed and therefore binding interactions with the pore, as reported here, should have a minimal effect on translocation dynamics.

We mention here two important points with respect to the above discussion. First, it is specific to single stranded DNA translocation through pores that are sufficiently narrow such that DNA translocates in single file. Analysis of DNA translocation through larger pores requires further considerations including the possibility of folded polymer conformations during translocation [86]. Second, we have avoided explicit definition of the polymer lengths and translocation velocities that separate the translocation regimes, given that these parameters depend significantly on the properties of the particular DNA-pore system under consideration. In general, however, the regimes (as described above) can be distinguished based on the degree to which DNA-pore binding interactions and the dynamical properties of the DNA chain, e.g. relaxation effects, effect translocation. In the case of single stranded DNA transport through α-HL our work indicates that translocation velocities that slow translocation rates to millisecond timescales per nucleotide (or slower), rates at which DNA-pore binding interactions play an important role in translocation, may be considered low velocity translocations. Similarly, Lubensky *et. al.* show that for DNA lengths up to ~300 nucleotides, in ssDNA translocation through α-HL, the polymer relaxes quickly such that relaxation effects in translocation may be ignored [31]. We therefore consider this to be a rough estimate of the crossover between short and long polymers.

3.3 Chapter conclusion

Emerging nanopore technologies for DNA analysis, including next-generation single molecule nanopore sequencing require that DNA translocate a nanometer-scale pore. Given this, understanding and characterizing the physics of DNA translocation through nanopores is critical to the design and optimization of these methods. We have presented detailed observations on the process by which short ssDNA moves through a nanometer-scale pore, at low velocity, through parallel observation of many ssDNA escape events from α-HL over an electrostatic energy barrier. Our results show that escape dynamics is subdiffusive and highly sequence dependent. We demonstrate that a three state model of the escape process involving stochastic sequence dependent binding interactions between DNA and the pore accurately reproduces the observed data, thereby showing that the origins of subdiffusive dynamics in translocation stem from DNA-pore binding interactions, consistent with a model of DNA translocation based on fractional dynamics. Our results indicate that as the DNA translocation velocity increases, such that nucleotides translocate the pore in microsecond timescales, the effect of DNA-pore interactions on translocation dynamics becomes minimal. In this case translocation dynamics should approach Markovian, consistent with experimental findings of high speed, short ssDNA translocation through α-HL [21,23,69].

Through our three state model we characterize the sequence dependent kinetics of the DNA escape process and present methods for characterizing the sequence dependent interaction energy between DNA and the pore. Given the highly sequence dependent dynamics of DNA translocation our results point to the possibility of using nucleotide dwell-time in the pore as a second metric by which to distinguish between individual nucleotides in

nanopore sequencing (in addition to nucleotide blockage current), a result that opens new avenues by which to improve the overall design and optimization of nanopore sequencing technology.

Our work, in combination with theory, experimental and simulation studies of DNA translocation by others indicates the presence of important regimes in ssDNA translocation through narrow pores in which translocation dynamics is governed by different processes depending on DNA length and translocation velocity. The analysis of translocation dynamics in this way, i.e. in terms of DNA length and velocity dependent regimes, provides a framework for organizing and interpreting the various results obtained in studies of DNA translocation dynamics.

Our work provides deeper insight into the complex process by which DNA moves through nanopores, understanding of which will enable the improved design and optimization of nanopore DNA analysis and sequencing technologies.

Chapter 4

Nanopore protein analysis

To date, nanopore sensing has been predominantly applied for DNA analysis due to its potential for low-cost, high-throughput DNA sequencing [29,30]. On the other hand, protein analysis with nanopores has only recently been undertaken, and is an area of active research and interest. The technique offers the opportunity to probe the biophysical properties of proteins at the single molecule level and has already shown much promise as a means by which to study the unfolding kinetics of proteins [40,41], characterize protein-pore interactions [42,43,44], and study the transport properties of proteins through pores [42,45,46]. It is therefore emerging as a new and powerful tool for studying protein biophysics, which in addition to potentially providing critically important insights could also lead to new biotechnology applications. This is particularly true for medical diagnostics, where nanopore technology offers the advantage of high sensitivity (i.e. target molecules can be detected at very low concentrations and at very low sample volumes) and minimal sample preparation (e.g. label-free detection) [29,30]. In this respect, nanopore analysis is an attractive candidate technology for highly sensitive protein and small molecule detection with potential applications in disease biomarker and pathogen detection [39].

In this chapter we begin with a brief review of nanopore analysis of proteins, highlighting the ability of this technique for probing their biophysical properties, and discuss its potential for investigating protein dynamics. We outline some challenges associated with nanopore protein analysis and briefly describe our methods to tackle these challenges. We

apply these methods to probe prion protein conformational dynamics inside a nanopore, a protein whose conversion into a misfolded isoform is associated with the pathogenesis of prion diseases in humans and animals. Before presenting our methods and results therefore, we provide a brief background on the prion protein and prion diseases.

4.1 Nanopore protein analysis: Review

The first demonstration of peptide translocation through a nanopore was carried out in 2004 with the transit of a series of peptides containing repeats of the sequence (Gly-Pro-Pro) through α-HL nanopores [128]. The authors demonstrated a correlation between the properties of the current during translocation (i.e. blockage amplitude and duration) and the structure of the translocating protein showing the ability to discriminate between single, double, and collagen-like triple helices based on translocation signatures [128]. These early results demonstrated the potential of the nanopore method for probing the structural and conformational properties of proteins at the single molecule level. Nanopore current is also a rich source of information on protein-pore interactions. Several studies have been carried out exploring interactions of polypeptides with wild-type α-HL, including α-helical [42,83,128] and β-hairpin [129] polypeptides. These investigations have shown that polypeptide translocation through α-HL significantly depends on the charge, length, folding and stability properties of the translocating molecule [49]. In this regard, a study combining both experimental work (nanopore sensing) and computational studies (Langevin dynamics simulations) showed that unfolded β-hairpins enter the pore in an extended conformation yielding fast translocation events [49,129], in contrast to structured β-hairpins, whose nanopore capture events occur slowly, and produce longer-lived current signatures [49,129].

Significant advances in understanding protein-pore interactions have come through protein engineering [130,131]. Chemical modification of the α-HL β-barrel can promote protein-pore interactions thereby influencing capture and translocation of proteins in the pore [132]. An insightful study on the effect of protein-pore interactions on nanopore capture and translocation of proteins involved probing the translocation of a protein containing a positively charged N-terminal sequence (similar to the peptide signal sequence that facilitates transport of proteins in the cell) and characterizing its interaction with negatively charged binding sites engineered in the α-HL channel [44,130]. The authors found that the electrostatic traps at the entry and exit of the pore significantly enhanced polypeptide capture and translocation [44], demonstrating the impact of protein-pore interactions on translocation kinetics. Moreover, equipping the engineered pore with aromatic binding sites was also shown to enhance capture and translocation of aromatic peptides while slowing down their translocation velocity [130,132]. The interaction affinity of negatively charged α-helical peptides with the pore can also be tuned by the structure and charge of the pore [83,130].

Single molecule analysis of protein folding-unfolding is another application in which nanopore sensing can yield important insights. In this regard, an early demonstration in 2007 involved observing protein denaturation via electrical detection with an α-HL nanopore as a function of the concentration of guanidinium-HCl (a chemical denaturing agent) [79]. The authors report observing both short and long blockades due to unfolded proteins and partially folded conformations respectively [79], demonstrating the potential of this technique for experimentally probing the protein unfolding pathway and the possibility of detecting rare events or intermediate states along this pathway that may otherwise be obscured by ensemble

methods. In a related study, the authors characterize event frequency as a function of guanidinium-HCL concentration and in this way reproduce the unfolding curves for both wild-type and mutant maltose binding protein showing an increased destabilization of the folded state in the case of the mutant [41].

Single point mutations in the primary structure of proteins may also affect stability and translocation properties of the protein [82]. The group of Jeremy Lee for example used α-HL to demonstrate that protein translocation of a histidine-containing protein is strongly dependent on single-site mutations that affect the protein's folding state [82].

These studies and others demonstrate the ability of nanopore methods for investigating the biophysical properties of proteins at the single molecule level. Given this, and the potential technological applications of this technique, there remains significant interest in its continued development and in finding new applications in the study of proteins [49,130]. Investigations of protein dynamics using nanopores, for example, is a particularly interesting application given the potential to probe long timescale conformational dynamics, which are known to be critically important to protein function [47,48]. Probing dynamics at the single molecule level, moreover, enables the study of rare and / or transient conformations that are typically hidden in ensemble analysis, e.g. via nuclear magnetic resonance [47,133] or X-ray crystallography [134,135]. In this way nanopore protein analysis can be developed into a complementary approach to bulk analysis of protein dynamics with the potential to provide new insights on the conformational motions of proteins and their relation to protein function.

A major challenge with nanopore protein analysis, particularly for probing long timescale dynamics, is that in contrast to uniformly charged biopolymers such as DNA, the

charge distribution of proteins and polypeptides can be highly irregular, positive, negative or neutral, significantly affecting the ability to capture and trap proteins in the pore. Moreover, proteins are folded polymers with a complex free-energy landscape encompassing numerous transition states and conformational sub-states [49]. This complexity coupled with fast transition state kinetics can complicate a detailed kinetic analysis of protein dynamics in the pore. Given this, we have developed methods by which to probe the complex conformational dynamics of single proteins with nanopores. Owing to their low charge density (relative to heavily charged biopolymers such as DNA and RNA) and thus decreased nanopore capture rate we carry out experiments using a salt concentration gradient across the pore which, through a combination of electric field enhancement around the entrance of the pore and osmotic flow [50,51], substantially increases the nanopore capture rate of small molecules in solution (relative to symmetric salt conditions) [50,51] thereby enabling nanopore analysis of proteins. Moreover, we develop machine-learning signal processing algorithms, based on Hidden Markov Models (HMMs), to analyze the complex conformational dynamics and kinetics of proteins, and we apply our methods in analysis of the conformational dynamics of cellular prion protein, a protein whose conversion into a misfolded isoform is responsible for the pathogenesis of prion diseases in humans and animals. Before presenting our methods and results we provide, in the following section, a brief background on the prion protein and prion diseases.

4.2 Prion protein and prion diseases: Background

Cellular prion protein (PrP^C) is a cell-surface protein primarily expressed in the central nervous system (CNS) that is membrane-bound at its C-terminal via

glycosylphosphatidylinositol (GPI) [136,137]. The sequence of the PRNP gene that codes for PrP^C is highly conserved throughout evolution in mammals [137], resulting in structural similarity of the protein across species. PrP^C is a ~200 residue protein, with a molecular weight of ~33kD. Its structure has been solved by NMR or X-ray crystallography for more than twenty different species [138] and is characterized by an unstructured N-terminus and a globular fold in the C-terminal domain (residues 121-230) consisting of three α-helices and a small anti-parallel β-sheet (figure 4-1). Its main function in the CNS is poorly understood [137], although evidence indicates that it plays a role in intracellular signaling, cell proliferation, and cellular homeostasis [139]. Experiments with PRNP knock-out mice indicate that PrP^C may confer some protective function of the brain against dementia and other neurodegenerative disorders associated with old-age [137].

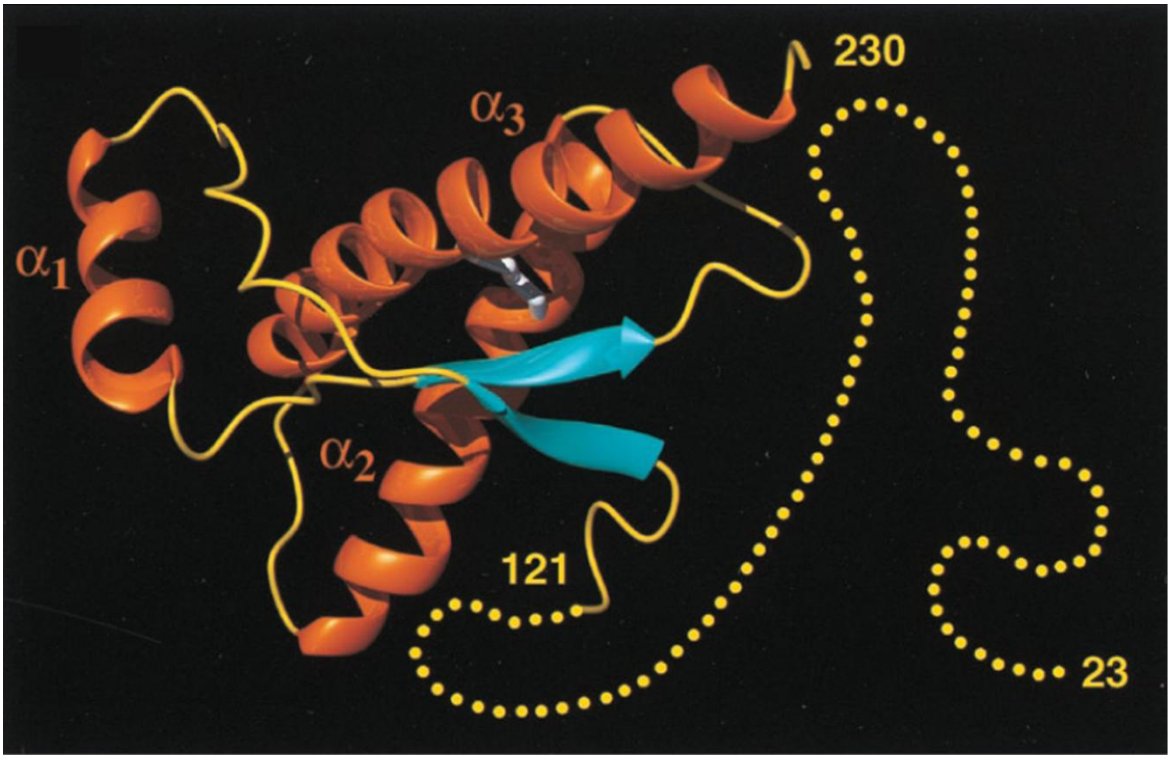


Figure 4-1. Three-dimensional structure of full-length human prion protein ($hPrP^{C}$ (23–230)). The C-terminal domain (residues 121-230) consists of three α-helices (orange) and a small anti-parallel β-sheet (cyan), connected by segments with non-regular secondary structure (yellow). The unstructured, flexible N-terminus (residues 23–121) is represented by the yellow dots. Image source: [140] (reproduced, with permission).

Despite a lack of complete understanding of the function of PrP^{C} in the CNS, misfolded PrP^{C}, designated PrP^{Sc} (or "PrP-Scrapie") is known to be the principal and possibly sole infectious agent responsible for the pathogenesis of prion diseases [55,56,57,141]. Prion diseases (also known as Transmissible Spongiform Encephalopathies) are a class of fatal neurodegenerative diseases affecting both humans and animals associated with the accumulation of PrP^{Sc} in the central nervous system [55,56,57,136]. The most common form of prion disease in humans is Creutzfeldt-Jakob disease (CJD) [137] , which can occur sporadically, through heredity or may be acquired through exposure to brain or nervous system tissue, either iatrogenically [137], or more recently by consumption of bovine

spongiform encephalopathy (mad cow disease) prions through consumption of tainted beef [136]. Sporadic CJD (sCJD) is the most common form of CJD accounting for ~85% of all cases, with an incidence rate of approximately one per million people per year worldwide [137]. The cause of sCJD is unknown, however, it has been proposed that the spontaneous formation of PrP^{Sc} may, in part, explain its occurrence. [136,137]. A stable and uniform incidence rate of non-hereditary forms of CJD in a diversity of countries coupled with the absence of a transmission pattern to account for many of the cases have been strong arguments in favor of the spontaneous occurrence of sCJD [137,142].

The clinical course of prion diseases entails rapidly progressing dementia, visual abnormalities, and cerebellar dysfunction including muscle incoordination, gait, and speech abnormalities, with death usually occurring within weeks or months of onset [137]. Definitive diagnosis of prion disease requires brain biopsy or necropsy and neurohistopathological analysis (there is no CNS inflammation of any conventional type) [58]. There are no effective treatments available for prion diseases.

For over a decade the protein-only hypothesis of PrP^{Sc} replication has been the predominant theory to explain the development and propagation of prion diseases [55,56,57,136,138]. The hypothesis implicates PrP^{Sc} as the principal and possibly sole infectious agent, capable of self-replication by post-translational interaction with PrP^{C} stimulating its conversion into PrP^{Sc} through modification of secondary and tertiary structure (a process known as template-directed misfolding) [55,56,57,136,138]. The mechanics of template-directed misfolding, however, are almost entirely unknown, stymied by the lack of an atomic-level structure for PrP^{Sc}, primarily due to its insolubility and tendency to aggregate [143,144], which has precluded structural determination by high-resolution techniques such

X-ray crystallography or nuclear magnetic resonance. Nevertheless, low-resolution structural data obtained through a variety of chemical, physical and biological techniques has yielded some information on the structure of PrP^{Sc} [138]. Circular Dichroism and Fourier Transform Infrared spectroscopy, for example, shows that PrP^{Sc} has an increased β-sheet content (and decreased α-helical content) relative to PrP^{C} [145] (PrP^{C} contains ~3% β-sheet and ~43% α-helix, while PrP^{Sc} contains ~34% β-sheet and ~20% α-helix [138]). Moreover, PrP^{Sc} is known to be resistant to proteinase K digestion from residues 90 to 230 [138], pinpointing the critical sub-sequence involved in misfolding. Information obtained via low-resolution techniques can aid in the building of structural models of the infectious protein by serving as constraints in model building.

In this regard several structural models of PrP^{Sc} have been developed [146], the most prominent of which are (figure 4-2): (a) Extension of the native PrP^{C} β-sheet [138,146],via recruitment of N-terminal residues and disruption of C-terminal α-helices, (b) a parallel in-register β-sheet [138,146], in which PrP molecules associate into a stacked β-sheet, and (c) a stacked β-helix whereby PrP^{Sc} adopts a left-handed β-helix, in which the conformation of α-helices 2 and 3 remain unchanged, while residues 90 to 160 refold to form a four-layered triangular helix [138,146].

Figure 4-2. Three structural models of PrP^{Sc}. (left) Extension of the native PrP^{C} β-sheet (shown are residues 90-230 of a single monomer) [138,146]. **(middle)** A parallel in-register β-sheet (shown are residues 160–220 of five stacked monomers) [138,146]. **(right)** A β-helix (shown are residues 90-230 of a single monomer) [138,146]. Image source: [146] (reproduced, with permission).

This diversity of structural models of PrP^{Sc} may hint at the possibility of multiple misfolded structures each arising from a local energetic minimum in the folding landscape of PrP [138]. This is supported by the existence of strains in prion disease with distinct phenotypes [138]. In this regard, investigations of PrP^{Sc} structure may yield a variety of potential solutions rather than a single definitive structure [138].

Development of effective diagnostic and treatment methods for prion diseases hinges on a detailed understanding and characterization of disease pathogenesis. Moreover, recent evidence indicates that the prion hypothesis (i.e. protein-only template-directed misfolding) may have general applicability to related neurodegenerative protein misfolding diseases (e.g. Alzheimer disease, Parkinson disease, and Amyotrophic Lateral Sclerosis) [147,148,149,150]. Understanding of conversion and prion propagation in prion disease may therefore shed light on the detailed pathogenesis of other 'prionoid'-like diseases.

A number of pathogenic PrP^{C} mutants (associated with familial prion diseases) exist that are characterized by an increased propensity for conversion into PrP^{Sc} and that differ from

wild-type by only a single amino-acid point mutation in their primary structure. These mutations are known to perturb the stability properties, flexibility and conformational dynamics of the protein [151,152,153]. Methods for probing the biophysical properties of wild-type and mutant prion protein and comparing, therefore, may provide insight into the mechanism of PrP^C conversion in disease.

In this regard, we sought to explore prion protein structure and dynamics, for both wild-type and pathogenic mutant PrP^C by single molecule nanopore sensing.

4.3 Nanopore analysis of wild-type and mutant prion protein

We electrophoretically capture individual wild-type and D178N mutant PrP^C molecules into an organic α-hemolysin nanopore (mutant D178N is a pathogenic PrP^C mutant) and show that these two proteins, which differ from each other by only a single amino-acid point mutation in their primary structure, exhibit easily distinguishable current signatures and kinetics inside the pore. We further demonstrate, with the use of Hidden Markov Model signal processing, accurate detection and discrimination between these two proteins at the single molecule level based on the kinetics of a single PrP^C capture event thereby demonstrating the ultimate sensitivity in discrimination by nanopore sensing. In addition, we present a four-state model to describe wild-type PrP^C kinetics in the pore which represents a first step in our investigation into characterizing the differences in kinetics and conformational dynamics between wild-type and D178N mutant PrP^C. These results demonstrate the potential of nanopore analysis for highly sensitive, real-time protein and

small molecule detection based on single molecule kinetics inside a nanopore, and show the utility of this technique as an assay to probe differences in stability between wild-type and mutant prion proteins. The methods we present enable the exploration of long timescale conformational motions of proteins at the single molecule level, making nanopore sensing a new tool for studying protein dynamics.

4.3.1 Materials and methods

PrP^C Constructs

PrP^C (both wild-type and mutant) was expressed and purified by the PrioNet Prion Protein & Plasmid Production Platform Facility (refer to Appendix D for details on expression and purification protocol). Truncated Syrian Hamster PrP^C (residues 120-232 – designated ShPrP(120-232)) was used in order to investigate the structure and dynamics of the PrP^C structural core. To facilitate capture of PrP^C inside the nanopore, the N-terminus of ShPrP(120-232) was adapted with four positively charged amino-acid residues (KKRR) (designated KKRR-ShPrP(120-232)). We expect these additional residues to have a minimal effect on the overall structure and stability of PrP^C based on previous studies whereby truncated PrP^C (of various lengths) was adapted with a 22 residue N-terminal fusion tag, which was found to have no influence on PrP^C structural stability [154,155,156]. Three PrP^C constructs were investigated in this study, namely: ShPrP(120-232), KKRR-ShPrP(120-232) and KKRR-ShPrP(120-232)-D178N (i.e. mutant PrP^C).

Nanopore experiments

α-hemolysin nanopores were formed using a method adapted from that of Akeson *et. al*. [22]. Briefly, a black lipid membrane of 1,2-diphytanoyl-sn-glycero-3-phosphocholine (Avanti Polar Lipids Inc., Alabaster, AL) and hexadecane (Sigma-Aldrich, St. Louis, MO) is formed across a 25μm PTFE aperture connecting two baths filled with electrolyte (figure 4-3. Refer to Appendix C for details on lipid bilayer formation).

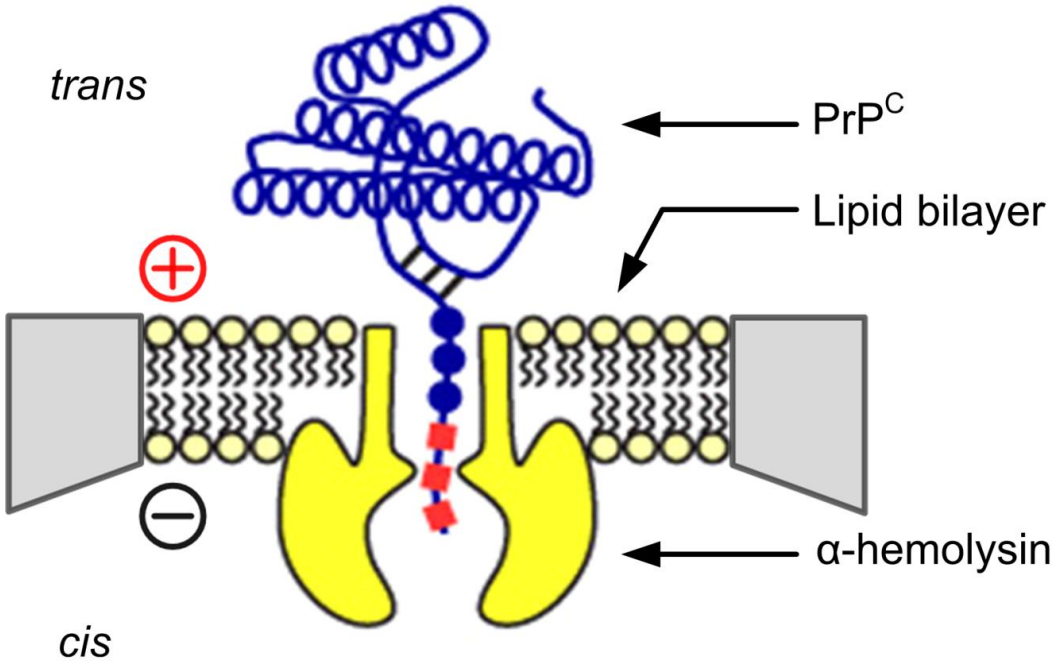


Figure 4-3. Capture of an individual PrP^C molecule into an α-HL nanopore. PrP^C is electrophoretically driven into the α-HL nanopore (voltage polarity given by the plus and minus signs) via its positively charged N-terminus. The *trans* chamber contains 0.3M KCl, 45mM $NaPO_4$, 10mM HEPES, pH 8.0 solution at a PrP^C concentration of 13.4μM. The *cis* chamber contains 3M KCl, 10mM HEPES, pH 8.0 solution. The salt-concentration gradient across the pore generates an osmotic flow from *trans*-to-*cis* and enhances the electric field around the entrance of the *trans*-side of the pore [50,51] thereby substantially increasing the nanopore-capture rate of PrP^C in solution relative to symmetric salt conditions [50,51]. Experiments were conducted (and maintained) at a temperature of 20°C ± 0.1°C.

Owing to the low charge density of KKRR-ShPrP(120-232) (relative to heavily charged biopolymers such as DNA and RNA), and thus decreased ability to capture PrP^C in the pore, experiments were conducted under asymmetric salt conditions which, through a combination of electric field enhancement around the entrance of the pore and osmotic flow [50,51], substantially increases the nanopore-capture rate of small molecules in solution (relative to symmetric salt conditions) [50,51]. PrP^C was preloaded on the *trans* side of the pore in 0.3M KCl, 45mM NaPO4, 10mM HEPES, pH 8.0 solution (final PrP^C concentration was ~13.4μM). We do not expect oligomerization, aggregation or precipitation of PrP^C under these solution conditions based on mass spectrometry results of KKRR-ShPrP(120-232) in solution and on previous studies with various full-length and truncated forms of PrP^C in high salt [157,158]. Our mass spectrometry data confirms the presence of monomeric PrP^C in solution and the absence of dimers (data not shown), suggesting the absence of higher-order oligomers as well. Evidence in the literature indicates that the aggregation propensity of PrP^C in the presence of high salt is due to interactions of anions in solution with glycine groups in the glycine-rich unstructured PrP^C N-terminus (i.e. residues 23-119), and is therefore a property of full-length PrP^C (i.e. PrPC(23-232)) [157,158]. Moreover, studies of truncated PrP^C (i.e. PrP^C(120-232)) in high salt buffer (0.5M NaCl) find no change in secondary structure content compared to salt-free buffer and do not report of aggregation or precipitation of PrP^C in solution [157]. In contrast to the *trans* side of the pore the *cis* side contained 3M KCl, 10mM HEPES, pH 8.0 solution. All experiments were conducted (and maintained) at a temperature of $20^0C \pm 0.1^0C$. An Axon Axopatch 200B patch clamp amplifier is used to measure the ionic current. Data is low-pass filtered at 10kHz by a 4-pole Bessel filter and sampled at 100kHz. Experiments were conducted with a single α-HL pore

incorporated in the lipid bilayer. Formation of an α-HL pore was done under symmetric salt conditions (i.e. 0.3M KCl on both the *cis* and *trans* sides of the pore) by injection of free subunits into solution (on the *cis* side of the pore) which subsequently self-assemble into heptameric, membrane-spanning pores or by injection of preformed heptameric α-HL into solution which spontaneously forms into a transmembrane pore in a lipid bilayer [159]. Confirmation of a single pore incorporation is achieved by applying a +100mV electric potential across the membrane (*trans* side positive) and observing a specific step-wise increase in the current (~+28pA at +100mV and ~-20pA at -100mV). Once this is observed, the RMS current noise on the pore (5kHz bandwidth using a Butterworth filter – independent of the 10kHz low-pass Bessel filter) is then probed at +100mV and +200mV. The RMS noise serves as a second metric for minimizing pore-to-pore variation between experiments. At +100mV the upper limit for the 5kHz noise, for our experimental setup, is ~0.80pA RMS, and at +200mV the upper limit is ~1.20pA RMS. If both the current and noise are within specification, the *cis* side of the pore is then perfused with 3M KCl, 10mM HEPES, pH 8.0 solution several times to ensure that any free, unbound α-HL has been removed from solution and to ensure that the salt concentration on the *cis*-side of the pore has been increased to 3M KCl. Custom-built data acquisition software (described elsewhere [119]) is used to apply large positive voltages (e.g. +160mV or greater) across the pore, and record the corresponding current and voltage, thereby electrophoretically driving individual PrP^C molecules in solution into the pore. Several thousand individual PrP^C capture events were recorded at each voltage in the range of +160mV to +240mV for each PrP^C construct.

PrP^C capture rate & KKRR-ShPrP (120-232) Event lifetime

To confirm that KKRR-ShPrP(120-232) enters the pore N-terminal first, we characterized and compared the nanopore capture rate as a function of voltage for KKRR-ShPrP(120-232), ShPrP(120-232), and a buffer only control (indicative of the pore gating rate under the given buffer conditions – i.e. asymmetric salt concentration). The capture rate at a given voltage, is determined by first calculating the survival probability, $P_s(t)$, of the open-channel state as a function of time. This represents the probability that the pore is open (i.e. unoccupied) at time *t* since the end of the last capture event. Capture of a molecule into the pore is defined, in terms of pore-current, when I/I0 crosses a threshold of 0.84, where *I* is the current through the pore and *I0* is the open-pore current. $P_s(t)$ is given by:

$$P_s(t) = 1 - \frac{i}{N} \tag{23}$$

where the open-pore event times are sorted in increasing order. *i* is the index of the i^{th} event (where the first event, i.e. the shortest event, has an index of 0) and *N* is the total number of events. The survival probability, in this case, exhibits exponential decay, and therefore can be least-squares fitted to an exponential function in order to extract the capture rate at a given voltage.

The capture rate for KKRR-ShPrP(120-232) and ShPrP(120-232) is exponentially dependent on voltage (figure 4-4a), consistent with the capture process being dominated by an energy barrier whereby, according to classical Kramer's theory, the applied voltage acts on the N-terminal positive charges (five in the case of KKRR-ShPrP(120-232) and one in the

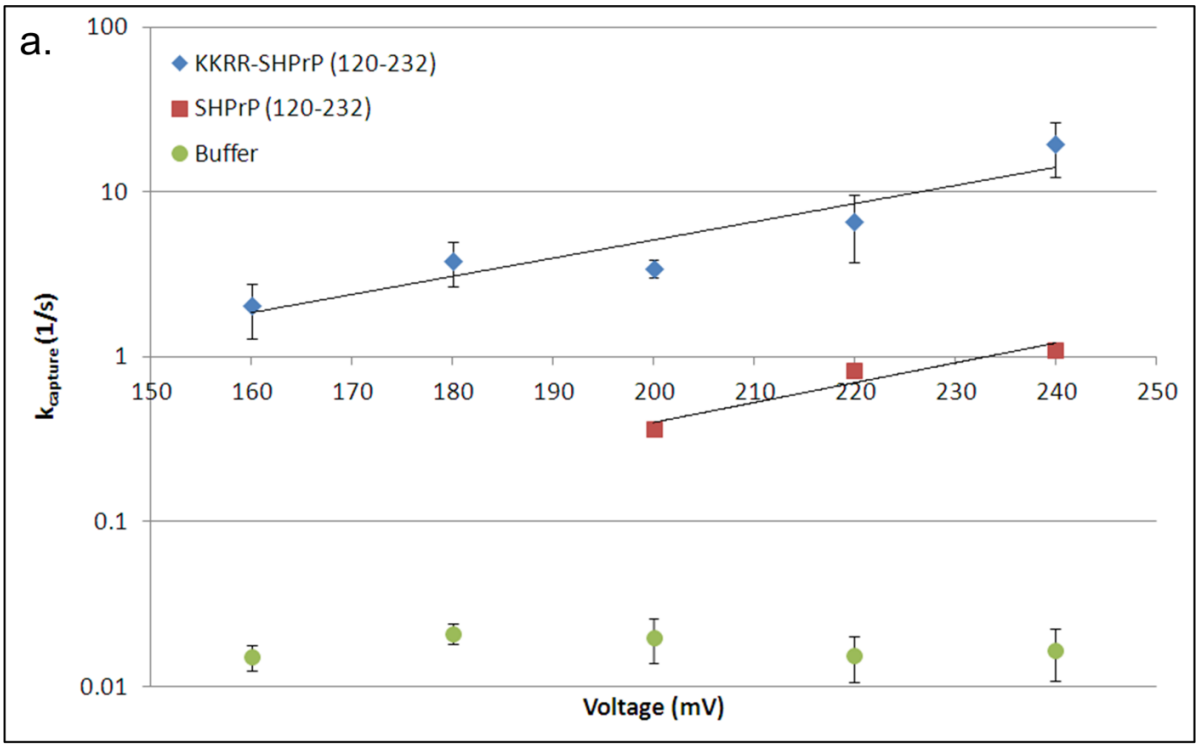
case of ShPrP(120-232)), decreasing the energy barrier height for entry into the pore, thereby exponentially increasing the rate of PrP^C capture. Moreover, the nanopore-capture rate of KKRR-ShPrP(120-232) is between one and one-and-a-half orders of magnitude higher (depending on voltage) than the capture rate of ShPrP(120-232), reflecting the greater charge density at the N-terminus in the case of KKRR-ShPrP(120-232). These results indicate that the large majority of captures of KKRR-ShPrP(120-232) involve threading of the N-terminus through the pore.

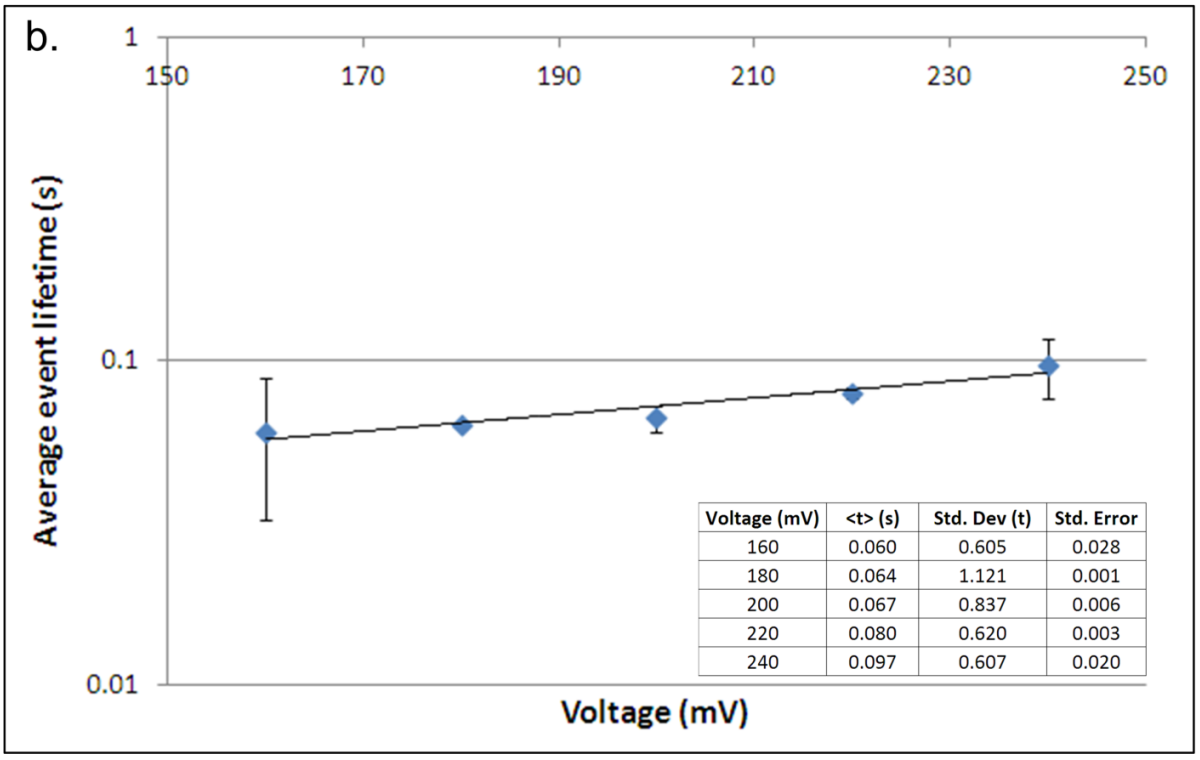
The average event lifetime for KKRR-ShPrP(120-232) as a function of voltage is shown in figure 4-4b and is determined from the normalized histogram of the log-event times (*ln(t)*) at a given voltage (i.e. *p[ln(t)]* - the probability distribution of log-event times. log-event times are used because event times span several orders of magnitude). The average event lifetime is given as:

$$\langle t \rangle = \frac{\int e^{\ln(t)} p[\ln(t)] d[\ln(t)]}{1 - \frac{N}{N_{tot}}} \qquad (24)$$

where N is the number of events that do not end in escape from the pore (e.g. events that are terminated by reversing the voltage and forcing PrP^C out of the pore. These events are not timed and therefore do not contribute to *p[ln(t)]*) and N_{tot} is the total number of events.

The event lifetime increases exponentially with voltage, consistent with PrP^C escape from the pore (as opposed to translocation) being the dominant mode of termination of an event, and requiring crossing an electrostatic energy barrier (governed by the applied voltage). The standard deviation of the event lifetime distribution shows the lifetime spanning several orders of magnitude indicating the presence of both short and very long time events (i.e. > 1s).





Voltage (mV)	<t> (s)	Std. Dev (t)	Std. Error
160	0.060	0.605	0.028
180	0.064	1.121	0.001
200	0.067	0.837	0.006
220	0.080	0.620	0.003
240	0.097	0.607	0.020

Figure 4-4. PrP^C capture rate and KKRR-ShPrP(120-232) event lifetime (3M KCl *cis*, 0.3M KCl *trans*). (a) Shown is the capture rate as a function of voltage for KKRR-ShPrP(120-232) (blue), ShPrP(120-232) (red) and the buffer-only control (green). In the case of the buffer-only control the capture rate represents the rate of pore gating as a function of voltage. Both KKRR-ShPrP(120-232) and ShPrP(120-232) exhibit capture kinetics that are exponentially dependent on voltage consistent with the applied voltage acting on the positively charged residues at the N-terminus (five in the case of KKRR-ShPrP(120-232) and one in the case of ShPrP(120-232)) to decrease the energy barrier height for entry into the pore, thereby exponentially increasing the capture rate. In addition, the capture rate for KKRR-ShPrP(120-232) is between one and one-and-a-half orders of magnitude higher than ShPrP(120-232) indicating that the large majority of captures of KKRR-ShPrP(120-232) involve threading of the N-terminus through the pore. Error bars represent the standard error on the mean, and were determined via bootstrapping in the case of ShPrP(120-232) and the buffer-only control, whereas in the case of KKRR-ShPrP(120-232) error bars were determined based on two separate datasets. **(b)** The average event lifetime for KKRR-ShPrP(120-232) increases exponentially with voltage consistent with PrP^C escape from the pore (as opposed to translocation) over an electrostatic energy barrier (governed by the applied voltage). The standard deviation of the event lifetime distribution indicates the presence of both short and very long time events (i.e. > 1s). Error bars represent standard error on the mean and were determined based on two separate datasets.

4.3.2 Results and discussion

Figure 4-5 shows the ionic current histograms for all PrP^C capture events from all voltages, with ionic current normalized by the open-pore current at a given voltage, for KKRR-ShPrP(120-232) and KKRR-ShPrP(120-232)-D178N (see Appendix E for sample wild-type and D178N mutant PrP^C capture events). Both histograms display multiple peaks, with varying degrees of overlap, indicative of complex PrP^C dynamics inside the pore. Moreover, the histograms exhibit clear distinguishable features (e.g. the near absence of the peak at I/I0 ~ 0 with respect to mutant PrP^C), indicating differences in conformational dynamics between the two proteins.

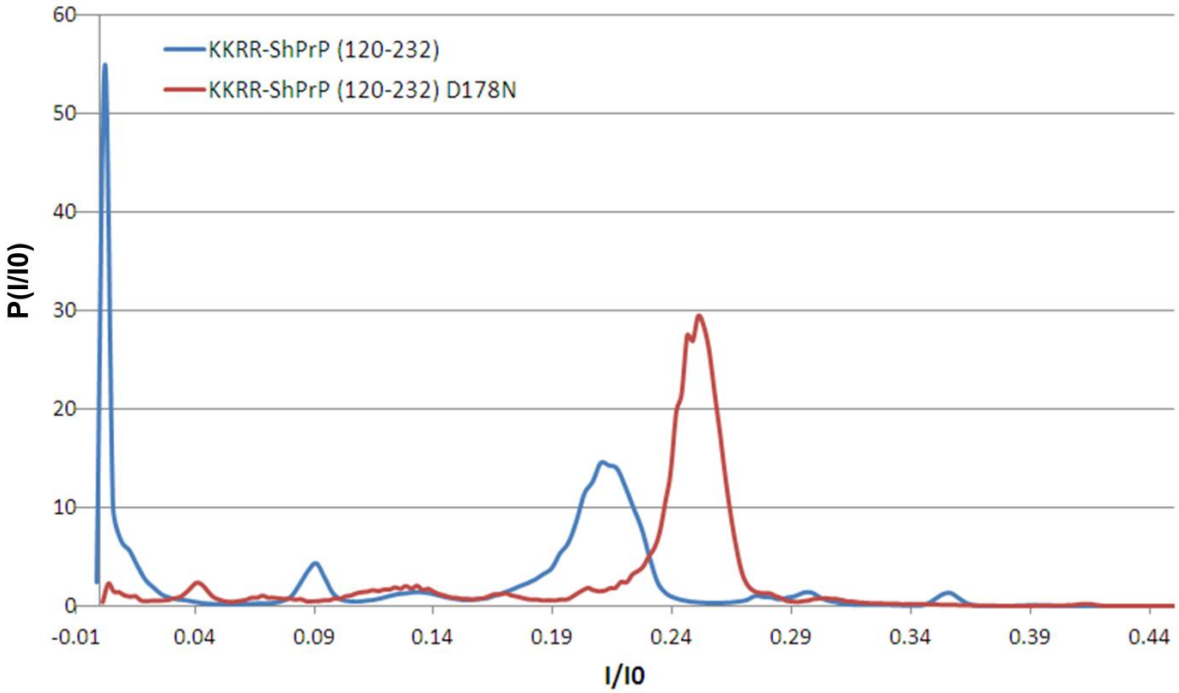
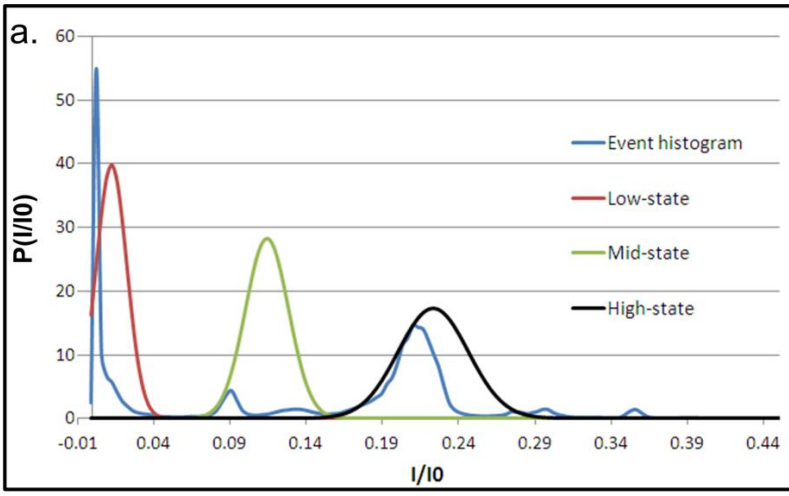


Figure 4-5. All-point current histogram for KKRR-ShPrP(120-232) and KKRR-ShPrP(120-232)-D178N. Ionic current histogram of all PrP^C capture events from all voltages, with ionic current normalized by the open-pore current (I0) at a given voltage, for KKRR-ShPrP(120-232) (blue) and KKRR-ShPrP(120-232)-D178N (red). Ionic current is median filtered to 2.99ms per data-point. The histograms exhibit multiple peaks with varying degrees of overlap indicative of complex PrP^C dynamics in the pore. Moreover, the histograms exhibit clear distinguishable features (e.g. the near absence of the peak at I/I0 ~ 0 with respect to KKRR-ShPrP(120-232)-D178N).

In order to model PrP^C kinetics in the pore and characterize signal statistics we developed a signal processing algorithm based on Hidden Markov Models (HMMs). HMM signal processing is a powerful technique by which to extract and characterize low-level signals buried in background noise [52,53,54] . The approach has been previously applied in characterizing the complex kinetics of DNA hairpins trapped in the α-HL nanopore [160,161]. HMM analysis requires an initial model for the system defined by the following parameters: State levels in terms of I/I0 (q), the initial condition (π), i.e. the probability that an event begins in a given state, the transition probabilities between states (A), and the noise

properties (the standard deviation on a Gaussian) of each state (*b*). These parameters represent best guesses and can be estimated, with respect to *q* and *b*, from the event histogram. Once the initial model is developed and provided with the corresponding data, the HMM operates through an expectation maximization (EM) algorithm improving upon the model parameters with each iteration until convergence to an optimal set of parameters that best describes the data, based on model likelihood, i.e. HMM analysis converges to the maximum likelihood model estimate (we refer the reader to [52,162] for details regarding HMM theory and its implementation). We model PrP^C kinetics in the pore as a three-state system. Our choice of three states to describe PrP^C kinetics is based on parsimony i.e. selecting a model with the fewest parameters that describes the data well. Moreover, our choice is based on wild-type PrP^C kinetics and in particular on a qualitative assessment of the form of the wild-type histogram. As mentioned previously, the histogram displays multiple peaks, with varying degrees of overlap. We find, however, that the peaks are concentrated into roughly three regimes. A simple description of the histogram, therefore, is one in which the current is split into a high, mid and low regime (i.e. three states - refer to figure 4-6a, which shows our initial wild-type PrP^C HMM model and best guess at the location and size of each regime). For ease of comparison between wild-type and D178N mutant PrP^C we model mutant kinetics in the pore as a three-state system as well. The initial and optimal (post-HMM processed) models are shown in figures 4-6 and 4-7 for KKRR-ShPrP(120-232) and KKRR-ShPrP(120-232)-D178N respectively.

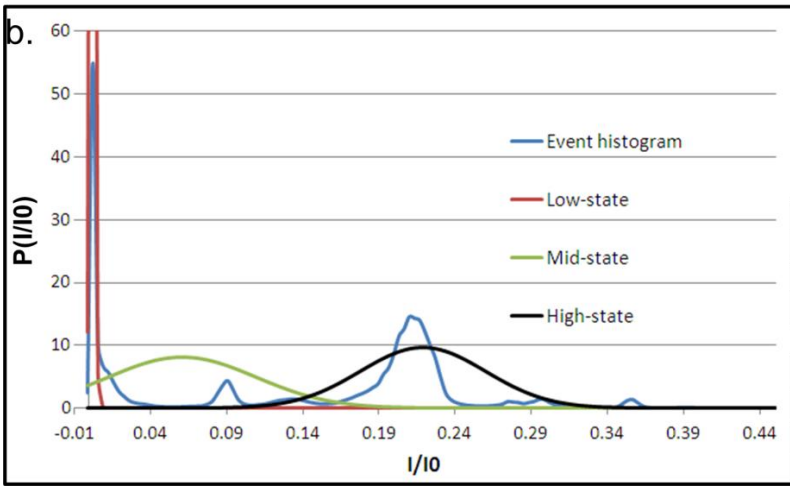
a.



State	π	q (I/I0)	b
Low	1/3	0.012	0.010
Mid	1/3	0.114	0.014
High	1/3	0.223	0.023

A	Low	Mid	High
Low	1/3	1/3	1/3
Mid	1/3	1/3	1/3
High	1/3	1/3	1/3

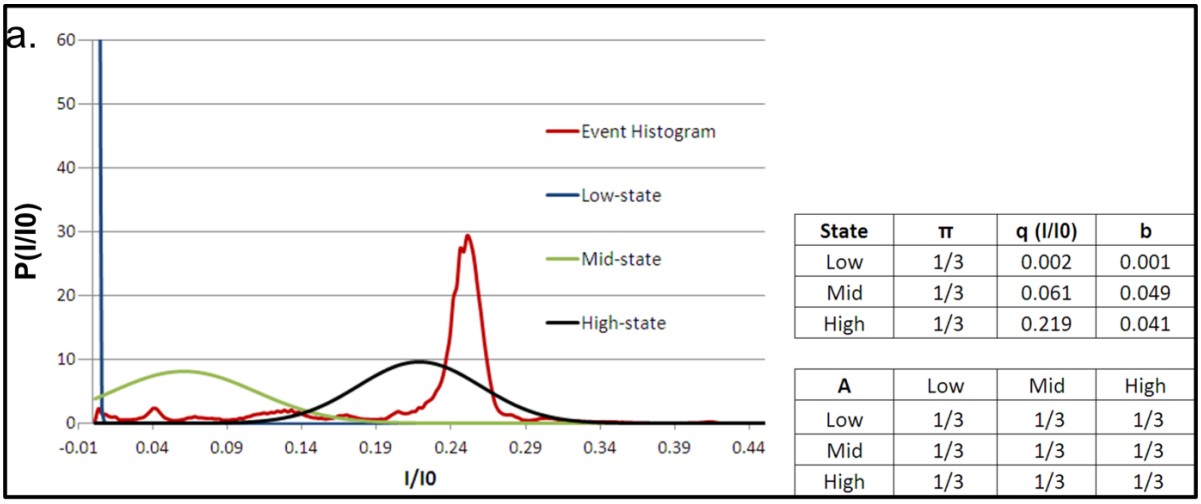
b.



State	π	q (I/I0)	b
Low	0.00006	0.002	0.001
Mid	0.014	0.061	0.049
High	0.986	0.219	0.041

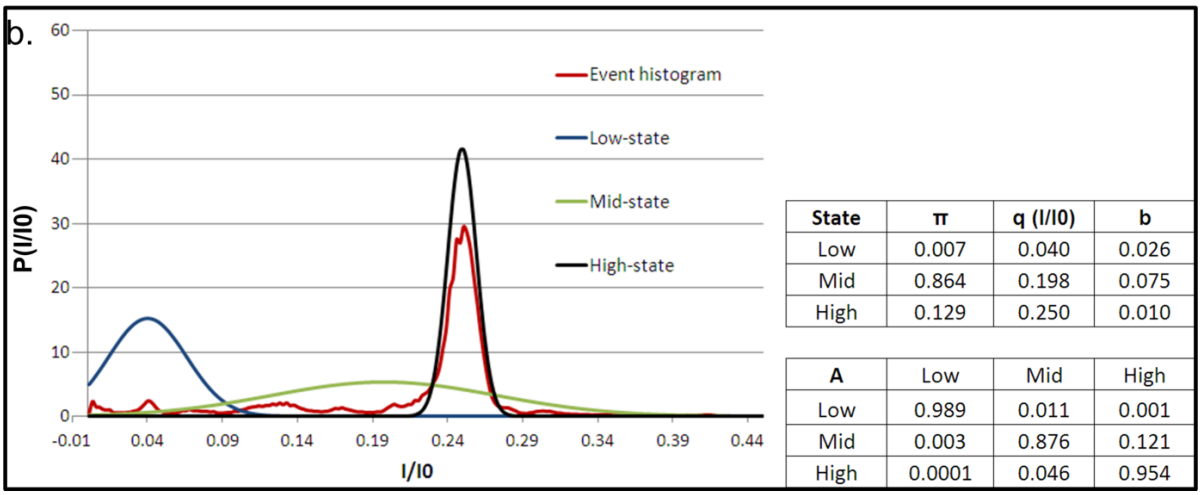
A	Low	Mid	High
Low	0.967	0.033	0.000
Mid	0.036	0.931	0.033
High	0.000	0.013	0.987

Figure 4-6. KKRR-ShPrP(120-232) event histogram and initial and optimal HMM models. (a) (Left) The KKRR-ShPrP(120-232) event histogram (blue) is divided into three regimes / states a high-state (black), mid-state (green) and low-state (red). The location of the peak and width of the distribution for each state in our initial model represents our best guess of the location and size of a given regime. **(Right)** The model parameters: π (i.e. the initial condition or probability that an event begins in a given state), q (the location of the peak of the Gaussian distribution, in terms of I/I0, for a given state), b (the standard deviation on the Gaussian of each state, which defines a state's noise properties), and A (the state-to-state transition probability matrix). In our initial model we assume ignorance of the probabilities and therefore assume π to be uniformly distributed (i.e. an event is assumed equally likely to begin in any of the three states). Similarly with the transition probability matrix A, we assume all transitions to be equally likely (e.g. if in the low-state there is an equal probability for remaining in the low-state as there is for transitioning into the mid-state or the high-state). **(b) (Left)** After 40 iterations of the EM algorithm the optimal three-state model that best describes the data (i.e. the maximum likelihood model estimate) is converged upon. The low-state, far from encompassing all of the low current (as was presumed in our initial model) is very narrow and well defined, while the mid and high states both broaden out (the peak of the mid-state also shifts to a deeper conductance level relative to the initial model). **(Right)** The corresponding optimal model parameters.



State	π	q (I/I0)	b
Low	1/3	0.002	0.001
Mid	1/3	0.061	0.049
High	1/3	0.219	0.041

A	Low	Mid	High
Low	1/3	1/3	1/3
Mid	1/3	1/3	1/3
High	1/3	1/3	1/3



State	π	q (I/I0)	b
Low	0.007	0.040	0.026
Mid	0.864	0.198	0.075
High	0.129	0.250	0.010

A	Low	Mid	High
Low	0.989	0.011	0.001
Mid	0.003	0.876	0.121
High	0.0001	0.046	0.954

Figure 4-7. KKRR-ShPrP(120-232)-D178N event histogram and initial and optimal HMM models. (a) (Left) The KKRR-ShPrP(120-232)-D178N event histogram (red), and the corresponding high-state (black), mid-state (green) and low-state (blue) that make up the initial model for HMM analysis. The location of the peak (*q*) and width of the distribution (*b*) for each state are the same as for the optimal KKRR-ShPrP(120-232) model (figure 4-6b). This choice for *q* and *b* serves to highlight how the individual states evolve and differ from that of wild-type PrP^C. **(Right)** The corresponding initial model parameters. Similar to the initial model for KKRR-ShPrP(120-232) (figure 4-6a) we assume ignorance of the probabilities and therefore assume π to be uniformly distributed. Likewise, with the transition probability matrix (*A*), we assume all transitions to be equally likely. **(b) (Left)** After 36 iterations of the EM algorithm the optimal three-state model that best describes the data (i.e. the maximum likelihood model estimate) is converged upon. The individual states (properties and kinetics) are significantly different from wild-type PrP^C (figure 4-6b), highlighting the importance of amino-acid residue D178 to the dynamics and structural stability of PrP^C **(Right)** The corresponding optimal model parameters.

The optimal three-state model for wild-type and mutant PrP^C (figures 4-6b and 4-7b respectively) reveal significant differences in both state properties and kinetics between the two proteins, highlighting the importance of amino acid residue D178 to the dynamics and structural stability of PrP^C. It is known that residue D178 in wild-type PrP^C stabilizes the protein through salt-bridge interactions with R164 (the C-terminus of β-strand 2) and by hydrogen bonding to Y128 (the N-terminal Tyr of β-strand 1) [163,164,165,166]. In mutant D178N therefore, these stabilizing forces are no longer present, the loss of which appears to significantly affect the conformational dynamics of mutant PrP^C in the pore.

The clear distinction between these two proteins also highlights the sensitivity of nanopore analysis in detecting changes in biomolecule structure and demonstrates the potential of using this technique for detection and identification of small molecules and proteins in solution based on differences in kinetics in a nanopore. In order to explore this potential we characterized protein-calling accuracy between wild-type and D178N mutant PrP^C at the single event level, based on kinetic differences in the pore (i.e. given a single event we characterized the accuracy with which it can be determined which protein, either wild-type or mutant, produced the event based on kinetics). In this regard we investigated two cases in particular:

1) Where all events are analyzed, regardless of event lifetime and
2) Where only those events that have a lifetime of ≥ 1s are analyzed

Case 2 allows us to characterize how protein-calling accuracy changes as we limit our study to long-time events (i.e. those events with long observation times and therefore better statistics for discriminating between the two proteins). In order to make individual calls given an event we developed a protein-calling algorithm based on HMM model likelihood. The method by which we characterize protein-calling accuracy and the results obtained are described in the following.

Case 1

Of all wild-type and mutant PrP^C events we first simulate a 50:50 mix dataset. Therefore 500 wild-type events and 500 mutant events are randomly selected and combined to form a simulated 50:50 mix of 1000 events. The remaining events for both wild-type and mutant form the training sets for the corresponding protein by which to build optimal HMM models (refer to figures F-1 and F-2 in Appendix F for the determined optimal wild-type and mutant models, respectively). Protein-calling accuracy is assessed by calling the individual events from the 50:50 mix dataset (i.e. the blind set of events). Individual events are called as either wild-type or mutant using our HMM protein-calling algorithm. The algorithm works as follows: Given an individual event and the optimal wild-type and mutant HMM models (as determined via HMM analysis of the respective training data) the algorithm calculates the likelihood function of the event given each model (i.e. $P(E|\lambda_{wild\text{-}type})$ and $P(E|\lambda_{mutant})$, where E is the event and λ represents a given model). The algorithm then calls an event as either wild-type or mutant based on maximum likelihood (i.e. the protein-call is based on whichever of the two determined likelihood functions is largest). The protein-calling results are given in table 4-1. The results are given in terms of the wild-type and mutant predictive value. This is

defined as the likelihood that an event which is called as either wild-type or mutant is called correctly (e.g. if the algorithm calls an event as mutant there is a 0.86 likelihood that the event is a mutant event).

Case 2

In characterizing protein-calling accuracy in case 2 the method is the same as described for case 1 with the difference being that in case 2 we are only interested in events that have an event lifetime of ≥ 1s. In other words, all events making up the training sets (wild-type and mutant) and the generated 50:50 mix dataset have an event lifetime of ≥ 1s (refer to figures F-3 and F-4 in Appendix F for the determined optimal wild-type and mutant models in this case, respectively). The total number of events for the 50:50 mix dataset is 176. The results, in terms of predictive value are given in table 4-1.

	Case 1	Case 2
# of events in 50:50 mix	1000	176
Wild-type predictive value	0.71	0.85
Mutant predictive value	0.86	0.90

Table 4-1. Wild-type and mutant predictive value. Case 1 refers to the situation whereby all events, regardless of event lifetime, are analyzed. The simulated 50:50 mix dataset forms the individual events to be protein-called by which wild-type and mutant predictive value is determined. These events are randomly selected from the total number of wild-type and mutant events. After selection (and removal from the total number of events), the remaining events form the training sets for both wild-type and mutant. Wild-type predictive value refers to the likelihood that an event is in fact a wild-type event given that the protein-calling algorithm has called it as wild-type. Similarly the mutant predictive value refers to the likelihood that an event is in fact a mutant event given that the protein-calling algorithm has called it as mutant. Case 2 refers to the case whereby only those events that have an event lifetime of $\geq$ 1s are analyzed (i.e. only long time events makeup the training sets for both wild-type and mutant and the generated 50:50 mix dataset). The predictive value, not surprisingly, improves when only considering long-time events which is primarily due to the fact that long-time events can be better assessed in terms of their kinetics than short-time events (i.e. the amount of data available to characterize an event is proportional to the event lifetime) thereby improving protein-calling accuracy.

These results show, not surprisingly, that predictive value (i.e. protein calling accuracy) improves when considering only long-time events. This is primarily due to the fact that long-time events can be better assessed in terms of their kinetics than short-time events (i.e. the amount of data available to characterize an event increases proportionately with the event lifetime) thereby improving protein-calling accuracy. In particular, the mutant at short times has a greater propensity for being called as wild-type than at long-times (i.e. mutant kinetics at short-times is less distinguishable from wild-type than at long-times). We note here that of all wild-type events, ~92% of them have an event lifetime of < 0.1s. Similarly of all D178N mutant events, ~66% have an event lifetime of < 0.1s. Therefore the majority of observed

events (the large majority in the case of wild-type PrP^C) are short-lived. Even in the case of short-lived events which are between 1 and 33 datapoints long (i.e. events are filtered to 2.99ms per datapoint), with a large proportion of events having an event lifetime of ≤ 0.01s (i.e. between 1 and 3 datapoints long - ~30% of events in the case of wild-type PrP^C), the results show that wild-type and mutant PrP^C are easily distinguished based on their kinetics in the pore. The results substantially improve, particularly with respect to wild-type predictive value, when only long-time events are considered.

As an alternative measure of the discriminatory power of this method to distinguish between wild-type and mutant PrP^C we characterize the Case 2 mutant receiver operating characteristic (ROC) curve. An ROC curve characterizes the performance of a binary classifier as a function of the discrimination threshold [167]. In particular, the area under the curve (AUC), is a frequently used performance metric by which to compare classification systems [167,168], whereby $AUC = 1$ corresponds to a perfect classifier and $AUC = 0.5$ corresponds to a classifier based on random guessing. The discrimination threshold, in our case, is the difference in log-likelihood (ΔL) between wild-type and mutant models for a given event, i.e. $\Delta L = ln(P(E|\lambda_{wild\text{-}type})) - ln(P(E|\lambda_{mutant}))$. The results of table 4-1, for example, correspond to $\Delta L = 0$ (i.e. if $\Delta L \geq 0$ the event is called as wild-type whereas if $\Delta L < 0$ the event is called as mutant). Figure 4-8 shows the mutant ROC curve as the discrimination threshold varies from $\Delta L = -40355$ to $\Delta L = 43352$. Also shown is the ROC curve for a classifier based on random guessing (i.e. no discriminatory power, $AUC = 0.5$).

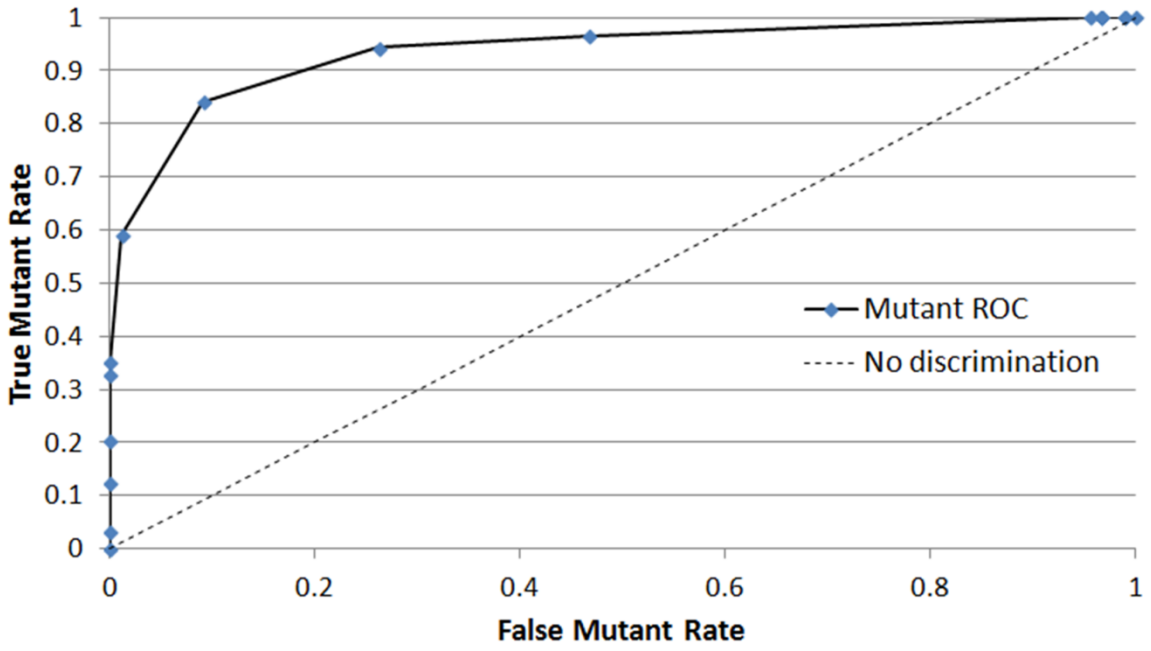


Figure 4-8. Mutant ROC curve vs. random guessing. The true mutant rate, equal to the classifier sensitivity, is a measure of the classifier's ability to identify mutant events correctly. The false mutant rate, equal to one minus the classifier specificity, is a measure of the degree to which events are falsely identified as mutant. The area under the mutant ROC curve (dot-solid line), a measure of the performance of the classifier [167,168], is ~0.935 indicative of excellent performance [169,170]. This can be compared to a classifier based on random guessing (dashed line), which has no discriminatory power with an AUC of 0.5.

The area under the mutant ROC curve, determined by trapezoid integration, is ~0.935 which forms a lower-bound estimate on the AUC [168]. A rough guide used to characterize the accuracy of a binary classifier is the traditional academic point system [169,170]: 0.90 -1 = excellent, 0.8 - 0.9 = good, 0.7 - 0.8 = fair, 0.6 - 0.7 = poor, 0.50 - 0.6 = fail, which by this standard indicates our classifier is an excellent performer.

These results demonstrate that nanopore analysis in combination with HMM signal processing can be used to detect and discriminate between wild-type and mutant PrP^C at the single event level based on their kinetics in the pore. These results therefore show the potential of using this technique as an assay to probe differences in stability between wild-

type and mutant prion proteins at the single molecule level, which opens up the possibility of studying small molecule-PrP^C interactions and the effects of these molecules on PrP^C stability as a possible screen for small molecules that improve the stability properties of the protein. Moreover, the ability to discriminate between two proteins that differ by only a single-amino acid point mutation demonstrates the sensitivity of this approach in detecting subtle changes in biomolecule structure, and points to the possibility of developing this technique for highly sensitive, real-time detection and identification of small molecules and proteins in solution, with potential applications in disease biomarker and pathogen detection.

We return now to a more detailed analysis of PrP^C kinetics in the pore with the goal of characterizing the kinetic differences between wild-type and D178N mutant PrP^C. We limit our discussion here specifically to the kinetics of wild-type PrP^C (see below for a discussion D178N mutant kinetics). The optimal model shown in figure 4-6b is a model of the kinetics of PrP^C in the pore over all voltages. To characterize the voltage-dependence of PrP^C kinetics the voltage-specific optimal model must be determined. This is done by HMM analysis of only those events at a given voltage whereby the optimal model (figure 4-6b) serves as the initial model for the voltage-specific HMM analysis, with the caveat that the state levels (q) and noise properties of each state (b) remain constant during the analysis (i.e. only π and A are updated during the voltage-specific HMM analysis). The voltage-specific HMM analysis therefore, improves the estimate of the initial condition and the transition probabilities for a given voltage. Given the voltage-specific optimal model and an individual event the most likely state sequence for the event is then determined (i.e. the event Viterbi path) via the Viterbi algorithm (refer to [162] regarding the theory and implementation of the Viterbi

algorithm). State properties, such as the lifetime distribution of each state, and state-to-state transition rates are then determined by analyzing the Viterbi path for all events. Specifically, given the Viterbi path for each event the transition rates between states are determined by first constructing the log-state time distribution for each state over all events. The exit rate for a particular state (i.e. the rate to transition out of a given state) is given by:

$$k_{exit} = \left(1 - N/N_{tot}\right)\left(\int e^{\ln(t)} p[\ln(t)] d[\ln(t)]\right)^{-1} \quad (25)$$

where $p[ln(t)]$ is the log-state time distribution for the given state (i.e. the normalized histogram of log-state times). N is related to the fact that the last occupied state in those events that do not end in escape from the pore is not timed but is nevertheless used in the calculation of the exit rate for that particular state. N is the number of instances, for a particular state, that were not timed and N_{tot} is the total number of instances of that state. Given the exit rate for a particular state the state-to-state transition rate is then given by,

$$k_{i \to j} = p_j k^i_{exit} = \frac{n_j}{n_{tot}} k^i_{exit} \quad (26)$$

where $k_{i \to j}$ is the transition rate from state i to j, k^i_{exit} is the exit rate out of state i, and p_j is the probability of transitioning into state j from state i. p_j is given by the number of

transitions into state *j* from state *i* (n_j) divided by the total number of transitions out of state *i* (n_{tot}), which includes escapes from the pore from state *i*.

Shown in figure 4-9 are the statistics of the mid-state (i.e. the transition rate from the mid-state to the high and low states as a function of voltage). For the mid-state transition rates given in figure 4-9 (bottom left) and figure 4-9 (bottom right) the mid-state event times are first separated into two groups:

Group 1: Mid-state event times whereby the mid-state was entered from the high-state

Group 2: Mid-state event times whereby the mid-state was entered from the low-state.

Exit rates for each group and state-to-state transition rates are then determined as given by equations 25 and 26 respectively. The results in figure 4-9 highlight the dependence of mid-state statistics on how it is entered. For example, the transition rate from the mid-state to the low-state differs by one-to-two orders of magnitude depending on if the mid-state is entered from the high-state versus if it is entered from the low-state.

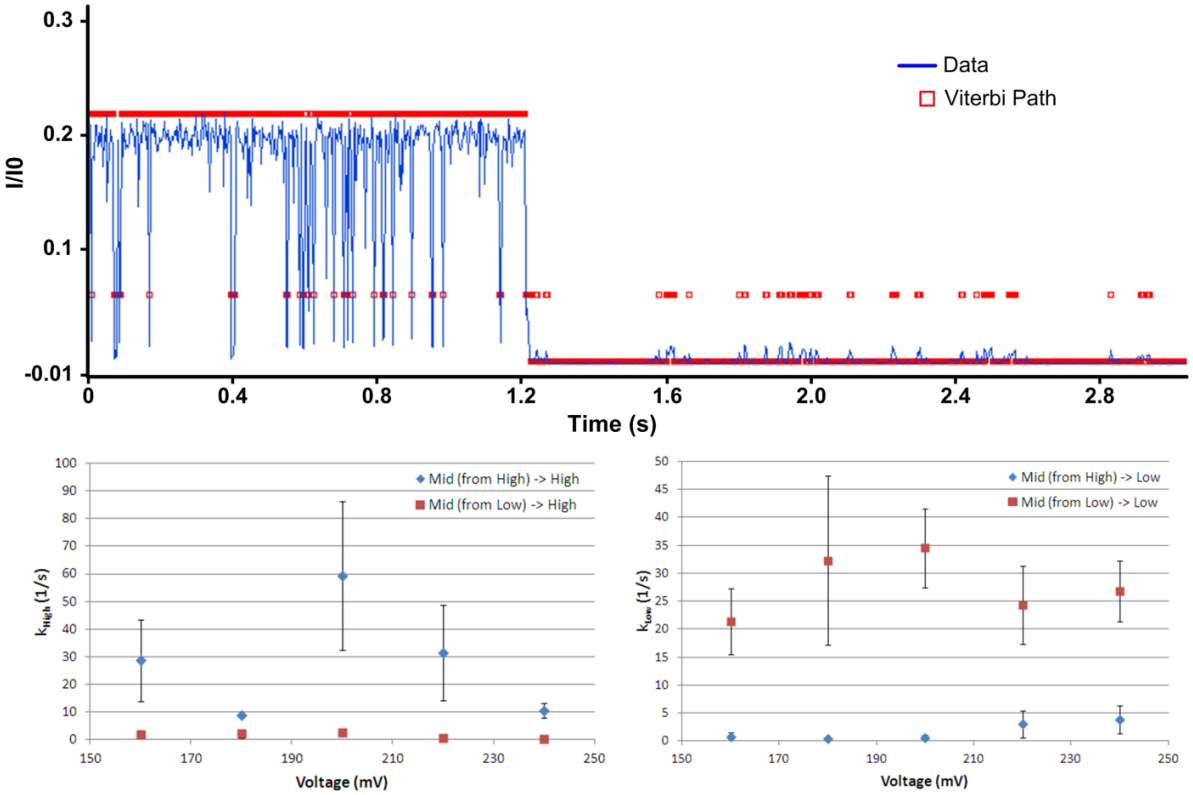


Figure 4-9. KKRR-ShPrP(120-232) mid-state statistics as a function of voltage. (Top) A sample KKRR-ShPrP(120-232) event (blue) with the most likely state sequence (i.e. Viterbi path) overlayed (red) as determined by a Viterbi analysis of the event. The sample event highlights the dependence of mid-state statistics (i.e. the transition rates out of the mid-state) on how the mid-state is entered. The event qualitatively shows that if the mid-state is entered from the high state then a transition back to the high-state is more likely than a transition into the low-state. Likewise, transitions into the mid-state from the low-state are more likely to return to the low-state as opposed to entering the high-state. **(Bottom left)** Mid-state transition rate into the high-state as a function of voltage, depending on how the mid-state is entered. If the mid-state is entered from the high-state (blue) the transition rate back into the high-state is between one and two orders of magnitude higher (depending on voltage) than the transition rate into the high-state when the mid-state is entered from the low-state (red). **(Bottom right)** Mid-state transition rate into the low-state as a function of voltage, depending on how the mid-state is entered. If the mid-state is entered from the low-state (red) the transition rate back into the low-state is between one and two orders of magnitude higher (depending on voltage) than the transition rate into the low-state when the mid-state is entered from the high-state (blue).

In general for a single state, we would expect the transition rates in figure 4-9 (bottom left) and figure 4-9 (bottom right) to be within an error bar of each other (assuming the rates

are Gaussian distributed). Given their degree of separation, between two and four error bars depending on voltage, these results indicate that the mid-state is more accurately modeled as two separate states a mid-high and a mid-low state, to distinguish between transitions into the mid-state from the high-state (mid-high state) versus mid-state transitions from the low-state (mid-low state). Given this, together with the results from the HMM analysis of the data (i.e. the state-to-state transition probabilities), we can model KKRR-ShPrP(120-232) kinetics in the pore as a four-state system (figure 4-10).

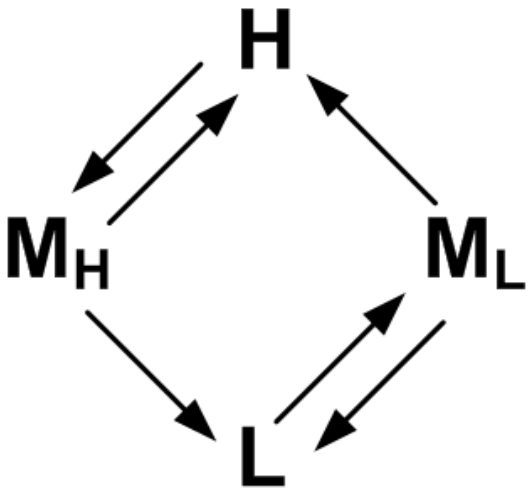


Figure 4-10. Four-state model characterizing KKRR-ShPrP(120-232) kinetics in the pore. H, M_H, L, and M_L refer to the high, mid-high, low, and mid-low states respectively (N.B. PrP^C can escape from the pore from each state, which is not explicitly shown in the four-state model). HMM analysis of KKRR-ShPrP(120-232) in combination with the mid-state analysis (refer to text) yields the information on how the states are connected.

Given the detailed kinetics of wild-type PrP^C in the pore, the voltage-dependence of all the state-to-state transition rates can be determined, which may yield information on the different conformations of PrP^C in the pore. For example, transitions that exhibit an exponential dependence on voltage (i.e. Arrhenius kinetics) indicate energy barrier crossing

processes and therefore yield clues on the types of conformations and conformational transitions that can makeup said processes. This together with computational modeling of PrP^C trapped inside the pore should reveal detailed information on the specific conformations of PrP^C in the pore. With respect to D178N mutant kinetics and how it compares with wild-type we find that given the substantial difference in state properties between these two constructs (i.e. figures 4-6b and 4-7b) no simple comparisons can be made. As mentioned previously, residue D178 plays an important role in maintaining the structural stability of PrP^C. This loss of stability, in the case of the mutant, likely enables it to adopt a variety of conformations inside the pore that are inaccessible to wild-type, which complicates the comparison between these two constructs, and hints at the need for additional states in a description of mutant kinetics in the pore. In order to make meaningful comparisons with wild-type PrP^C, therefore, a more detailed analysis of mutant kinetics, including additional states in the initial HMM model, is required.

4.4 Chapter conclusion

Nanopore protein analysis is emerging as a powerful tool for studying protein biophysics. The technique enables label-free analysis of individual proteins in a near-native environment making it an attractive technology for medical diagnostics. It can also yield significant insights on the properties of proteins, particularly rare and / or transient phenomena that are typically hidden by ensemble methods. An exciting potential application of nanopore sensing is in the study of protein dynamics. Single molecule methods for probing the long timescale motions of proteins, known to be critically important to protein function, can yield significant insights that may otherwise be unattainable by other methods.

In this way, nanopore sensing can potentially become a complementary tool to existing methods for the analysis of protein dynamics. Challenges to nanopore sensing of proteins include improving the capture and trapping of low-charge density proteins in the pore and development of sophisticated data analysis methods to characterize the complex kinetics and conformational dynamics of proteins. We have presented methods that address both of these challenges (namely nanopore analysis in the presence of a salt concentration gradient and HMM signal processing) and have applied these methods to probe the structure and dynamics of wild-type and D178N mutant PrP^C captured in an α-HL nanopore. We have shown that these two proteins (proteins that differ by only a single amino-acid point mutation) exhibit easily distinguishable current signatures and kinetics inside the pore and have demonstrated accurate detection and discrimination between these two proteins at the single molecule level based on the kinetics of a single PrP^C capture event. This method of protein analysis may be useful as an assay to probe differences in stability between wild-type and mutant prion proteins at the single molecule level, opening up the possibility to study small molecule-PrP^C interactions and their effects on PrP^C stability as a possible screen for small molecules that improve the stability properties of the protein. Moreover, our results demonstrate the sensitivity of nanopore analysis in detecting subtle changes in biomolecule structure and show its potential for highly sensitive, real-time protein and small molecule detection and identification based on single molecule kinetics inside a nanopore with potential applications in disease biomarker and pathogen detection. In addition, we developed a four-state model to characterize wild-type PrP^C kinetics in the pore which represents a first step in our investigation on characterizing the differences in kinetics and conformational dynamics between wild-type and D178N mutant PrP^C, a comparison of

which may ultimately yield clues into the molecular mechanism of PrP^C conversion in disease. We have presented methods to enable the detailed analysis of the kinetics and conformational dynamics of single proteins using nanopores, which points to the potential of using this technique to probe the molecular properties of other clinically relevant proteins.

Chapter 5

Conclusion

Single molecule methods are of fundamental importance to understanding and characterizing the detailed properties and dynamics of biological systems. Nanopore sensing is a relatively new single molecule technique offering unique advantages compared to other methods, specifically the possibility of label-free detection and analysis of individual biomolecules in a near-native environment. Nanopore methods have therefore raised much excitement in the scientific community and have attracted significant investment from both industry and government that continues to this day [171]. This is particularly true for nanopore DNA analysis where nanopore technology is being actively developed for low-cost, high-throughput whole-genome sequencing with the potential to revolutionize the practice of medicine by enabling genetic analysis to become common-place in diagnostics and therapeutics.

The aim of this thesis work was to advance the field of nanopore sensing, extending its applications and enabling the development of improved methods for analysis and sequencing of DNA. In working toward this aim we have achieved three fundamental advances:

Firstly, we have provided an improved understanding of how DNA translocates nanopores, particularly in cases where the DNA resides in the pore for substantial amounts of time, leading to subdiffusive translocation dynamics. We showed that translocation, in this

case, is highly sequence dependent and that stochastic DNA-pore binding interactions leads to subdiffusive motion of DNA through the pore. The model and methods presented to characterize translocation dynamics may aid in the design of mechanisms to slow DNA in its passage through a nanopore for the purpose of improving the resolution and accuracy of DNA sequencing.

Secondly, in demonstrating that translocation dynamics are highly sequence dependent, our results raise the possibility of using nucleotide dwell-time in the pore as an additional metric to nucleotide blockage current to distinguish individual nucleotides in nanopore sequencing, opening up new avenues by which to optimize the accuracy and performance of nanopore DNA sequencing technologies.

Finally, we have demonstrated the ability to capture and trap individual proteins in a nanopore for sufficient time to extract and characterize a signal related to protein dynamics as it undergoes thermally driven changes in its conformation with respect to the pore. We show that protein dynamics are influenced by subtle changes in protein sequence, and that the resulting signal can thus be used to identify individual proteins in solution, specifically distinguishing a mutant form of the prion protein from that of wild-type.

It is a testament to the advances made in this thesis, that by the end of the experimental program we pursued, it became a routine and simple procedure to capture a single molecule of DNA or protein in the pore, study it for sufficient time to extract information specific to that individual molecule, and expel it from the pore to allow the process to be repeated on another molecule, extracting information from individual molecules in solution one at a time. We believe this analysis methodology is a glimpse of future analytical methods, that shed

light on details of heterogeneous populations, by analyzing individual molecules one at a time.

.

..

Bibliography

1. Walter NG, Huang C, Manzo AJ, Sobhy MA (2008) Do-it-yourself guide: how to use the modern single-molecule toolkit. *Nat. Methods*. **5**: 475-489.

2. Cornish PV, Ha T (2007) A survey of single-molecule techniques in chemical biology. *ACS Chem. Biol*. **2**: 53-61.

3. Ritort F (2006) Single-molecule experiments in biological physics: methods and applications. *J. Phys.: Condens. Matter*. **18**: R531-R583.

4. Greenleaf WJ, Woodside MT, Block SM (2007) High-resolution, single-molecule measurements of biomolecular motion. *Annu. Rev. Biophys. Biomol. Struct*. **36**: 171-190.

5. Neuman KC, Nagy A (2008) Single-molecule force spectroscopy: optical tweezers, magnetic tweezers, and atomic force microscopy. *Nat. Methods*. **5**: 491-505.

6. Strick T, Allemand J, Croquette V, Bensimon D (2001) The manipulation of single biomolecules. *Phys. Today*. 46-51.

7. Svoboda K, Block SM (1994) Biological applications of optical forces. *Annu. Rev. Biophys. Biomol. Struct*. **23**: 247-285.

8. Ashkin A, Dziedzic JM, Bjorkholm JE, Chu S (1986) Observation of a single-beam gradient force optical trap for dielectric particles. *Opt. Lett*. **11**: 288-290.

9. Roy R, Hohng S, Ha T (2008) A practical guide to single-molecule FRET. *Nat. Methods*. **5**: 507-516.

10. Finzi L, Gelles J (1995) Measurement of lactose repressor-mediated loop formation and breakdown in single DNA molecules. *Science*. **267**: 378-380.

11. Pouget N, Dennis C, Turlan C, Grigoriev M, Chandler M, et. al. (2004) Single-particle tracking for DNA tether length monitoring. *Nucleic Acids Res*. **32**: e73.

12. Wanunu M (2012) Nanopores: A journey towards DNA sequencing. *Phys. Life Rev*. **9**: 125-158.

13. Ma L, Cockroft SL (2010) Biological nanopores for single molecule biophysics. *ChemBioChem*. **11**: 25-34.

14. DeGuzman VS, Lee CC, Deamer DW, Vercoutere WA (2006) Sequence-dependent gating of an ion channel by DNA hairpin molecules. *Nucleic Acids Res.* **34**: 6425-6437.

15. Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, et. al. (2012) Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* **30**: 349-353.

16. Lin J, Kolomeisky A, Meller A (2010) Helix-coil kinetics of individual polyadenylic acid molecules in a protein channel. *Phys. Rev. Lett.* **104**: 158101.

17. Graham MD (2003) The Coulter principle: Foundation of an industry. *J. Lab. Autom.* **8**: 72-81.

18. Wiggin M, Tropini C, Tabard-Cossa V, Jetha NN, Marziali A (2008) Nonexponential kinetics of DNA escape from α-hemolysin nanopores. *Biophys J.* **95**: 5317-5323.

19. Jetha NN, Feehan C, Wiggin M, Tabard-Cossa V, Marziali A (2011) Long dwell-time passage of DNA through nanometer-scale pores: Kinetics and sequence dependence of motion. *Biophys. J.* **100**: 2974-2980.

20. Wiggin MJ (2009) Genotyping by nanopore force spectroscopy: Method development and evaluation for clinical diagnostics (PhD Thesis). Retrieved from https://circle.ubc.ca/.

21. Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotides using a membrane channel. *Proc. Natl. Acad. Sci. USA.* **93**: 13770-13773.

22. Akeson M, Branton D, Kasianowicz JJ, Brandin E, Deamer DW (1999) Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophys J.* **77**: 3227-3233.

23. Meller A, Nivon L, Brandin E, Golovchenko J, Branton D (2000) Rapid nanopore discrimination between single polynucleotide molecules. *Proc. Natl. Acad. Sci. USA.* **97**: 1079-1084.

24. Metzker ML (2005) Emerging technologies in DNA sequencing. *Genome Res.* **15**: 1767-1776.

25. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed November 21, 2013.

26. Cordero P, Ashley EA (2012) Whole-genome sequencing in personalized therapeutics. *Clin. Pharmacol. Ther*. **91**: 1001-1009.

27. Pasche B, Absher D (2011) Whole-genome sequencing. A step closer to personalized medicine. *JAMA*. **305**: 1596-1597.

28. National Human Genome Research Institute (2013) Revolutionary genome sequencing technologies - the $1000 genome. Retrieved from http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-10-014.html

29. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, et. al. (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol*. **26**: 1146-1153.

30. Venkatesan BM, Bashir R (2011) Nanopore sensors for nucleic acid analysis. *Nat. Nanotechnol*. **6**: 615-624.

31. Lubensky DK, Nelson DR (1999) Driven polymer translocation through a narrow pore. *Biophys. J*. **77**: 1824-1838.

32. Metzler R, Klafter J (2003) When translocation dynamics becomes anomalous. *Biophys. J*. **85**: 2776-2779.

33. Redner S (2001) A guide to first-passage processes. Cambridge University Press, Cambridge, UK.

34. Bates M, Burns M, Meller A (2003) Dynamics of DNA molecules in a membrane channel probed by active control techniques. *Biophys. J*. **84**: 2366-2372.

35. Metzler R, Klafter J (2000) The random walk's guide to anomalous diffusion: A fractional dynamics approach. *Phys. Rep*. **339**: 1-77.

36. Metzler R, Klafter J (2004) The restaurant at the end of the random walk: recent development in the description of anomalous transport by fractional dynamics. *J. Phys. A: Math. Gen*. **37**: R161-R208.

37. Chatterjee D, Cherayil BJ (2010) Anomalous reaction-diffusion as a model of nonexponential DNA escape kinetics. *J. Chem. Phys.* **132**: 025103 (5pp).

38. Huttenhain R, Malmstrom J, Picotti P, Aebersold R (2009) Perspective of targeted mass spectrometry for protein biomarker verification. *Curr. Opin. Chem. Biol*. **13**: 1-8.

39. Stastna M, Van Eyk JE (2012) Secreted proteins as a fundamental source for biomarker discovery. *Proteomics*. **12**: 722-735.

40. Payet L, Martinho M, Pastoriza-Gallego M, Betton JM, Auvray L, et. al. (2012) Thermal unfolding of proteins probed at the single molecule level using nanopores. *Anal. Chem*. **84**: 4071-4076.

41. Merstorf C, Cressiot B, Pastoriza-Gallego M, Oukhaled A, Betton JM, et. al. (2012) Wild-type, mutant protein unfolding and phase transition detected by single nanopore recording. *ACS Chem. Biol*. **7**: 652-658.

42. Movileanu L, Schmittschmitt JP, Scholtz JM, Bayley H (2005) Interactions of peptides with a protein pore. *Biophys. J*. **89**: 1030-1045.

43. Mohammad MM, Movileanu L (2008) Excursion of a single polypeptide into a protein pore: Simple physics but complicated biology. *Eur. Biophys. J*. **37**: 913-925.

44. Mohammad MM, Prakash S, Matouschek A, Movileanu L (2008) Controlling a single protein in a nanopore through electrostatic traps. *J. Am. Chem. Soc*. **130**: 4081-4088.

45. Soskine M, Biesemans A, Moeyaert B, Cheley S, Bayley H, et. al. (2012) An Engineered ClyA Nanopore Detects Folded Target Proteins by Selective External Association and Pore Entry. *Nano Lett*. **12**: 4895-4900.

46. Pastoriza-Gallego M, Rabah L, Gibrat G, Thiebot B, Gisou van der Gout F, et. al. (2011) Dynamics of unfolded protein transport through an aerolysin pore. *J. Am. Chem. Soc*. **133**: 2923-2931.

47. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. *Natur*e. **450**: 964-972.

48. Yang H, Luo G, Karnchanaphanurach P, Louie T, Rech I, et. al. (2003) Protein conformational dynamics probed by single-molecule electron transfer. *Science*. **302**: 262-266.

49. Movileanu L (2009) Interrogating single proteins through nanopores: Challenges and opportunities. *Trends Biotechnol*. **27**: 333-341.

50. Wanunu M, Morrison W, Rabin Y, Grosberg AY, Meller A (2010) Electrostatic focusing of unlabelled DNA into nanoscale pores using a salt gradient. *Nat. Nanotechnol*. **5**: 160-165.

51. Hatlo MM, Panja D, Roij RV (2011) Translocation of DNA molecules through nanopores with salt gradients: The role of osmotic flow. *Phys. Rev. Lett*. **107**: 068101.

52. Chung SH, Moore JB, Xia L, Premkumar LS, Gage PW (1990) Characterization of single channel currents using digital signal processing techniques based on hidden markov models. *Phil. Trans. R. Soc. Lond. B*. **329**: 265-285.

53. Churbanov A, Winters-Hilt S (2008) Clustering ionic flow blockade toggles with a mixture of HMMs. *BMC Bioinf.* **9**:S13.

54. Venkataramanan L, Sigworth FJ (2002) Applying hidden markov models to the analysis of single ion channel activity. *Biophys. J.* **82**: 1930-1942.

55. Collinge J (2001) Prion diseases of humans and animals: Their causes and molecular basis. *Annu. Rev. Neurosci*. **24**: 519-550.

56. Prusiner SB (1982) Novel proteinaceous infectious particles cause scrapie. *Science.* **216**: 136-144.

57. Prusiner SB (1998) Prions. *Proc. Natl. Acad. Sci. USA.* **95**: 13363-13383.

58. Coulthart MB, Cashman NR (2001) Variant Creutzfeldt-Jakob disease: A summary of current scientific knowledge in relation to public health. *Can. Med. Assoc. J.* **165**: 51-58.

59. Coulter WH. US Patent #2,656,508 (1953).

60. Madampage CA (2011) The first step towards the development of an electrophoretic prion detector (PhD Thesis). Retrieved from http://hdl.handle.net/10388/etd-08082011-162258.

61. Mehta S, Hunter DJ, Mugusi FM, Spiegelman D, Manji KP, et. al. (2009) Perinatal outcomes, including mother-to-child transmission of HIV, and child mortality and their association with maternal vitamin D status in Tanzania. *J. Infect. Dis*. **200**: 1022-1030.

62. Bhat UG, Gartel AL (2008) Differential sensitivity of human colon cancer cell lines to the nucleoside analogs ARC and DRB. *Int. J. Cancer*. **122**: 1426-1429.

63. Ramon R, Sawadogo D, Koko FS, Noba V, Likikouet R, et. al. (1999) Haematological characteristics and HIV status of pregnant women in Abidjan, Cote d'Ivoire, 1995-96. *Trans. R. Soc. Trop. Med. Hyg*. **93**: 419-422.

64. Deblois RW, Bean CP, Wesley RKA (1977) Electrokinetic measurements with submicron particles and pores by resistive pulse technique. *J. Colloid. Interface. Sci*. **61**:323–35.

65. Sakmann B, Neher E (1995) Single-Channel Recording, 2nd ed., Plenum Press, New York.

66. Bezrukov SM, Vodyanoy I, Parsegian VA (1994) Counting polymers moving through a single ion channel. *Nature*. **370:** 279–281.

67. Smeets RMM, Keyser UF, Krapf D, Wu M, Dekker NH, et. al. (2006) Salt dependence of ion transport and DNA translocation through solid-state nanopores. *Nano Lett*. **6**: 89-95.

68. Chang H, Kosari F, Andreadakis G, Alam MA, Vasmatzis G, et. al. (2004) DNA-mediated fluctuations in ionic current through silicon oxide nanopore channels. *Nano Lett.* **4**: 1551-1556.

69. Meller A, Branton D (2002) Single molecule measurements of DNA transport through a nanopore. *Electrophoresis*. **23**: 2583-2591

70. Meller A, Nivon L, Branton D (2001) Voltage-driven DNA translocations through a nanopore. *Phys. Rev. Lett*. **86**: 3435-3438.

71. Howorka S, Siwy Z (2009) Nanopore analytics: Sensing of single molecules. *Chem. Soc. Rev.* **38**: 2360-2384.

72. Song L, Hobaugh MR, Shustak C, Cheley S, Bayley H, et. al. (1996) Structure of Staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science*. **274**: 1859-1866.

73. Aksimentiev A, Schulten K (2005) Imaging α-hemolysin with molecular dynamics: Ionic conductance, osmotic permeability, and the electrostatic potential map. *Biophys. J*. **88**: 3745-3761.

74. Stahl C, Kubetzko S, Kaps I, Seeber S, Engelhardt H, et. al. (2001) MspA provides the main hydrophilic pathway through the cell wall of Mycobacterium smegmatis. *Mol. Microbiol*. **40**:451-464.

75. de Zoysa RS, Krishantha DMM, Zhao Q, Gupta J, Guan X (2011) Translocation of single stranded DNA through the α-hemolysin protein nanopore in acidic solutions. *Electrophoresis*. **32**: 3034-3041.

76. Maglia G, Henricus M, Wyss R, Li Q, Cheley S, et. al. (2009) DNA strands from denatured duplexes are translocated through engineered protein nanopores at alkaline pH. *Nano Lett*. 9: 3831-3836.

77. Haque F, Li J, Wu H, Liang X, Guo P (2013) Solid state and biological nanopore for real time sensing of single chemical and sequencing of DNA. *Nano Today*. **8**: 56-74.

78. Kang X, Gu L, Cheley S, Bayley H (2005) Single protein pores containing molecular adapters at high temperatures. *Angew. Chem. Int. Ed*. **44**: 1495-1499.

79. Oukhaled G, Mathe J, Biance AL, Bacri L, Betton JM, et. al. (2007) Unfolding of proteins and long transient conformations detected by single nanopore recording. *Phys. Rev. Lett*. **98**: 158101-1 - 158101-4.

80. Rincon-Restrepo M, Mikhailova E, Bayley H, Maglia G (2011) Controlled translocation of individual DNA molecules through protein nanopores with engineered molecular brakes. *Nano Lett.* **11**: 746-750.

81. Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, et. al. (2010) Nanopore DNA sequencing with MspA. *Proc. Natl. Acad. Sci. USA.* **107**: 16060-16065.

82. Stefureac R, Waldner L, Howard P, Lee JS (2008) Nanopore analysis of a small 86-residue protein. *Small*. **4**: 59-63.

83. Stefureac R, Long Y, Kraatz H, Howard P, Lee JS (2006) Transport of α-helical peptides through α-hemolysin and aerolysin pores. *Biochemistry*. **45**: 9172-9179.

84. Payet L, Martinho M, Pastoriza-Gallego M, Betton J, Auvray L, et. al. (2012) Thermal unfolding of proteins probed at the single molecule level using nanopores. *Anal. Chem*. **84**: 4071-4076.

85. White RJ, Ervin EN, Yang T, Chen X, Daniel S, et. al. (2007) Single ion channel recordings using glass nanopore membranes. *J. Am. Chem. Soc*. **129**: 11766-11775.

86. Tabard-Cossa V, Trivedi D, Wiggin M, Jetha NN, Marziali A (2007) Noise analysis and reduction in solid-state nanopores. *Nanotechnology*. **18**: 305505 (6pp).

87. Meller A (2003) Dynamics of polynucleotide transport through nanometer-scale pores. *J. Phys.: Condens. Matter*. **15**: R581-R607.

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

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

90. Fologea D, Uplinger J, Thomas B, McNabb DS, Li J (2005) Slowing DNA translocation in a solid-state nanopore. *Nano Lett.* **5**: 1734-1737.

91. de Zoysa RSS, Jayawardhana DA, Zhao Q, Wang D, Armstrong DW, et. al. (2009) Slowing DNA translocation through nanopores using a solution containing organic salts. *J. Phys. Chem. B*. **113**: 13332-13336.

92. Mitchell N, Howorka S (2008) Chemical tags facilitate the sensing of individual DNA strands with nanopores. *Angew. Chem. Int. Ed*. **47**: 5565-5568.

93. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, et. al. (2012) Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol*. **30**: 344-348.

94. Deamer D (2010) Nanopore analysis of nucleic acids bound to exonucleases and polymerases. *Annu. Rev. Biophys*. **39**: 79-90.

95. Clarke J, Wu H, Jayasinghe L, Patel A, Reid S, et. al. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol*. **4**: 265-270.

96. Tsutsui M, Taniguchi M, Yokota K, Kawai T (2010) Identifying single nucleotides by tunneling current. *Nat. Nanotechnol*. **5**: 286-290.

97. Huang S, He J, Chang S, Zhang P, Liang F, et. al. (2010) Identifying single bases in a DNA oligomer with electron tunneling. *Nat. Nanotechnol*. **5**: 868-873.

98. Tanaka H, Kawai T (2009) Partial sequencing of a single DNA molecule with a scanning tunneling microscope. *Nat. Nanotechnol*. **4**: 518-522.

99. Zwolak M, Di Ventra M (2008) Physical approaches to DNA sequencing and detection. *Rev. Mod. Phys*. **80**: 141-165.

100. Ling XS, Bready B, Pertsinidis A (2007) Hybridization-assisted nanopore sequencing of nucleic acids. US patent application no. 2007 0190542.

101. McNally B, Singer A, Yu Z, Sun Y, Weng Z, et. al. (2010) Optical recognition of converted DNA nucleotides for single molecule DNA sequencing using nanopore arrays. *Nano Lett*. **10**: 2237-2244.

102. Luan B, Peng H, Polonsky S, Rossnagel S, Stolovitsky G, et. al. (2010) Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. *Phys. Rev. Lett*. **104**: 238103-1 - 238103-4.

103. Nelson T, Zhang B, Prezhdo OV (2010) Detection of nucleic acids with graphene nanopores: Ab initio characterization of a novel sequencing device. *Nano Lett*. **10**: 3237 - 3242.

104. Min S, Kim W, Cho Y, Kim K (2011) Fast DNA sequencing with a graphene-based nanochannel device. *Nat. Nanotechnol*. **6**: 162-165.

105. Merchant CA, Healy K, Wanunu M, Ray V, Peterman N, et. al. (2010) DNA translocation through graphene nanopores. *Nano Lett*. **10**: 2915-2921.

106. Maitra RD, Kim J, Dunbar WB (2012) Recent advances in nanopore sequencing. *Electrophoresis*. **33**: 3418-3428.

107. Muthukumar M (2011) Polymer translocation. Boca Raton: CRC Press. 353 p.

108. Metzler R, Klafter J (2000) From a generalized Chapman-Kolmogorov equation to the fractional Klein-Kramers equation. *J. Phys. Chem. B*. **104**: 3851-3857.

109. Bouchaud JP, Potters M (2003) Theory of financial risk and derivative pricing: From statistical physics to risk management. New York: Cambridge University Press. 379 p.

110. Chaudhury S, Cherayil BJ (2008) A model of anomalous chain translocation dynamics. *J. Phys. Chem. B*. **112**: 15973-15979.

111. Chuang J, Kantor Y, Kardar M (2001) Anomalous dynamics of translocation. *Phys. Rev. E*. **65**: 011802 (8pp).

112. Panja D (2009) Response of single polymers to localized step strains. *Phys. Rev. E*. **79**: 011803 (9pp).

113. Panja D, Barkema GT (2008) Passage times for polymer translocation pulled through a narrow pore. *Biophys. J*. **94**: 1630 - 1637.

114. Panja D, Barkema GT, Ball RC (2007) Anomalous dynamics of unbiased polymer translocation through a narrow pore. *J. Phys.: Condens. Matter*. **19**: 432202 (8pp).

115. Tian P, Smith GD (2003) Translocation of a polymer chain across a nanopore: A Brownian dynamics simulation study. *J. Chem. Phys.* **119**: 11475 - 11483.

116. Bruskov V (1978) Specificity of interaction of nucleic acid bases with hydrogen bond forming amino acids. *Stud. Biophys. (Berlin)*. **67S**:43-44.

117. Wallace EVB, Stoddart D, Heron AJ, Mikhailova E, Maglia G, et. al. (2010) Identification of epigenetic DNA modifications with a protein nanopore. *Chem. Commun*. **46**: 8195-8197.

118. Tropini C, Marziali A (2007) Multi-nanopore force spectroscopy for DNA analysis. *Biophys J*. **92**:1632-1637.

119. Jetha NN, Wiggin M, Marziali A (2009) Nanopore force spectroscopy on DNA duplexes, in *Micro and Nano Technologies in Bioanalysis* (Lee, J.W., and Foote, R.S., ed.), Humana Press, Totowa, NJ.

120. Tabard-Cossa V, Wiggin M, Trivedi D, Jetha NN, Dwyer JR, et. al. (2009) Single molecule bonds characterized by solid-state nanopore force spectroscopy. *ACS Nano*. **3**:3009-3014.

121. Henrickson SE, Misakian M, Robertson B, Kasianowicz JJ (2000). Driven DNA transport into an asymmetric nanometer-scale pore. *Phys. Rev. Lett*. **85**: 3057-3060.

122. Liebovitch LS, Toth TI (1991) Distributions of activation energy barriers that produce stretched exponential probability distributions for the time spent in each state of the two state reaction A ↔ B. *B. Math. Biol.* **53**:443-455.

123. Liebovitch LS, Fischbarg J, Koniarek JP (1987) Ion channel kinetics: A model based on fractal scaling rather than multistate markov processes. *Math. Biosci.* **84**: 37-68.

124. Austin RH, Beeson KW, Eisenstein L, Frauenfelder H, Gunsalus IC (1975) Dynamics of ligand-binding to myoglobin. *Biochemistry*. **14**: 5355-5373.

125. Meerschaert MM, Nane E, Xiao Y (2008) Correlated continuous time random walks. *arXiv*: 0809.1612.

126. Sakaue T (2010) Sucking genes into pores: Insight into driven translocation. *Phys. Rev. E*. **81**: 041808 (6pp).

127. Kantor Y, Kardar M (2004) Anomalous dynamics of forced translocation. *Phys. Rev. E*. **69**: 021806 (12pp).

128. Sutherland TC, Long Y, Stefureac R, Bediako-Amoa I, Kraatz H, et. al. (2004) Structure of peptides investigated by nanopore analysis. *Nano Lett*. **4**: 1273-1277.

129. Goodrich CP, Kirmizialtin S, Huyghues-Despointes BM, Zhu A, Scholtz JM, et. al. (2007) Single-molecule electrophoresis of β-hairpin peptides by electrical recordings and Langevin dynamics simulations. *J. Phys. Chem. B*. **111**: 3332-3335.

130. Oukhaled A, Bacri L, Pastoriza-Gallego M, Betton J, Pelta J (2012) Sensing proteins through nanopores: Fundamental to applications. *ACS Chem. Biol*. **7**: 1935-1949.

131. Bayley H, Jayasinghe L (2004) Functional engineered channels and pores (Review). *Mol. Membr. Biol*. **21**: 209-220.

132. Wolfe AJ, Mohammad M, Cheley S, Bayley H, Movileanu L (2007) Catalyzing the translocation of polypeptides through attractive interactions. *J. Am. Chem. Soc*. **129**: 14034-14041.

133. Gobl C, Tjandra N (2012) Application of solution NMR spectroscopy to study protein dynamics. *Entropy*. **14**: 581-598.

134. Schotte F, Soman J, Olson JS, Wulff M, Anfinrud PA (2004) Picosecond time-resolved X-ray crystallography: Probing protein function in real-time. *J. Struct. Biol*. **147**: 235-246.

135. Cammarata M, Levantino M, Schotte F, Anfinrud PA, Ewald F, et. al. (2008) Tracking the structural dynamics of proteins in solution using time-resolved wide-angle X-ray scattering. *Nat. Methods*. **5**: 881-886.

136. Aguzzi A, Calella AM (2009) Prions: Protein aggregation and infectious diseases. *Physiol. Rev*. **89**: 1105-1152.

137. Sharma S, Mukherjee M, Kedage V, Muttigi MS, Rao A (2009) Sporadic Creutzfeldt-Jakob disease - A review. *Int. J. Neuroscience*. **119**: 1981-1994.

138. Guest WC (2012) Template-directed protein misfolding in neurodegenerative disease (PhD Thesis). Retrieved from https://circle.ubc.ca/.

139. Nicolas O, Gavin R, del Rio JA (2009) New insights into cellular prion protein (PrP^{C}) functions: The "ying and yang" of a relevant protein. *Brain Res. Rev*. **61**: 170-184.

140. Zahn R, Liu A, Luhrs T, Riek R, von Schroetter C, et. al. (2000) NMR solution structure of the human prion protein. *Proc. Natl. Acad. Sci. USA*. **97**: 145-150.

141. Soto C (2011) Prion hypothesis: the end of the controversy? *Trends Biochem. Sci.* **36**: 151-158.

142. Belay ED (1999) Transmissible spongiform encephalopathies in humans. *Annu. Rev. Microbiol*. **53**: 283-314.

143. Caughey B, Kocisko DA, Raymond GJ, Lansbury Jr. PT (1995) Aggregates of scrapie-associated prion protein induce the cell-free conversion of protease-sensitive prion protein to the protease-resistant state. *Chem. Biol*. **2**: 807-817

144. Caughey B, Raymond GJ, Kocisko DA, Lansbury Jr. PT (1997) Scrapie infectivity correlates with converting activity, protease resistance, and aggregation of scrapie-associated prion protein in guanidine denaturation studies. *J. Virol*. **71**: 4107-4110.

145. Pan KM, Baldwin M, Nguyen J, Gasset M, Serban A, et. al. (1993) Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. *Proc. Natl. Acad. Sci. USA*. **90**: 10962–10966.

146. Guest WC, Plotkin SS, Cashman NR (2011) Toward a mechanism of prion misfolding and structural models of PrP^{Sc}: Current knowledge and future directions. *J. Toxicol. Environ. Health Part A*. **74**: 154-160.

147. Aguzzi A (2009) Cell biology: Beyond the prion principle. *Nature*. **459**: 924–925.

148. Aguzzi A, Rajendran L (2009) The transcellular spread of cytosolic amyloids, prions, and prionoids. *Neuron*. **64**: 783-790.

149. Grad LI, Guest WC, Yanai A, Pokrishevsky E, O'Neill MA, et. al. (2011) Intermolecular transmission of superoxide dismutase 1 misfolding in living cells. *Proc. Natl. Acad. Sci. USA*. **108**: 16398-16403.

150. Soto C (2012) Transmissible proteins: Expanding the prion heresy. *Cell*. **149**: 968-977.

151. van der Kamp MW, Daggett V (2010) Pathogenic mutations in the hydrophobic core of the human prion protein can promote structural instability and misfolding. *J. Mol. Biol*. **404**: 732-748.

152. Vanik DL, Surewicz WK (2002) Disease-associated F198S mutation increases the propensity of the recombinant prion protein for conformational conversion to scrapie-like form. *J. Biol. Chem*. **277**: 49065-49070.

153. Meli M, Gasset M, Colombo G (2011) Dynamic diagnosis of familial prion diseases supports the β2-α2 loop as a universal interference target. *PLoS One*. **6**: e19093.

154. Julien O, Chatterjee S, Bjorndahl TC, Sweeting B, Acharya S, et. al. (2011) Relative and regional stabilities of the hamster, mouse, rabbit, and bovine prion proteins toward urea unfolding assessed by nuclear magnetic resonance and circular dichroism spectroscopies. *Biochemistry*. **50**: 7536-7545.

155. Bjorndahl TC, Zhou G, Liu X, Perez-Pineiro R, Semenchenko V, et. al. (2011) Detailed biophysical characterization of the acid-induced PrP^{C} to PrP^{β} conversion process. *Biochemistry*. **50**: 1162-1173.

156. Perez-Pineiro R, Bjorndahl TC, Berjanskii MV, Hau D, Li L, et. al. (2011) The prion protein binds thiamine. *FEBS J*. **278**: 4002-4014.

157. Nandi PK, Leclerc E, Marc D (2002) Unusual property of prion protein unfolding in neutral salt solution. *Biochemistry*. **41**: 11017-11024.

158. Apetri AC, Surewicz WK (2003) Atypical effect of salts on the thermodynamic stability of human prion protein. *J. Biol. Chem*. **278**: 22187-22192.

159. Movileanu L, Cheley S, Howorka S, Braha O, Bayley H (2001) Location of a constriction in the lumen of a transmembrane pore by targeted covalent attachment of polymer molecules. *J. Gen. Physiol*. **117**: 239-251.

160. Vercoutere WA, Winters-Hilt S, DeGuzman VS, Deamer D, Ridino SE, et. al. (2003) Discrimination among individual Watson-Crick base pairs at the termini of single DNA hairpin molecules. *Nucleic Acids Res*. **31**: 1311-1318.

161. Winters-Hilt S, Vercoutere W, DeGuzman VS, Deamer D, Akeson M, et. al. (2003) Highly accurate classification of watson-crick basepairs on termini of single DNA molecules. *Biophys. J*. **84**: 967-976.

162. Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*. **77**: 257-286.

163. Alonso DO, DeArmond SJ, Cohen FE, Daggett V (2001) Mapping the early steps in the pH-induced conformational conversion of the prion protein. *Proc. Natl. Acad. Sci. USA*. **98**: 2985-2989.

164. Riek R, Wider G, Billeter M, Hornemann S, Glockshuber R, et. al. (1998) Prion protein NMR structure and familial human spongiform encephalopathies. *Proc. Natl. Acad. Sci. USA*. **95**: 11667-11672.

165. Zuegg J, Gready JE (1999) Molecular dynamics simulations of human prion protein: Importance of correct treatment of electrostatic interactions. *Biochemistry*. **38**: 13862-13876.

166. Barducci A, Chelli R, Procacci P, Schettino V (2005) Misfolding pathways of the prion protein probed by molecular dynamics simulations. *Biophys. J*. **88**: 1334-1343.

167. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**: 861-874.

168. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. **30**: 1145-1159.

169. Scheeres K, Knoop H, Meer VDJ, Bleijenberg G (2009) Clinical assessment of the physical activity pattern of chronic fatigue syndrome patients: a validation of three methods. *Health Qual. Life Out.* **7**:29.

170. Kruger IM, Kruger MC, Doak CM, Kruger A (2012) Cut-off values of distal forearm bone density for the diagnosis of central osteoporosis in black postmenopausal South African women. *JEMDSA*. **17**: 78-83.

171. National Human Genome Research Institute (2013) New NIH awards focus on nanopore technology for DNA sequencing. Retrieved from http://www.genome.gov/27554929

172. Healy K, Schiedt B, Morrison AP (2007) Solid-state nanopore technologies for nanopore-based DNA analysis. *Nanomedicine*. **2**: 875-897.

173. Oukhaled A, Cressiot B, Bacri L, Patoriza-Gallego M, Betton J, et. al. (2011) Dynamics of completely unfolded and native proteins through solid-state nanopores as a function of electric driving force. *ACS Nano*. **5**: 3628-3638.

174. Fologea D, Ledden B, McNabb DS, Li J (2007) Electrical characterization of protein molecules by a solid-state nanopore. *Appl. Phys. Lett*. **91**: 053901 (3pp).

175. Han A, Creus M, Schurmann G, Linder V, Ward TR, et. al. (2008) Label-free detection of single protein molecules and protein-protein interactions using synthetic nanopores. *Anal. Chem*. **80**: 4651-4658.

176. Kowalczyk SW, Dekker C (2012) Measurement of the docking time of a DNA molecule onto a solid-state nanopore. *Nano Lett*. **12**: 4159-4163.

177. Wanunu M, Sutin J, McNally B, Chow A, Meller A (2008) DNA translocation governed by interactions with solid-state nanopores. *Biophys. J*. **95**: 4716-4725.

178. Langecker M, Pedone D, Simmel FC, Rant U (2011) Electrophoretic time-of-flight measurement of single DNA molecules with two stacked nanopores. *Nano Lett*. **11**: 5002-5007.

179. Yusko EC, Johnson JM, Majd S, Prangkio P, Rollings RC, et. al. (2011) Controlling protein translocation through nanopores with bio-inspired fluid walls. *Nat. Nanotechnol*. **6**: 253-260.

180. Tabard-Cossa V (2013) Instrumentation for low-noise high-bandwidth nanopore recording in *Engineered Nanopores for Bioanalytical Applications* (Edel JB, Albrecht T, ed.). Elsevier Inc.

181. Merchant CA, Healy K, Wanunu M, Ray V, Peterman N, et. al. (2010) DNA translocation through graphene nanopores. *Nano Lett*. **10**: 2915-2921.

182. Schneider GF, Kowalczyk SW, Calado VE, Pandraud G, Zandbergen HW, et. al. (2010) DNA translocation through graphene nanopores. *Nano Lett*. **10**: 3163-3167.

183. Hall AR, Scott A, Rotem D, Mehta KK, Bayley H, et. al. (2010) Hybrid pore formation by directed insertion of α-haemolysin into solid-state nanopores. *Nat. Nanotechnol*. **5**: 874-877.

184. Nakane J, Wiggin M, Marziali A (2004) A nanosensor for transmembrane capture and identification of single nucleic acid molecules. *Biophys J*. **87**: 615-621.

185. Shin S, Luchian T, Cheley S, Braha O, Bayley H (2002) Kinetics of a reversible covalent-bond-forming reaction observed at the single molecule level. *Angew. Chem. Int. Ed*. **41**: 3707 - 3709.

186. Luchian T, Shin S, Bayley H (2003) Kinetics of a three-step reaction observed at the single-molecule level. *Angew. Chem. Int. Ed*. **42**: 1925 - 1929.

Appendix A

Sequencing costs per human genome

The increasing need for rapid, inexpensive whole-genome sequencing has driven the development of new technologies that outperform traditional Sanger chain-termination methods in terms of speed and cost. These second- and third-generation sequencing technologies, stimulated in large-part by the $1,000 genome program instituted by the National Institutes of Health in 2004 [28,30], are expected to revolutionize medicine. Since the program's inception the cost for sequencing a whole human genome has fallen roughly according to Moore's Law[‡] until around 2008 when sequencing costs fell dramatically, out-pacing Moore's Law, which represented approximately the time when the sequencing centers transitioned from Sanger-based chain termination sequencing to second generation DNA sequencing technologies (figure A-1) [25].

[‡] Moore's Law is a useful metric for measuring the progress of technology improvements [25], which describes a trend observed in computing hardware involving the doubling of computing power every two years [25]. Technologies whose development trajectory follow Moore's Law are regarded as doing very well.

Appendix A. Sequencing costs per human genome

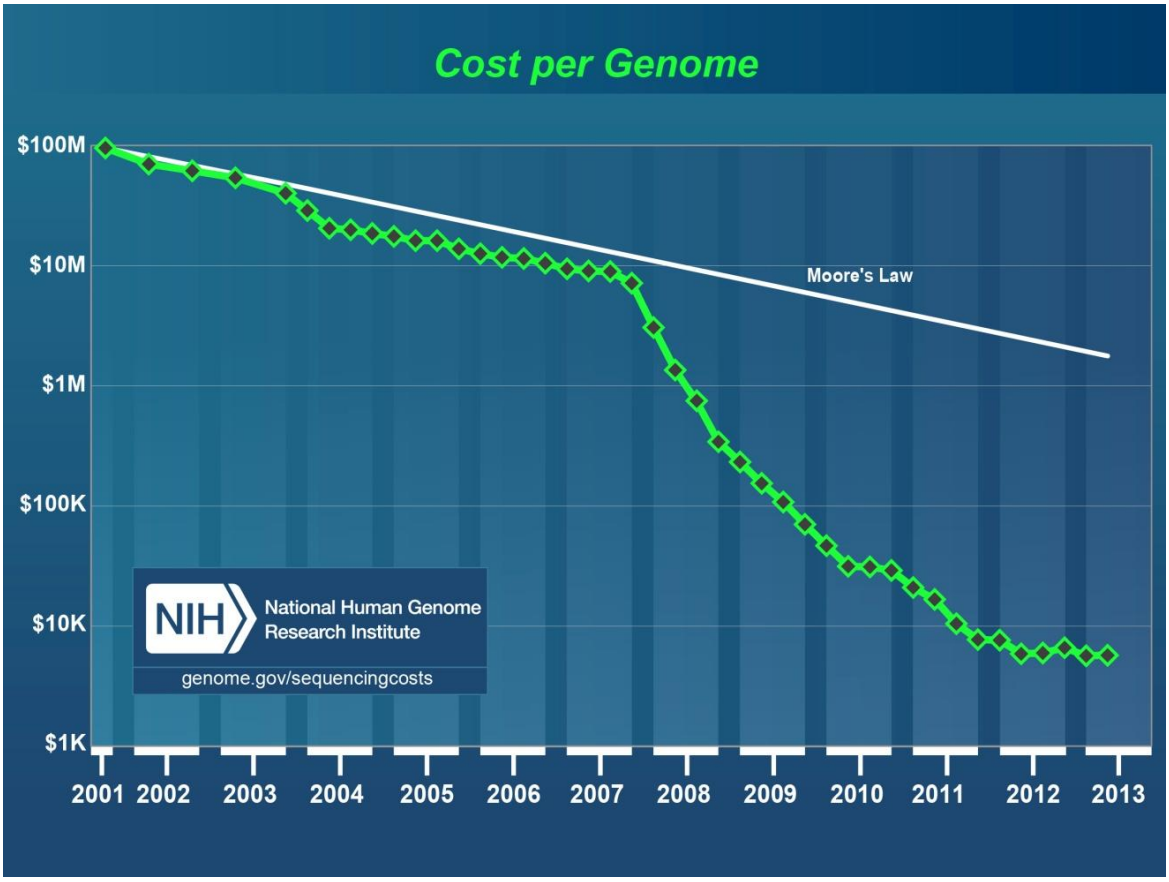


Figure A-1. Sequencing costs per human genome. The cost for sequencing a single human genome decreased roughly according to Moore's Law (white line) up until around 2008 when sequencing costs fell dramatically. This represents around the same time when sequencing centers transitioned from Sanger-based chain termination sequencing to second generation DNA sequencing technologies [25]. Image source: [25] (reproduced, with permission).

The task remains to reduce sequencing costs below $1000 per human genome, and in this regard nanopore sensing is arguably the most highly anticipated of several DNA sequencing technologies to meet this challenge.

Appendix B

Synthetic nanopores

Synthetic nanopores have become increasingly popular in nanopore sensing, offering a way to overcome many of the challenges associated with organic pores (refer to section 2.1 for discussion regarding the drawbacks of organic pores in nanopore sensing). Synthetic (or solid state) nanopores can be made by drilling a hole through a thin membrane of a suitable material (e.g. silicon nitride) with a high-energy electron beam and / or focused ion beam [20,172] (figure B-1). Synthetic pores offer some key advantages over their organic counterparts including tunable size [20,172] (e.g. by material deposition [172], or by exposure to ion, electron or photon beams which can contract or expand the pore [172]), ease of incorporation into instruments for device fabrication, stability over long periods of time (e.g. months) [20,86], durability under extreme conditions (e.g. low pH or very high voltages), and the potential to control their geometry and surface chemistry [172]. Solid-state nanopores have therefore yielded significant insights into the biophysical properties and translocation dynamics of both DNA and proteins [e.g. 86,173-178] particularly under conditions in which nanopore sensing with organic pores is impractical. Synthetic pores, however, also suffer drawbacks including low signal-to-noise ratios (they exhibit much greater noise than α-hemolysin [86]), and strong interactions of DNA and proteins with the pore, specifically silicon nitride nanopores, to the point where biomolecules "stick" to the pore surface causing irreversible blockages [20,179].

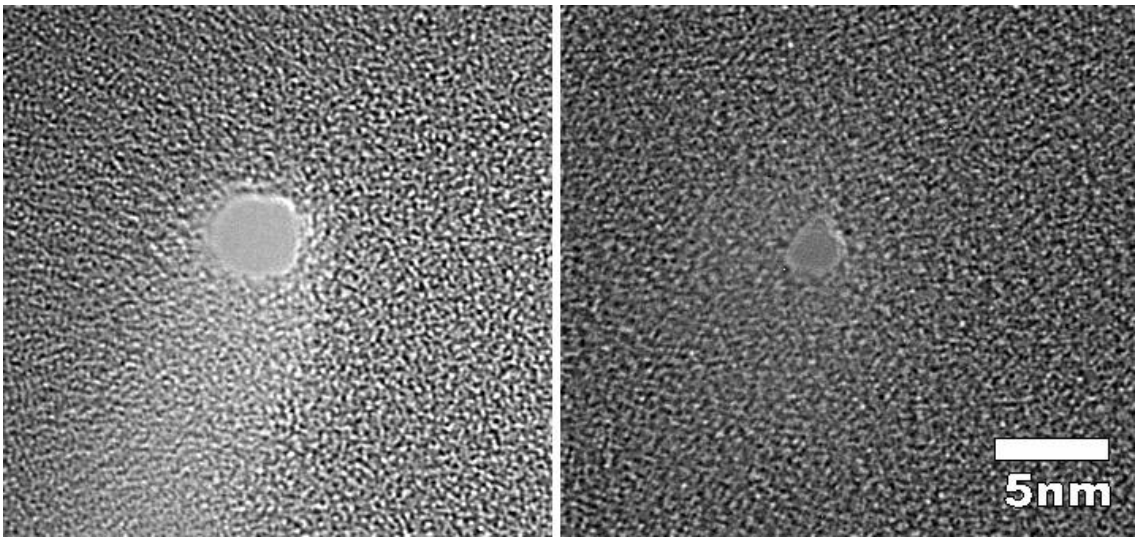


Figure B-1. Scanning electron micrograph of SiNx synthetic nanopores. (left) ~4 nm diameter nanopore. **(right)** ~2 nm diameter nanopore. Images and pores produced by Dhruti Trivedi.

The low signal-to-noise characteristics of silicon nitride nanopores, which is primarily a product of flicker noise in the low frequency regime and dielectric loss and capacitance at higher frequencies [86,180], can be partially mitigated by coating the membrane with a low dielectric material such as polydimethylsiloxane (PDMS) [86]. This has been shown to significantly reduce the high frequency noise to levels comparable to that of α-hemolysin (α-HL) [86]. In this way, Tabard-Cossa *et. al.* showed significantly improved performance of SiNx nanopores for DNA analysis demonstrating the ability to distinguish between DNA translocation events that differ by the conformation of the translocating molecule [86].

Other membrane materials have also served as substrates for synthetic pores. More recently, nanopores in graphene monolayers have been drilled using a tightly focused electron beam [181, 182]. Graphene nanopores, which are a single atom thick, are exciting prospects for application in high-speed nanopore DNA sequencing. Given the thickness of

the membrane, the pore current in the presence of DNA is expected to be a function of a minimal number of nucleotides occupying, or within the vicinity of the pore, thereby potentially optimizing pore spatial resolution for nanopore sequencing. Moreover, theoretical calculations of DNA translocation through graphene nanopores have indicated the possibility of single base resolution by exploiting the conductivity properties of graphene for probing the translocating strand electrically in the transverse direction [103]. While many issues remain to be resolved with respect to realizing graphene's potential in nanopore sensing (e.g. the very low signal-to-noise ratio exhibited by graphene nanopores [181,182]), the ability to form pores in graphene and the recent demonstration of DNA translocation through these pores shows promise for the material in nanopore DNA analysis.

While solid-state pores resolve many of the issues associated with organic pores, they suffer from problems that are in general resolved by their organic counterparts (e.g. the structural irreproducibility of synthetic pores). Given the inverse problems these classes of pores exhibit, several groups have recently explored the possibility of forming hybrid pores (i.e. hybrid organic and synthetic pores to ideally exploit the advantages of each class of pore while mitigating the disadvantages) [183]. In this regard it has been shown that by tagging preformed α-HL pores with double stranded DNA (dsDNA), the DNA-pore complex can be electrophoretically driven into a synthetic nanopore. In this way, the α-HL pore docks into the synthetic pore while the dsDNA subsequently dissociates leaving an α-HL pore firmly positioned within a synthetic nanopore (figure B-2) [183]. In this configuration, α-HL was shown extremely robust to both mechanical and electrical perturbations (i.e. the pore remained active at very high electric potentials and over several days [183]) while

maintaining its low noise characteristics. Given these encouraging initial results, research efforts are ongoing to improve the technique for forming hybrid pores. These results however, point to this new class of pore as a potentially optimal solution in nanopore sensing.

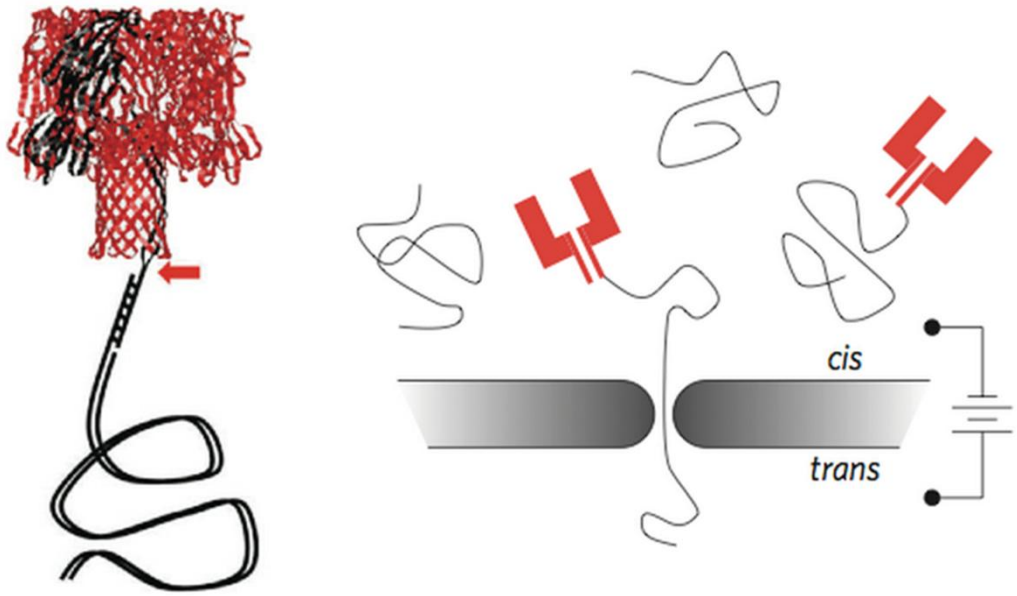


Figure B-2. Formation of hybrid nanopores. (left) A DNA-pore complex is formed by tagging a preformed α-hemolysin nanopore with dsDNA. **(right)** A DNA-pore complex in solution can be electrophoretically driven into a silicon nitride nanopore thereby forming a hybrid organic-synthetic nanopore upon dissociation of the dsDNA tag. Image source: [183] (reprinted by permission from Macmillan Publishers Ltd: Nature Nanotechnology [183], copyright 2010).

Appendix C

Forming an α-hemolysin nanopore for single molecule analysis

Nanopore analysis of single molecules can be performed by measuring the modulation in ionic current passing through the nanopore while an individual biomolecule such as DNA or RNA is resident in, translocating, or otherwise interacting with the pore. The corresponding current signature has been shown to reveal properties of the biomolecule and information on its interactions with the pore. The α-hemolysin nanopore remains the pore of choice, particularly for single molecule analysis of nucleic acids, because of its internal dimensions, hydrophilicity and low noise characteristics. In this appendix we present a detailed protocol for forming a robust α-hemolysin nanopore (or multiple nanopores) for single molecule analysis.

C.1 Introduction

Nanopore single molecule analysis is based on the electrical detection of individual molecules, electrophoretically driven into a nanometer-scale pore, by measurement of the modulation of ionic current as each biomolecule enters, resides in, and exits the pore. Of the available organic pores, α-hemolysin (α-HL), a heptameric protein that self-assembles as a pore in a lipid bilayer, is the most commonly used for studies of nucleic acids. Its dimensions

(in particular its ~1.4 nm diameter limiting aperture) [72], internal hydrophilicity, and low-noise characteristics make it ideally suited for this purpose.

Single molecule DNA analysis using nanopores was first reported in 1996 [21]. In this seminal paper, Kasianowicz *et al.* detected individual strands of RNA and DNA by the transient blockage of ionic current as they were driven through an α-hemolysin nanopore. The authors confirmed the relationship between molecule translocation in the pore and current blockage events, and showed a correlation between blockage event times and the molecule length. Nanopore current was later shown to display a high degree of sensitivity to the nature of the molecule blocking the pore, as homopolymers that differ only in sequence were distinguished from each other based on the nucleic acid – α-HL pore interactions that ensue during translocation [22].

Since then, there has been an explosion of studies and techniques developed using nanopores. These include single molecule force spectroscopy [118,184], studies of nucleic acid dynamics [18,19,160,161], progress towards single molecule sequencing (see section 2.2), single molecule chemistry [185,186], and protein analysis (see section 4.1). A force spectroscopy technique employing multiple α-HL nanopores incorporated into a single lipid bilayer has also been developed for rapid single nucleotide polymorphism detection [119].

In this appendix we present the protocol for forming single or multiple α-hemolysin nanopores. The ease with which an α-hemolysin nanopore can be formed is strongly correlated with the quality of the lipid bilayer in which it incorporates. We therefore also present considerable detail on methods for forming robust lipid bilayers.

C.2 Materials

C.2.1 Aperture construction in PTFE / FEP tubing (forming the U-tube)

1. Constantan® thermocouple wire, 0.001” diameter (OMEGA Engineering Inc.).
2. Polytetrafluoroethylene (PTFE) / Fluorinated Ethylene-Propylene (FEP) dual-shrink tube 0.036” inner diameter (ID), (Small Parts Inc.).
3. Instant adhesive (Loctite® 401).
4. 24 AWG solid wire (insulation removed).
5. Varitemp® heat gun (MASTER® Appliance).

C.2.2 Silver-chloride (Ag/AgCl) electrodes

1. Silver wire (Alfa Aeser), 1mm (0.04”) diameter, 99.9% purity.
2. 0.06” ID and 0.130” ID PTFE/FEP dual-shrink tube (Small Parts Inc.).
3. Bleach (6% sodium hypochlorite).
4. Connector pins appropriate to the patch clamp amplifier head stage (e.g. MaxBit® 1 mm diameter, male, crimp pins if using the Axopatch 200B, Molecular Devices Corp.). Refer to **Subheading C.2.5** regarding the patch clamp amplifier.
5. 24 AWG solid wire.
6. Varitemp® heat gun (MASTER® Appliance).
7. Imperial® wetordry® sanding paper (3M Inc.).

C.2.3 Black lipid bilayer formation

1. KCl buffer: 1 M potassium chloride (KCl), 10 mM HEPES, and 1 mM EDTA, pH 8.0. Filter through a 0.1 µm vacuum filter (e.g. Stericup® filter units, Millipore Corporation.). Adjust the pH using 2 M potassium hydroxide (KOH). Store at room temperature.
2. 10 mg/mL solution of 1,2-diphytanoyl-*sn*-glycero-3-phosphocholine, powdered form, (Avanti Polar Lipids) in chloroform. Store lyophilized lipid and lipid chloroform solution at -20°C. (*see* **C.4 Note 1**).
3. Single bristle lipid brush **(figure C-1)** made from a 10 µL pipette tip, a paint brush bristle, and a small stick from a cotton swab. (*see* **C.4 Note 2**).
4. Varitemp® heat gun (MASTER® Appliance).
5. Three NORM-JECT® syringe connectors **(figure C-2)**. Using the heat gun, shrink 0.130" ID PTFE / FEP dual shrink tube (Small Parts Inc.) around stainless steel hypodermic needles (with plastic hubs), 14, and 15 gauge. (Small Parts Inc.). (*see* **C.4 Note 3**).
6. 1-Hexadecene, (Sigma-Aldrich). Store at room temperature.
7. Concentrated nitric acid (HNO_3).
8. Hexane. Store at room temperature.
9. Ethanol. Store at room temperature.
10. 50 µL glass capillary pipettes (e.g. Drummond® Wiretrol micropipettes, VWR International).
11. Oilless vacuum pump (24" Hg, McMaster-Carr®) and vacuum chamber (VWR).

12. Stirring hot plate (Corning®).
13. Polyvinyl chloride (PVC) gloves (Winter Monkey Grip®).
14. 6 M Sodium Hydroxide (NaOH) solution.
15. Safety glasses.

Figure C-1. Lipid brush. A single paint brush bristle formed into a loop extends out the end of the pipette tip.

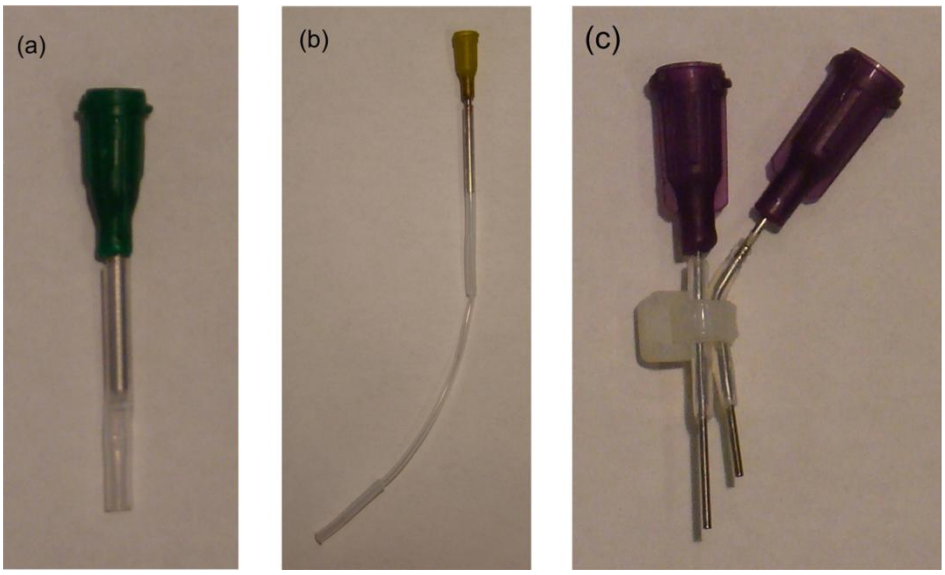


Figure C-2. The syringe connectors. (a) Connector for flushing solutions through the U-tube (step 4 of section C.3.3.). **(b)** Two of this type of connector should be made. One used for pushing air through the U-tube and one for pushing KCl buffer (refer to section C.3.3.). **(c)** Tygon®-syringe connector for flushing out and replacing KCl buffer (step 6 of section C.3.4).

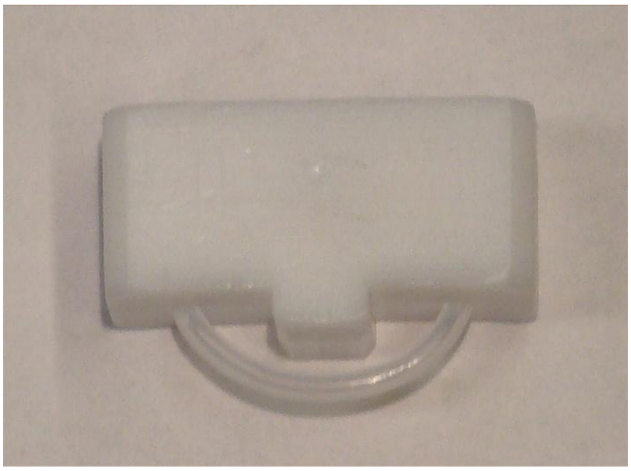
C.2.4 α-hemolysin pore (or multiple pore) formation

1. α-Hemolysin from *Staphylococcus aureus* (Sigma-Aldrich). Follow MSDS precautions when handling α-hemolysin, which is highly toxic. Mix lyophilized α-hemolysin powder with a 1:1 glycerol to water solution at 1mg/mL. Store powdered α-hemolysin and solution at -20°C. The solution should be kept for a maximum of 3 months.
2. Venturi pump with a 0.25” ID Tygon® hose attached.
3. Tygon®-syringe connector **(figure C-2c)**. The connector is made by bending the steel of a 21 gauge hypodermic needle by approximately 30° and zip tying it to a second 21 gauge hypodermic needle (the bottom of the bent hypodermic needle should not be flush with that of the straight needle). It is helpful to heat shrink some PTFE/FEP tubing around a portion of each needle to improve the tightness of the tying.

C.2.5 Experimental apparatus

1. PTFE cell **(figure C-3)**.
2. Aluminum cell housing **(figure C-4)** to secure the PTFE cell, Ag/AgCl electrodes, and light source. The housing should be grounded to reduce electrical noise.
3. Vibration isolation table for mechanical noise reduction. Light load air suspension tables are adequate (e.g. LW series tables, Newport Corporation).
4. A microscope light source. (e.g. a high intensity white LED & DC battery or a fiber-Lite high intensity illuminator, model 170D, with flexible fiber optic cable, Dolan-Jenner Industries Inc.).
5. 40x magnification, stereomicroscope with swing arm stand (e.g. Leica S6E, Leica Microsystems).
6. 100 MHz bandwidth, 1.25 GS/s (giga-samples per second), 2 channel oscilloscope (e.g. TDS 3012B, Tektronix).
7. At least one faraday cage over the cell housing to eliminate external electromagnetic fields.
8. Thermoelectric module (peltier) and temperature controller (TE Technology Inc.) for performing temperature dependence studies. (*see* **C.4 Note 4**).
9. Low noise, patch clamp amplifier (e.g. Axopatch 200B – Molecular Devices). The signal from the patch clamp amplifier may need to be grounded, which can be accomplished by connecting the signal ground to the vibration isolation table.

Appendix C. Forming an α-HL nanopore for single molecule analysis



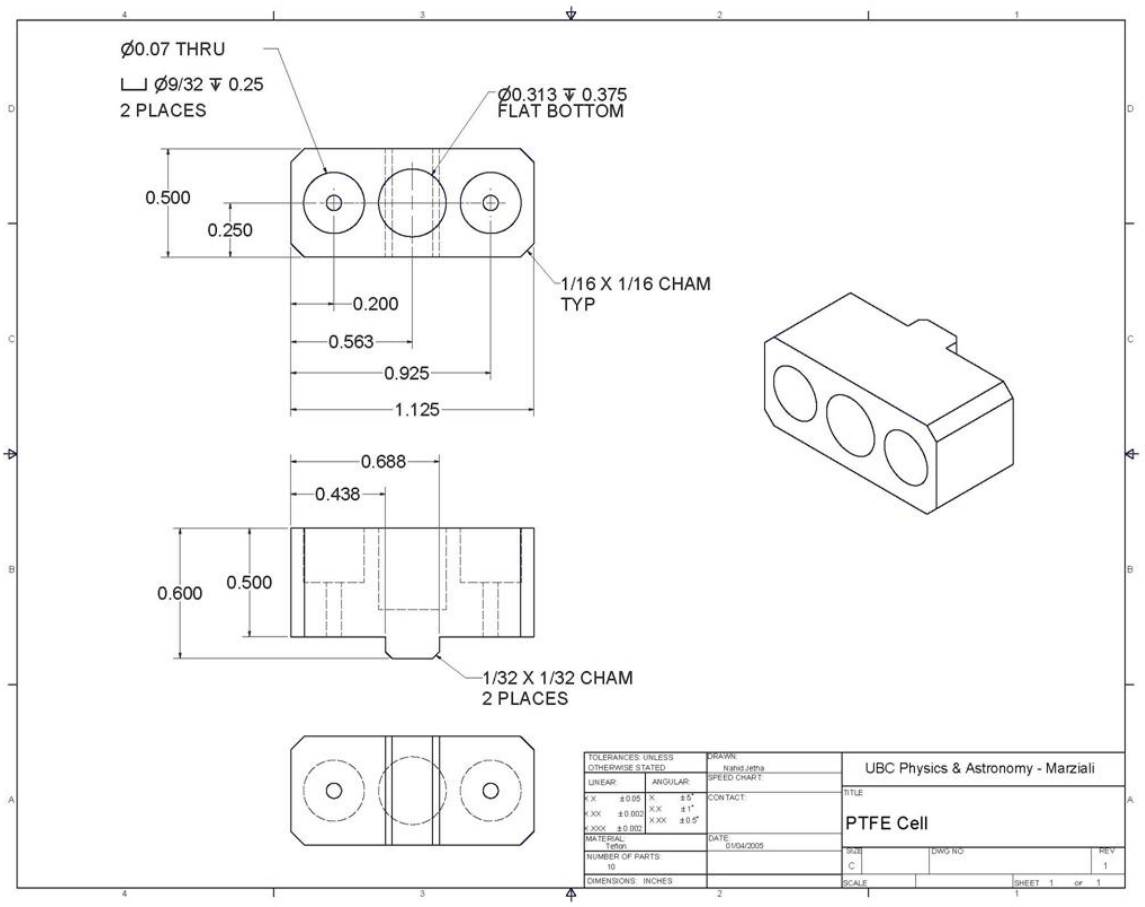


Figure C-3. Cell. (top) PTFE cell. **(middle)** Front view of the PTFE cell with the U-tube in place. **(bottom)** Mechanical drawing of the PTFE cell.

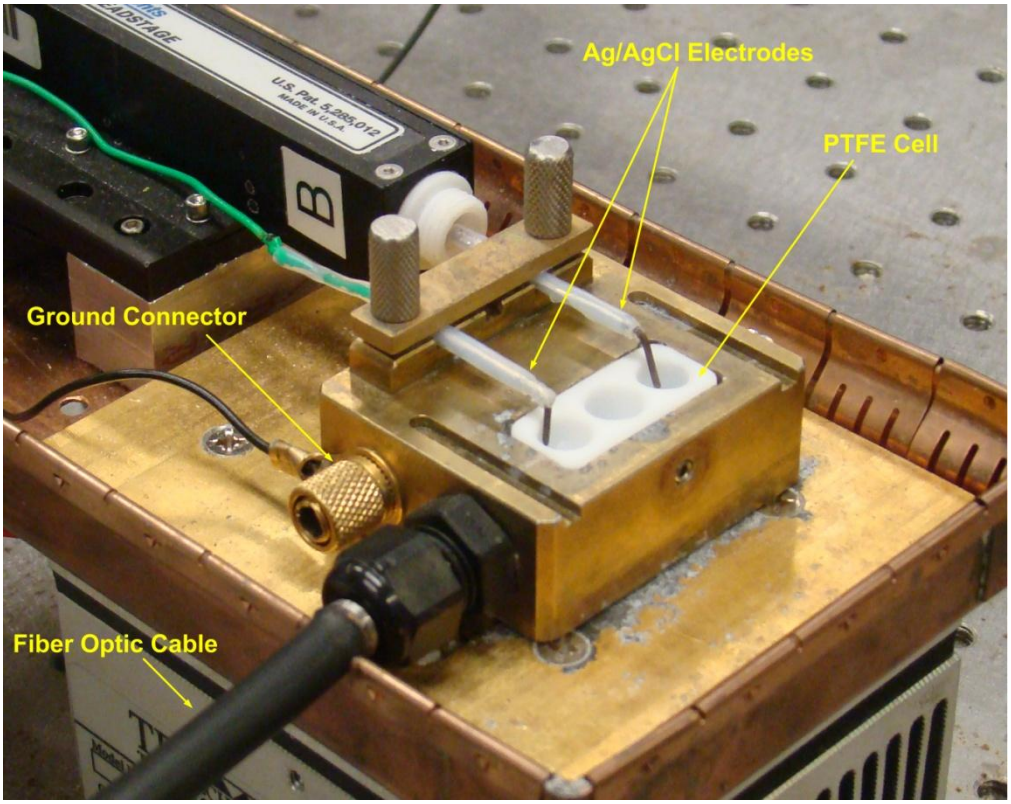


Figure C-4. The experimental apparatus with components in place. The aluminum cell housing has been gold plated to remove the oxide layer and improve the attenuation of electromagnetic noise. A first faraday cage is placed overtop of the electrodes and PTFE cell (positioned by the grooves in the aluminum cell housing). A second faraday cage encloses the entire system.

C.3 Methods

C.3.1 Aperture construction in PTFE / FEP tubing (forming the U-tube)

1. Wrap a piece of the thermocouple wire around the solid wire several times and glue using instant adhesive. 1-2” of thermocouple wire should extend past the end of the solid wire.

2. Position the solid wire inside a 6” long piece of 0.036” ID PTFE/FEP tubing such that only a small portion of the thermocouple wire protrudes from the end.
3. Using the heat gun, shrink ~0.25” of tubing around the thermocouple wire until it closes completely.
4. Use a new razor blade to cut the tubing near the end of the closed portion, preserving ~1/16” of the closed portion on the end of the tubing (to ensure that the aperture is 0.001in in diameter).
5. Gently pull the solid wire from the back end of the tubing. Be careful not to break the thermocouple wire in the tubing. To check that the aperture has been created properly, push water through the back end of the tubing using a NORM-JECT® syringe (with the appropriate syringe connector, **figure C-2a**). A very fine stream of water will spray from a good aperture.
6. Insert the tubing into the cell and bend it into the U-shaped formation as shown in figure C-3. (*see* **C.4 Note 5**).

C.3.2 Silver-chloride (Ag/AgCl) electrode fabrication

1. Insulate a 1” long section of 1mm diameter silver wire by heat shrinking 0.06” ID PTFE/FEP using the heat gun. Leave approximately 0.25” of silver overhanging on each end (the silver wire should be approximately 1.50” long). Repeat this step using the 0.130” ID PTFE/FEP tubing.
2. For the power electrode **(figure C-5a)**, solder one end of the silver wire to a crimp pin, then heat shrink a piece of PTFE/FEP tubing around the soldered portion of the pin.

3. For the ground electrode **(figure C-5b)**, solder an approximately 8” long piece of solid wire to one end of the silver wire, then heat shrink a piece of PTFE/FEP tubing around the soldered portion of the wire. Solder the other end of the solid wire to a crimp pin and heat shrink a piece of PTFE/FEP tubing around the soldered portion of the pin.
4. Bend the free end of the silver wire by 90° (for both electrodes) and immerse the electrodes into bleach solution for ~30min to create a layer of AgCl on the electrode surface. This step should be repeated before every experiment.

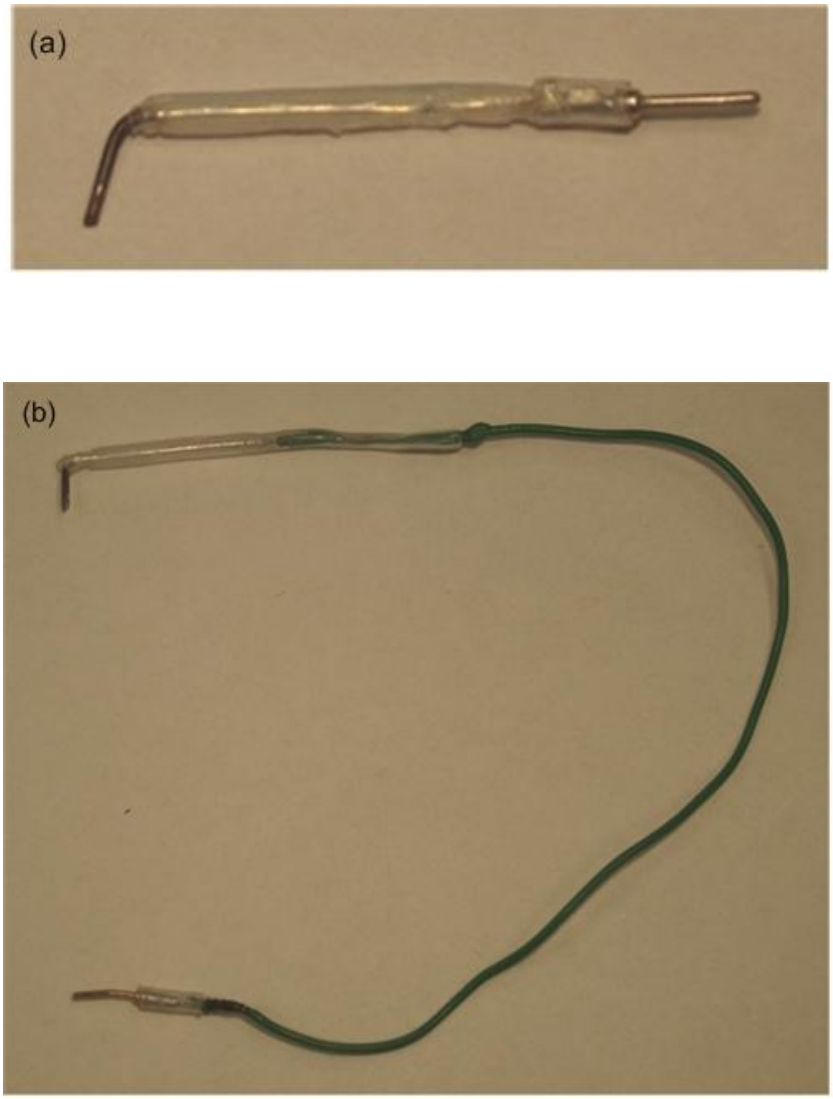


Figure C-5. Electrodes. (a) The power electrode positioned on the *trans* side (the non-aperture side) of the PTFE cell. **(b)** The ground electrode positioned on the *cis* side (aperture side).

C.3.3 Forming a black lipid bilayer

1. The day before forming a nanopore, set some lipid out to dry in test tubes. Clean three test tubes by blowing pressurized air in them to remove any dust that may have settled on the bottom of the tube. Transfer 50 µL of 10 mg/mL 1,2-diphytanoyl-*sn*-glycero-3-phosphocholine / chloroform solution into each test tube (using a glass capillary pipette). Let the lipid dry overnight in a fume hood. (*see* **C.4 Note 6**).

2. The day of the experiment, turn on the patch clamp amplifier. Allow it to warm up for 1 hour before beginning to record data.
3. Perform this step in a fume hood; use PVC gloves, a lab coat, and safety glasses. Immerse a PTFE cell in a solution of ~15% nitric acid (v/v) in dH_20. Bring the solution to a boil and then dump it into a basin (prefilled with ~0.5 L of water; neutralize with 6 M NaOH). Repeat this boiling procedure two more times with deionized water only.
4. Sequentially flush the U-tube of the cell with ~200 µL of each of the following solutions: dH_2O, ethanol and hexane. To flush the tubing load a NORM-JECT® syringe (fitted with the appropriate connector, **figure C-2a**) with the corresponding solution and push the solution through the tubing from the *trans* side (i.e. the non-aperture end). A thin stream of liquid should spray from the aperture. To prevent introducing air bubbles, remove any air trapped at the forefront of the syringe prior to connecting it. (*see* **C.4 Note 7**).
5. Push any remaining solution out of the tube with air (using a dry NORM-JECT® syringe connector) and place the cell in the vacuum chamber. Pump vacuum for approximately 5 minutes, or until the remaining solution inside the U-tube has evaporated. (*see* **C.4 Note 8**).
6. Using a glass pipette, aliquot 250 µL of hexane into one of the test tubes of dried lipid to dissolve the lipid. Cap the tube to prevent evaporation.
7. Remove the silver-chloride electrodes from the bleach solution and rinse them with deionized water.
8. Clean the lipid brush with an ethanol soaked KimWipe®. Allow the brush to dry.

9. Securely fasten the PTFE cell into the aluminum cell housing **(figure C-4)**. With the aid of the stereomicroscope and illuminator, pipette 1.5 µL of the hexane-lipid solution (from step 6) onto the aperture of the U-tube. This coats the surface and inner walls of the tube with a lipid monolayer. Repeat this once more after the hexane has evaporated.
10. Using a NORM-JECT® syringe & syringe connector **(figure C-2b)**, push air through the U-tube from the *trans* side to remove any lipid solution inside the tube. Note the location of efflux of the hexane-lipid solution as this is the location of the aperture. Once the tube is visually dry place the aluminum cell housing (with the PTFE cell securely fastened) into the vacuum chamber and pump vacuum for approximately five minutes. (*see* **C.4 Note 8**).
11. Wear safety glasses for this step. Aliquot 5 µL - 6 µL of 1-hexadecene to another test tube of dried lipid. Allow the 1-hexadecene to coat and soften the lipid for approximately 20 – 40 seconds. Invert the test tube; gently and repeatedly tap the base of the tube to remove the excess 1-hexadecene from the lipid. Be careful not to shatter the test tube. Store the tube in the inverted position, to prevent the 1-hexadecene from collecting at the bottom (*see* **C.4 Note 9**).
12. Install the electrodes **(figure C-4)** and connect them to the head stage of the patch clamp amplifier (The *cis* side – i.e. the aperture side - is ground). Connect the aluminum cell housing to ground (through a connection to the vibration isolation table). Set the output gain on the patch clamp amplifier to one. Ensure that the applied voltage is zero, the pipette capacitance compensation is minimized, and that the whole cell capacitance compensation and the series resistance compensation are off (these are the parameters

for the Axopatch 200B). The patch clamp amplifier should be set to low pass filter at 10kHz (if performing nanopore force spectroscopy, refer to [119])

13. Using a NORM-JECT® syringe and syringe connector **(figure C-2b)**, gently push the KCl buffer or analyte solution (e.g. DNA solution in 1 M KCl, if performing nanopore force spectroscopy, refer to [119]) through the U-tube from the *trans* side (be wary of air bubbles and ensure that none are introduced into the tubing). (*see* **C.4 Note 7**).
14. Fill the *cis* and *trans* sides of the PTFE cell with KCl buffer. Fill both the wells such that the electrodes from each particular side are fully immersed in the solvent.
15. Apply a 5 mV square wave potential using the patch clamp amplifier (Seal Test toggle of the Axopatch 200B). Current through an open tube will appear as a square wave; if the tube contains an obstruction (e.g. an air bubble), there will be a zero DC bias with spikes during potential steps, due to the capacitance of the obstruction. (*see* **C.4 Note 7**).
16. Adjust the offset potential to zero the current (Pipette Offset knob of the Axopatch 200B). Adjust the scale (and the trigger) of the oscilloscope to view a single cycle of the seal test. Monitor the current for a few minutes; significant, continual current drift indicates cell leaks or depletion of the AgCl layer on the electrodes. (*see* **C.4 Note 10**).
17. Using the lipid brush transfer a small amount of lipid from the lipid/1-hexadecene test tube, to the *cis* side of the U-tube (i.e. the end containing the aperture). When collecting the lipid keep the test tube in the inverted position at all times. When transferring the lipid, transfer it onto the edge of the U-tube (away from the aperture). After transferring the lipid wipe the bristle with a KimWipe®. (*see* **C.4 Note 11**).
18. Using the lipid brush, mix the lipid for approximately 20–30 seconds until it becomes noticeably less viscous. Move the lipid to the edge of the U-tube, and clean excess lipid

from the aperture. It is important not to excessively mix the lipid as it will become too soft for practical use. (*see* **C.4 Note 12**).

19. Using a NORM-JECT® syringe and syringe connector, push a small amount of KCl buffer (or analyte solution, if performing nanopore force spectroscopy, refer to [119]) through the U-tube from the *trans* side to clear the inside of lipid. If it is cleared then the capacitive spike should disappear. (*see* **C.4 Note 13**).
20. Form a lipid bilayer. Using a 20 µL micropipettor, blow an air bubble across the aperture and suck it back up. If a bilayer is formed a capacitive spike will appear on the oscilloscope.
21. Check for leaks by applying a 100 mV potential across the bilayer. Ionic current of <1pA indicates a good bilayer. Current higher than this indicates leaky bilayers, or leaks in the cell itself. (*see* **C.4 Note 14**).
22. Check the mechanical properties of the bilayer by applying an increasing potential until the bilayer breaks, causing the current to instantaneously saturate. A good bilayer will burst at a voltage of 450 mV to 550 mV. Bilayers bursting at <300mV are generally unstable and should not be used (*see* **C.4 Note 15**). Membranes bursting at >600 mV occur when a thick lipid blockage occludes the tube; such a blockage is not a bilayer and cannot be used to form a nanopore (*see* **C.4 Note 16**). Reform a bilayer and repeat this step a couple of more times to ensure that the bilayers consistently burst between 450 mV and 550 mV.

C.3.4 α-hemolysin pore (or multiple pore) formation

1. Prepare a NORM-JECT® syringe by filling it with KCl buffer.

2. For forming a single nanopore, prepare a diluted solution of α-toxin by transferring 3 µL of the α-hemolysin solution (the α-hemolysin dissolved in glycerol/water) with 50 µL of KCl buffer. For forming multiple nanopores mix 3 µL of the α-hemolysin solution with 20 µL of KCl.
3. Form a lipid bilayer (refer to step 20 of **Subheading C.3.3**) and add 3 µL of the diluted α-toxin solution to the *cis* side of the PTFE cell (the side containing the aperture of the U-tube). Add the α-toxin away from the aperture, and mix it into the KCl buffer by refluxing with the micropipette. (*see* **C.4 Note 17**).
4. Set the potential to 100 mV and wait for channel insertion. An α-hemolysin nanopore will conduct a current of 90 pA to 100 pA at 100 mV and a current of -70 pA to -79 pA at -100 mV (open channel current values for a single pore in 1 M KCl). Check both polarities of the potential before proceeding (it is possible for a secondary structure of the protein to incorporate into the lipid bilayer in which case the conducted current will reside outside of the above ranges. In this case burst the bilayer and reform it).
5. For forming multiple nanopores wait until 50 to 150 nanopores incorporate into the lipid bilayer before proceeding (the conducted current should be approximately 50 to 150 times the single channel values, respectively).
6. Flush out and replace the KCl buffer on the *cis* side of the PTFE cell. Attach the syringe from step 1 to the longer needle of the Tygon®-syringe connector; squirt some liquid out to remove any air bubbles. Connect the other needle to the hose of the venturi pump, and turn it on. Insert the longer needle into the *cis* well and inject the clean KCl solution while simultaneously removing excess liquid through the shorter needle connected to the venturi pump. Make sure that the liquid level on the *cis* side of the PTFE cell does not

drop below the top surface of the U-tube otherwise the pore(s) will disappear. Perform this step quickly to prevent further pores from incorporating into the bilayer. Repeat this step three times. (*see* **C.4 Note 18**).

7. Turn off the 100 mV applied potential. Zero the current if required. (*see* **C.4 Note 19**)

C.4 Notes

1. Avoid transferring chloroform (and hexane) with plastic, which can contaminate the solutions and affect the quality of the lipid. Wherever possible, use glass pipettes or capillary pipettes.
2. To make the brush, form a U-shaped loop with a single bristle from a Sable artist's paint brush, and insert both free ends of the bristle into a 10 µL pipette tip leaving a loop ~2 mm long. Glue the loop in place using instant adhesive.
3. 0.06" ID PTFE/FEP dual shrink tube may also be used depending on the outer diameter of the hypodermic needles. When heat shrinking the tubing the goal is to shrink the suspended portion such that when placed around the native 0.036" ID PTFE/FEP dual shrink tube it forms a snug fit.
4. Because the experiments are sensitive to electrical and mechanical noise, the peltier device should be liquid cooled as opposed to air cooled by a fan. Liquid hard-drive coolers that pump liquid through a water block (to which the peltier should be fastened) is a good non-invasive cooling option.
5. It can be difficult to insert the tubing into the through hole of the PTFE cell because the fit is tight. Heat shrink ~1 ½" at the non-aperture end of the PTFE/FEP tubing and insert that end ("guide end") into the through hole. Use pliers to pull on the guide end to pull

the rest of the tubing through. Once the tube is in place, cut the excess tube from the non-aperture end.

6. Lipid takes at least 4-5 hours to dry once transferred. Lipid oxidizes over time; do not keep dried lipid for more than 2 days.
7. If air bubbles do get trapped inside the tube, and are obstructing the flow of liquid, push solution through the *cis* side of the tubing (i.e. the aperture end). It is crucial if doing this, to remove any air trapped at the forefront of the syringe. After clearing the obstruction, return to pushing solution through the *trans* side. If the flow remains impeded after numerous attempts, restart with a different PTFE cell.
8. The syringe connector used to push air through the U-tube should always remain dry (i.e. for pushing solution through the U-tube use a different connector).
9. Extending this time beyond 40 seconds will likely result in over-softening of the lipid.
10. Depleted electrodes are white or have white patches present, in which case they need to be replated in bleach. If this is not the case, then gently disassemble the PTFE cell (without emptying it) and check for leaks or spills; wipe any KCl solution from the cell or apparatus and reassemble. In general the current will drift slightly over time, in which case it should be re-zeroed. The Offset current should be re-zeroed periodically, e.g. every 10 minutes during experiments.
11. Avoid contact between the harvested lipid and any 1-hexadecene residing on the walls of the test tube. Excess 1-hexadecene may over-soften the lipid.
12. When moving the lipid back to the edge of the U-tube, leave a very small amount of lipid near the aperture to act as a reservoir for bilayer formation.

13. If the lipid clears from the aperture but spontaneously re-enters than it may imply that the lipid is too soft. Spontaneous clogging of the aperture may be remedied by pushing solution through the tubing from the *trans* side with one hand and using the lipid brush with the other to push lipid away from the aperture. Refer to **C.4 Note 15** for a possible remedy to the excess softening of the lipid.
14. If a substantial leakage current is observed, re-mix the lipid (refer to step 18 of **Subheading C.3.3**). If this does not solve the problem, then check for leaks or spills (*see* **C.4 Note 10**).
15. If the bilayer bursts at potentials less than 300 mV it can imply one of two things. Either the lipid has not been mixed well enough, in which case re-mix the lipid (refer to step 18 of **Subheading C.3.3**). The alternative is that the lipid is too soft (which may be a product of over-mixing or too much 1-hexadecene). If this is the case than transfer another fraction of lipid (step 17 from **Subheading C.3.3**, alternatively aliquot 1-hexadecene into a third test tube of dry lipid and transfer this lipid to the *cis* side of the U-tube), which can be used as a mop to collect the soft lipid that surrounds the aperture replacing it with fresh, unmixed lipid. This new lipid need not be mixed thoroughly as required in step 18 of **Subheading C.3.3**.
16. Clear the U-tube of lipid (refer to step 19 of **Subheading C.3.3**).
17. Do not add the α-toxin too close to the aperture, to prevent runaway nanopore formation. If this occurs then flush out and replace the KCl buffer on the *cis* side of the cell (refer to step 6 of **Subheading C.3.4**). If, after 10 minutes, no nanopores have formed then add another 2 μL of the diluted α-toxin solution to the *cis* side of the cell and mix appropriately. Burst and reform the bilayer.

18. It is a good idea to have the Tygon®-syringe connector inserted into the Tygon® tubing of the venturi pump and attached to the NORM-JECT® syringe prior to reaching this step (step 6 of **Subheading C.3.4**). Once a pore incorporates in the bilayer (e.g., in the case when making a single nanopore) there is limited time to flush out and replace the KCl solution before a second pore forms. It is important to avoid spilling any solution around the cell, as this creates leaks from the *cis* to the *trans* chamber.
19. When cleaning up (at the end of performing any experiments), empty the *cis* and *trans* contents of the PTFE cell into the basin (step 3, **Subheading C.3.3**) and place the cell into a separate container than the clean PTFE cells (cleaned in the manner described in step 3, **Subheading C.3.3**). Clean the electrodes by lightly sanding the surface with the wetordry® sanding paper, and then re-immerse them in bleach.

Appendix D

PrP^C expression and purification

The gene for ShPrP(120-232) fragment was purchased from DNA2.0. All reagents were purchased from Sigma and Fisher except for the nickel-NTA resin which was purchased from Qiagen. Mutagenesis was performed using standard protocol from QuikChange Site-Directed Mutagenesis Kit (Stratagene).

All genes (ShPrP(120-232), KKRR-ShPrP(120-232), and KKRR-ShPrP(120-232)-D178N) were inserted into pET15b(+) vector between *XhoI* and *EcoRI* restriction sites. Expression prion constructs were verified by DNA sequencing. Recombinant proteins were expressed in *Eschericia coli BL21(DE3)* host cells in LB medium supplemented with 100 μg/ml ampicillin. The cultures were grown at 37^{o}C until an OD_{600} = 0.8 was reached and then induced with 1.0 mM IPTG for 18 hr at 37^{o}C. After post-induction incubation the cells were harvested by centrifugation for 20 min at 9060 x g (4^{o}C). All prion proteins were isolated from the inclusion bodies and purified via refolding on the Ni-NTA column. The cell pellets from 1L of the LB medium with expressed prion protein was re-suspended in 25ml of the lysis buffer (50mM Tris, 100mM NaCl, 5mM EDTA, 0.1% NaN_3, pH 8.0) and lysed by sonication (using Fisher Scientific Ultrasonic Dismembrator Model 500 with microtip probe for 6 X15sec pulses at 40% power) on ice with addition of PMSF and Triton X-100 to final concentrations of 0.1mM and 0.5% respectively. The mixture was incubated on ice for 15min followed by centrifugation at 17400 x g for 20 min (4^{o}C). The supernatant was removed and inclusion bodies were cleaned by two washing steps with lysis buffer containing 0.5% Triton

X-100. Triton X-100 was removed by washing the pellet twice with lysis buffer only. Inclusion bodies were centrifuged at 17400 x g for 20 min (4^oC) after every wash and supernatant was removed. Clean protein pellet was resuspended in 30ml of the solubilization buffer (10mM Tris, 100mM sodium phosphate, 10mM reduced glutathione, 8M Urea, pH 8.0) and left for overnight incubation at room temperature with next centrifugation at 48400 x g for 1hr at 25^oC.

The Ni-NTA resin (40ml bed volume) was pre-equilibrated with solubilization buffer and mixed with clarified supernatant. The mixture was incubated (with slow rotation) at room temperature for 10 min and packed into the column. Unbound proteins were removed by column wash with 5 bed volumes of the solubilization buffer. Prion protein was refolded on a column by decreasing Urea concentration from 8M to 0M. Linear gradient of solubilization buffer and refolding buffer (10mM Tris, 100mM sodium phosphate, pH 8.0) was applied to the column. Non-specific bound proteins were washed out of the column with 5 bed volumes of the washing buffer (10mM Tris, 100mM sodium phosphate, 50mM Imidazole, pH 8.0). Prion protein was eluted from the column with elution buffer (10mM Tris, 100mM sodium phosphate, 500mM Imidazole, pH 5.8). Fractions containing protein were identified by SDS-PAGE gel and mixed. His-tag was removed from the prion protein using Enterokinase Cleavage Capture Kit (Novagen). Purified prion proteins were then concentrated using an Amicon Ultra-15 centrifugal device.

Protein concentration was determined using an extinction coefficient at 280nm (ShPrP(120-232) extinction coefficient is $20525 M^{-1}cm^{-1}$, KKRR-ShPrP(120-232) extinction coefficient is $20525 M^{-1}cm^{-1}$, and KKRR-ShPrP(120-232)-D178N extinction coefficient is

20525$M^{-1}cm^{-1}$). Extinction coefficients were calculated by Protparam program based on every specific recombinant prion amino acid sequence.

Appendix E

Sample PrP^C capture events

KKRR-ShPrP(120-232)

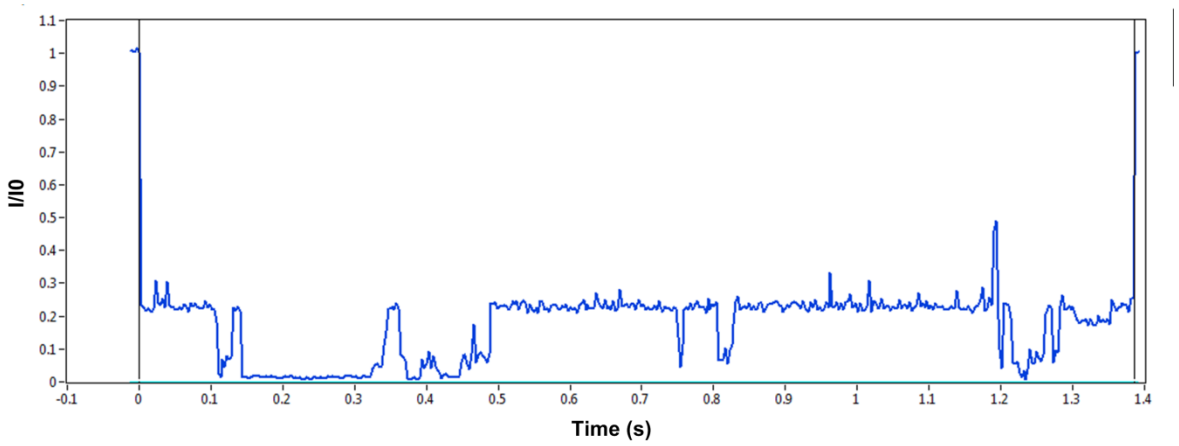


Figure E-1. Sample wild-type PrP^C capture event 1. Current through the pore upon capture and trapping of wild-type PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

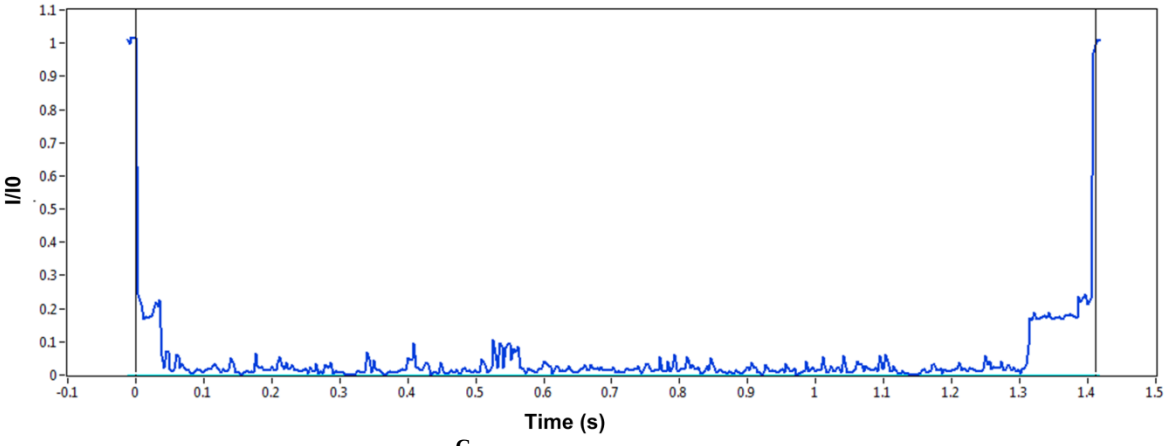


Figure E-2. Sample wild-type PrP^C capture event 2. Current through the pore upon capture and trapping of wild-type PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

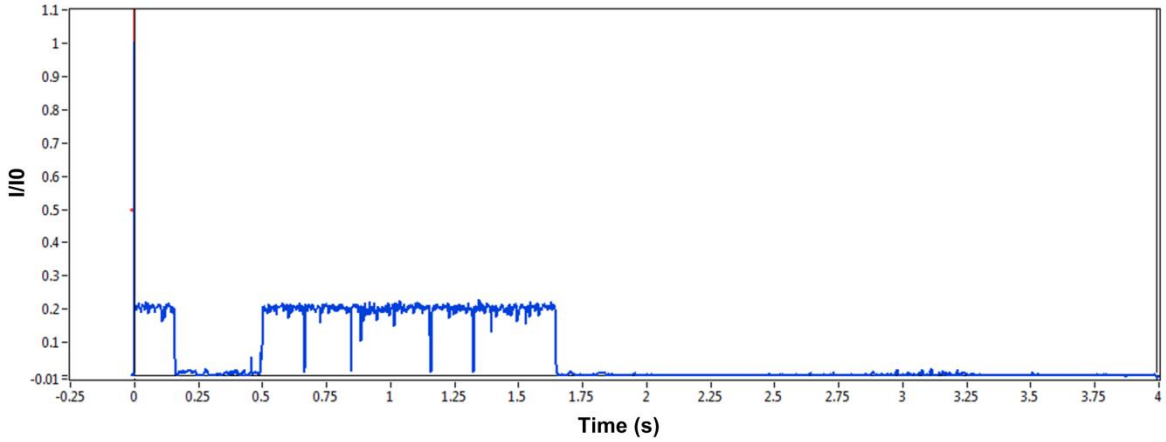


Figure E-3. Sample wild-type PrP^C capture event 3. Current through the pore upon capture and trapping of wild-type PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

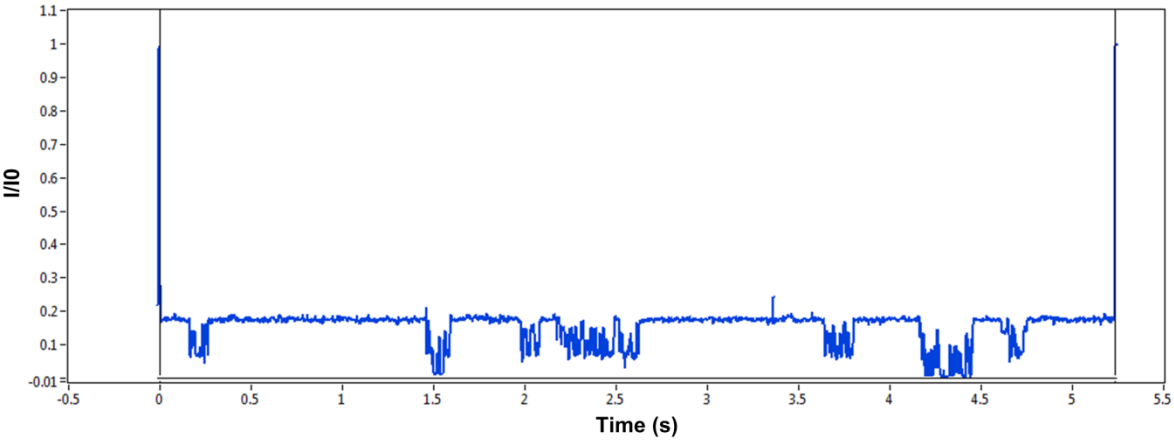


Figure E-4. Sample wild-type PrP^C capture event 4. Current through the pore upon capture and trapping of wild-type PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

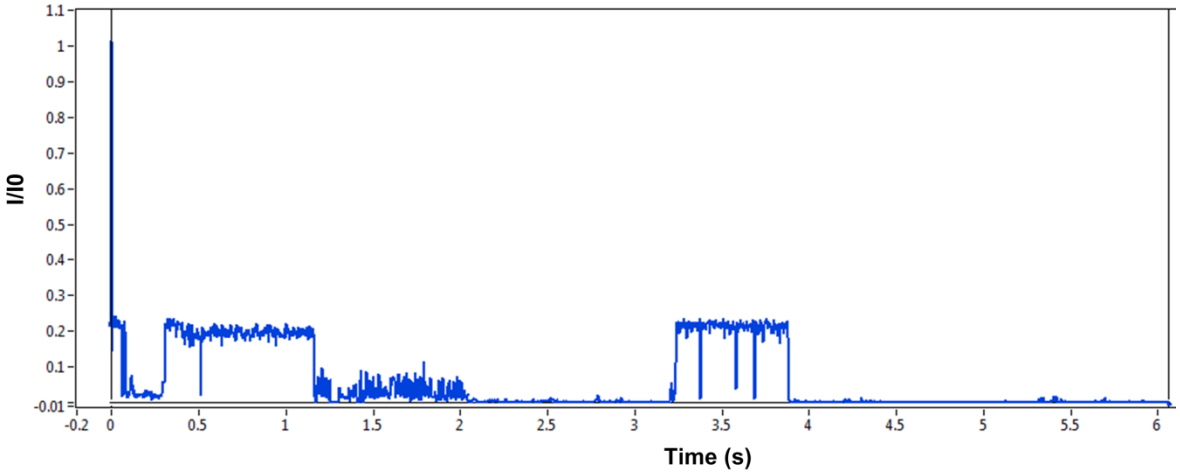


Figure E-5. Sample wild-type PrP^C capture event 5. Current through the pore upon capture and trapping of wild-type PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

KKRR-ShPrP(120-232)-D178N

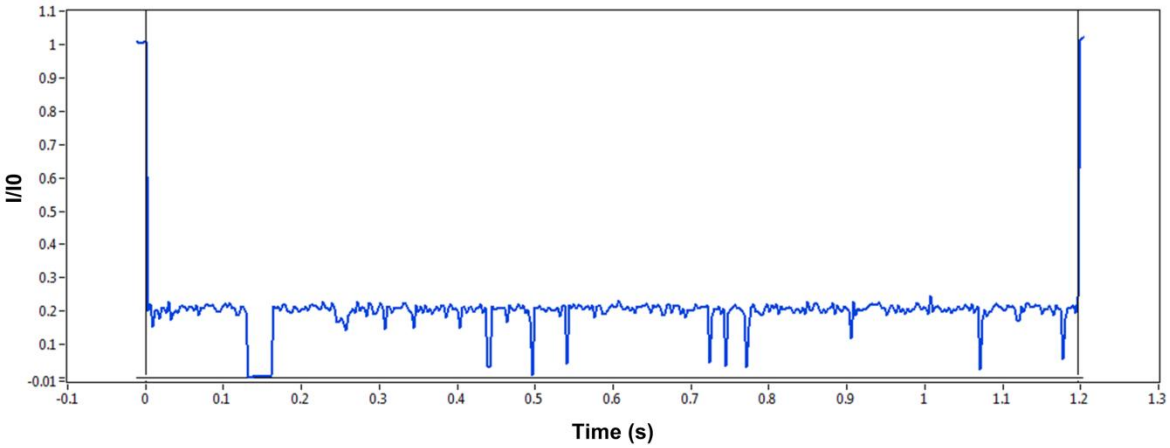


Figure E-6. Sample D178N mutant PrP^C capture event 1. Current through the pore upon capture and trapping of mutant PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

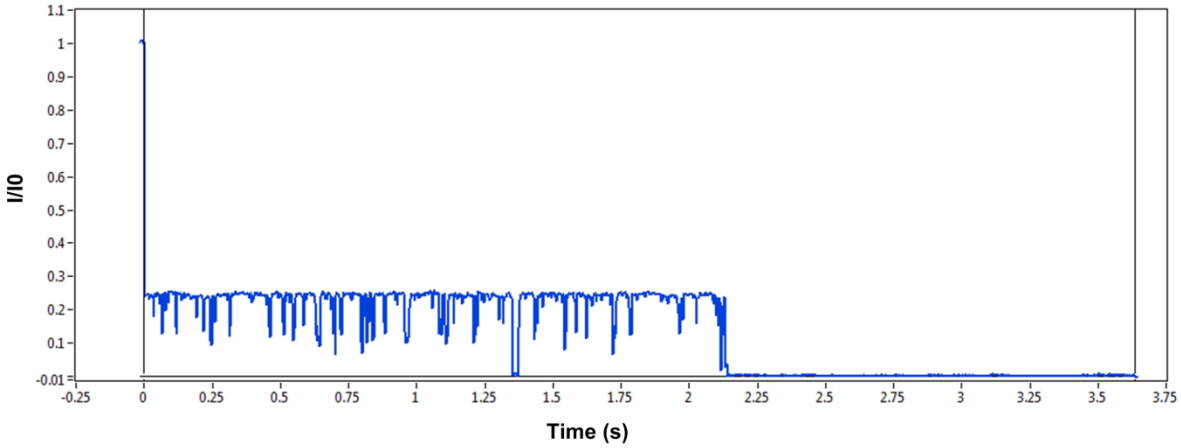


Figure E-7. Sample D178N mutant PrP^C capture event 2. Current through the pore upon capture and trapping of mutant PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

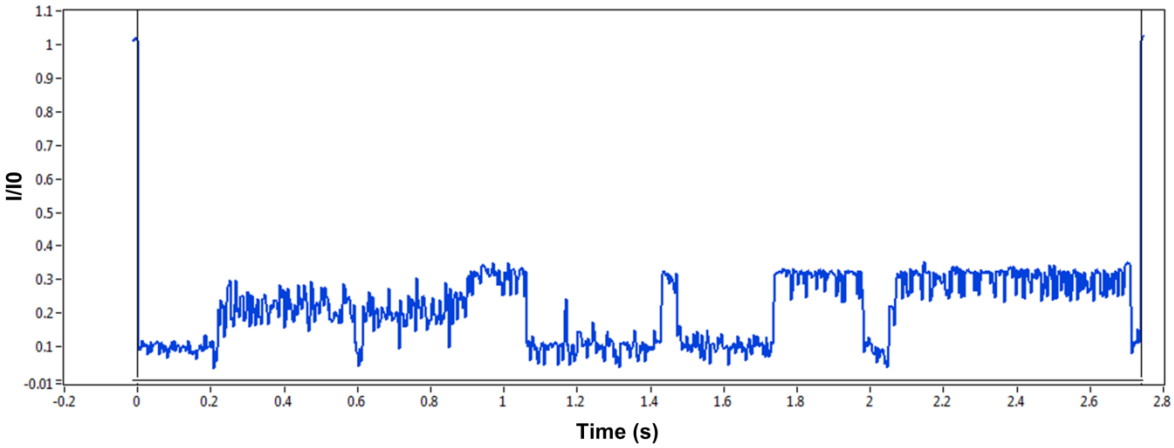


Figure E-8. Sample D178N mutant PrP^C capture event 3. Current through the pore upon capture and trapping of mutant PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

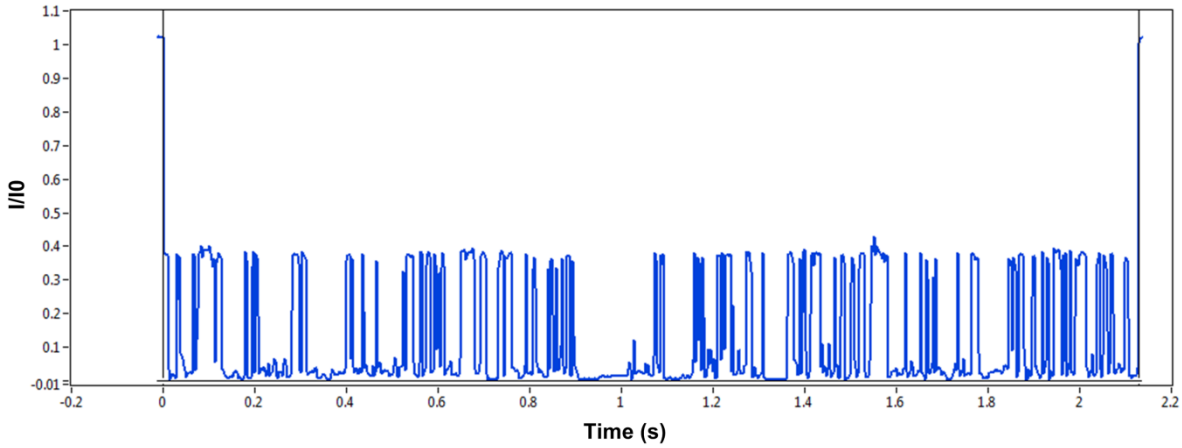


Figure E-9. Sample D178N mutant PrP^C capture event 4. Current through the pore upon capture and trapping of mutant PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

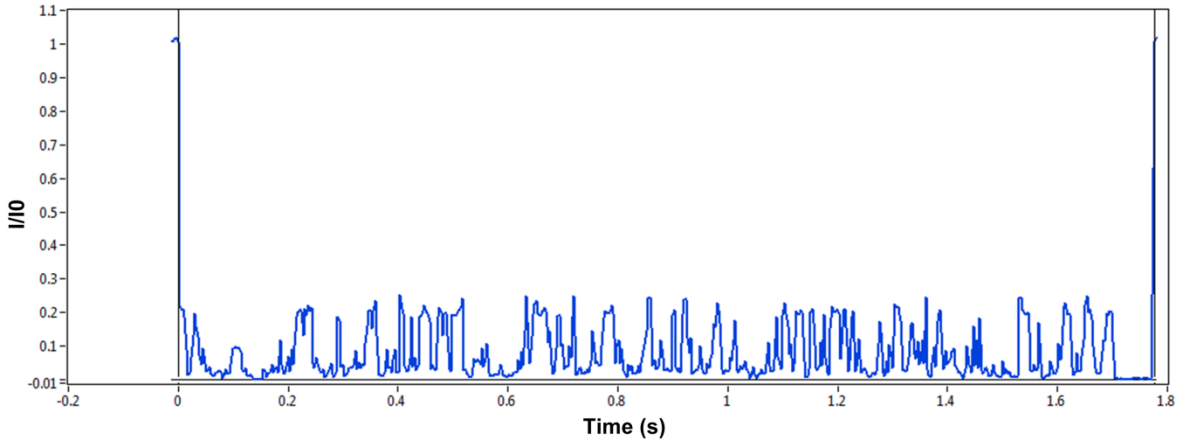


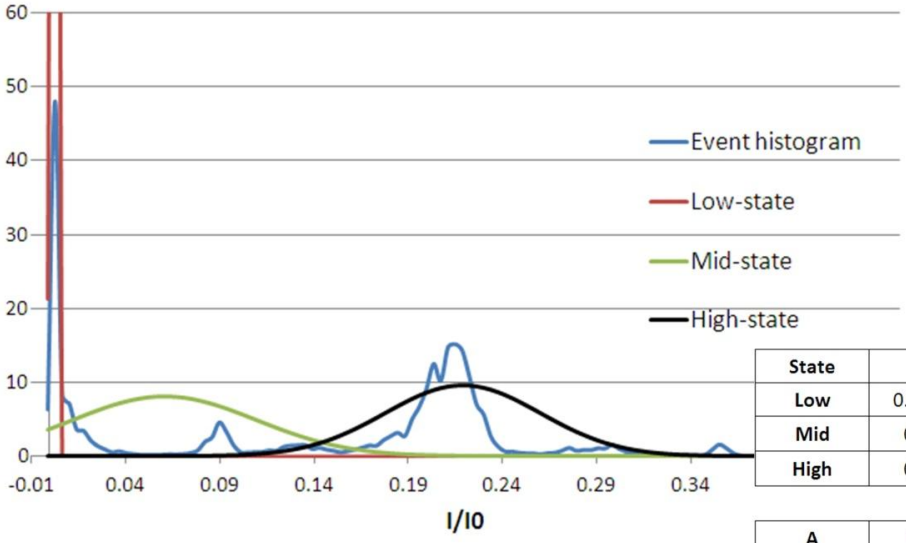
Figure E-10. Sample D178N mutant PrP^C capture event 5. Current through the pore upon capture and trapping of mutant PrP^C. Current is expressed as I/I0 where I0 is the open pore current.

Appendix F

PrP^C protein calling: Case 1 and case 2 wild-type and mutant HMM models

Protein calling case 1: Optimal KKRR-ShPrP(120-232) and KKRR-ShPrP(120-232)-D178N models

KKRR-ShPrP(120-232)

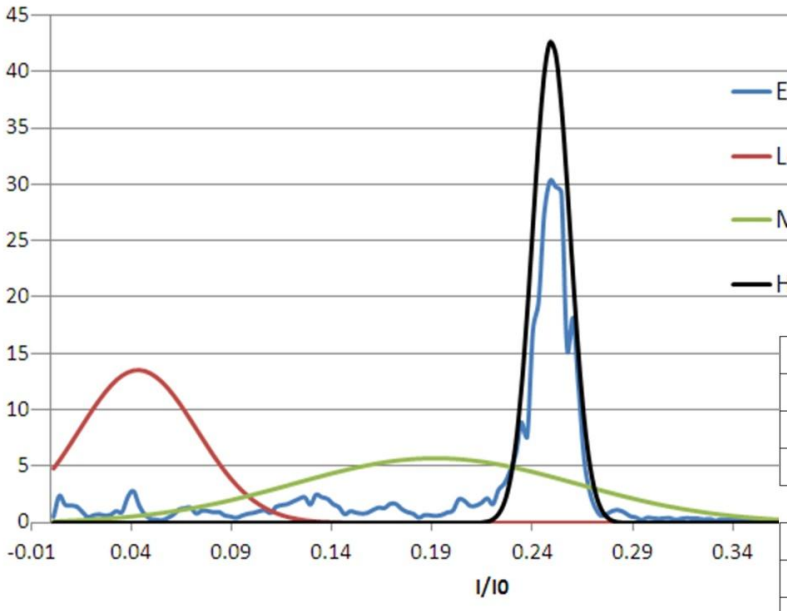


State	π	q (I/I0)	b
Low	0.00006	0.002	0.001
Mid	0.014	0.061	0.049
High	0.986	0.219	0.041

A	Low	Mid	High
Low	0.967	0.033	0.000
Mid	0.036	0.931	0.033
High	0.000	0.013	0.987

Figure F-1. KKRR-ShPrP (120-232) event histogram and optimal HMM model (case 1). (Top left) KKRR-ShPrP(120-232) event histogram of the training data (blue), with the high (black Gaussian), mid (green Gaussian), and low states (red Gaussian), as defined by the optimal HMM model to the training data, overlayed. **(Bottom right)** The optimal three-state HMM model to the training set. Initial condition (π), I/I0 state levels (q), noise properties of each state (b – i.e. the standard deviation on the Gaussian for each state), and the state-to-state transition probability matrix (A).

KKRR-ShPrP(120-232)-D178N



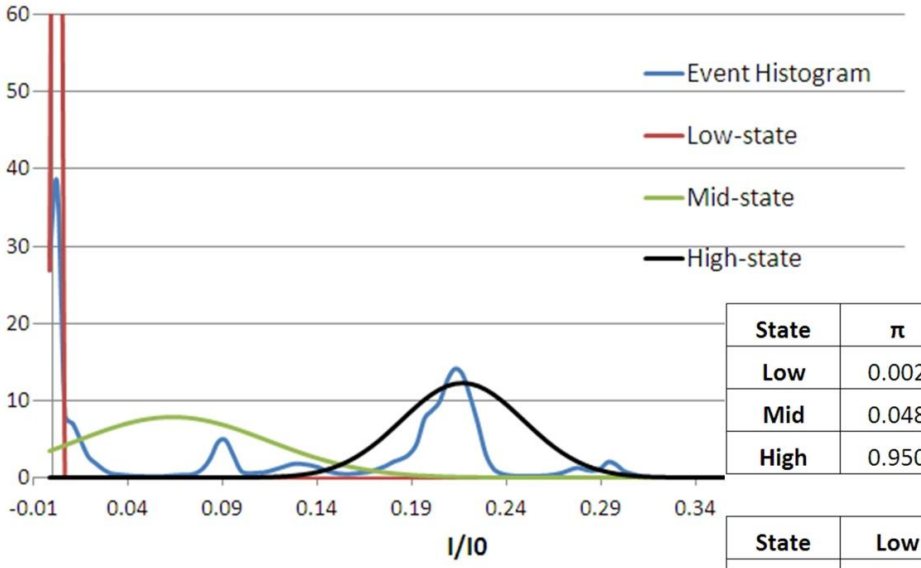
State	π	q (I/I0)	b
Low	0.007	0.044	0.029
Mid	0.860	0.192	0.071
High	0.133	0.250	0.009

State	Low	Mid	High
Low	0.988	0.011	0.001
Mid	0.004	0.880	0.116
High	0.0001	0.045	0.955

Figure F-2. KKRR-ShPrP (120-232)-D178N event histogram and optimal HMM model (case 1). (Top left) KKRR-ShPrP(120-232)-D178N event histogram of the training data (blue), with the high (black Gaussian), mid (green Gaussian), and low states (red Gaussian), as defined by the optimal HMM model to the training data, overlayed. **(Bottom right)** The optimal three-state HMM model to the training set. Initial condition (π), I/I0 state levels (q), noise properties of each state (b – i.e. the standard deviation on the Gaussian for each state), and the state-to-state transition probability matrix (A).

Protein calling case 2: Optimal KKRR-ShPrP(120-232) and KKRR-ShPrP(120-232)-D178N models

KKRR-ShPrP(120-232)

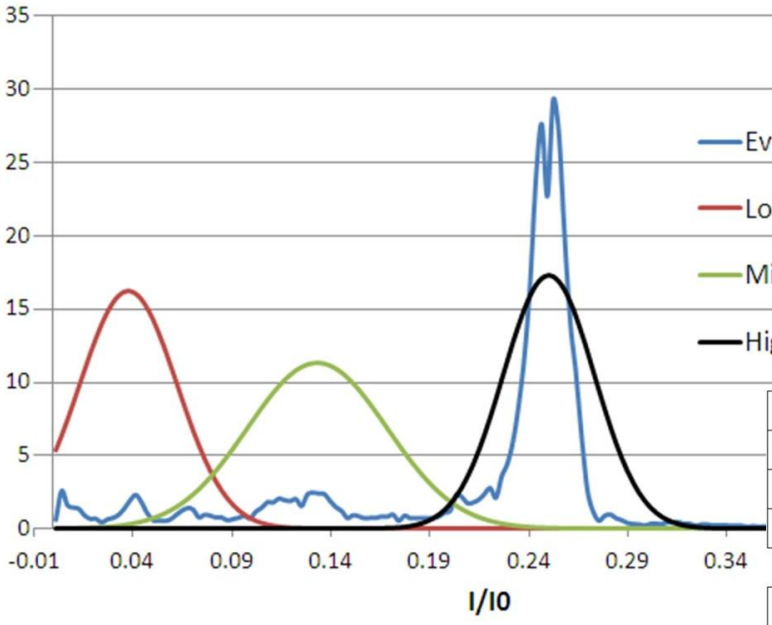


State	π	q (I/I0)	b
Low	0.002	0.002	0.001
Mid	0.048	0.064	0.051
High	0.950	0.217	0.032

State	Low	Mid	High
Low	0.967	0.033	0.000
Mid	0.036	0.933	0.031
High	0.000	0.017	0.983

Figure F-3. KKRR-ShPrP (120-232) event histogram and optimal HMM model (case 2). (Top left) KKRR-ShPrP(120-232) event histogram of the training data (blue), with the high (black Gaussian), mid (green Gaussian), and low states (red Gaussian), as defined by the optimal HMM model to the training data, overlayed. **(Bottom right)** The optimal three-state HMM model to the training set. Initial condition (π), I/I0 state levels (q), noise properties of each state (b – i.e. the standard deviation on the Gaussian for each state), and the state-to-state transition probability matrix (A).

KKRR-ShPrP(120-232)-D178N



State	π	q (I/I0)	b
Low	0.036	0.038	0.025
Mid	0.118	0.133	0.035
High	0.846	0.250	0.023

State	Low	Mid	High
Low	0.981	0.015	0.005
Mid	0.009	0.837	0.155
High	0.001	0.029	0.970

Figure F-4. KKRR-ShPrP(120-232)-D178N event histogram and optimal HMM model (case 2). (Top left) KKRR-ShPrP(120-232)-D178N event histogram of the training data (blue), with the high (black Gaussian), mid (green Gaussian), and low-states (red Gaussian), as defined by the optimal HMM model to the training data, overlayed. **(Bottom right)** The optimal three-state HMM model to the training set. Initial condition (π), I/I0 state levels (q), noise properties of each state (b – i.e. the standard deviation on the Gaussian for each state), and the state-to-state transition probability matrix (A).