UNIVERSITY OF CALIFORNIA, SAN DIEGO

Nanoscale Transport of Electrons and Ions in Water

A Dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Physics with a Specialization in Computational Science

by

Paul Christopher Boynton

Committee in charge:

Professor Massimiliano Di Ventra, Chair
Professor Olga Dudko
Professor Michael Holst
Professor Katja Lindenberg
Professor Oleg Shpyrko

2016

The Dissertation of Paul Christopher Boynton is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2016

DEDICATION

To my mom and dad and my lovely wife

EPIGRAPH

Basic research is what I am doing

when I don't know what I am doing.

–Wernher von Braun

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I want to thank my advisor, Max Di Ventra, for always being enthusiastic about my work and for pushing me to accomplish my goals. His passion for research undoubtedly affected my work, and I thank him for his honesty, understanding, and sense of humor. I would also like to thank my committee members: Olga Dudko, Michael Holst, Katja Lindenberg, and Oleg Shpyrko.

Thanks to all of my group members, past and present, for thinking and talking about physics with me. A special thanks to Matt Krems for showing me the ropes when I first joined the group and for telling me to use Emacs and Python. Thanks to my first office mate, Jim Wilson, for sharing many laughs and making me think about physics and life in new ways. Also thanks to my last office mate, Forrest Sheldon, for many great conversations about both work and play.

Thanks to my parents for being so supportive throughout all of graduate school. I always knew I could call home for useful advice, which is a comfort I'm very grateful for. I'm also thankful that they gave me space (and time) when work had to be done.

Thanks to my wife Alyssa, for ultimately making me a happier person. After working a late night I could always rely on her uplifting spirit and a helping hand. And of course I'm thankful for all of the meals she's made during this time, when working hours are not very well defined. I would also like to thank Alyssa's parents, Jack and Sherry, for flying us back to Miami many times in recent years, providing much needed breaks from work.

Lastly, thanks to all of the good friends I've made here in San Diego. They kept me honest and made the last five and a half years a more joyful experience. And of course thanks to the Art of Espresso for providing an awesome place to get great

coffee and socialize with friends and colleagues.

Chapter 2, in full, is a reprint of material as it appears in P. Boynton and M. Di Ventra, "Probing water structures in nanopores using tunneling currents", *Phys. Rev. Lett.*, vol. 111, no. 21, p. 216804, 2013. The dissertation author was the primary investigator and author of this publication.

Chapter 3, in full, is a reprint of material as it appears in P. Boynton, A. Balatsky, I. Schuller, and M. Di Ventra, "Improving sequencing by tunneling with multiplexing and cross-correlations", *J. Comput. Electron.*, vol. 13, no. 4, pp. 794-800, 2014. The dissertation author was the primary investigator and author of this publication.

Chapter 4, in full, is a reprint of material as it appears in P. Boynton and M. Di Ventra, "Sequencing proteins with transverse ionic transport in nanochannels", 2015 (under peer-review). arXiv:1509.04772 [physics.bio-ph]. The dissertation author was the primary investigator and author of this publication.

Appendix A, in full, is a reprint of material as it appears in P. Boynton and M. Di Ventra, "Sequencing proteins with transverse ionic transport in nanochannels", 2015 (under peer-review). arXiv:1509.04772 [physics.bio-ph]. The dissertation author was the primary investigator and author of this publication.

VITA

2010	Bachelor of Science in Physics and Applied Mathematics *magna cum laude*, University of Miami
2012	Master of Science in Physics, University of California, San Diego
2016	Doctor of Philosophy in Physics with a Specialization in Computational Science, University of California, San Diego

PUBLICATIONS

P. Boynton and M. Di Ventra, "Sequencing proteins with transverse ionic transport in nanochannels", 2015 (under peer-review). arXiv:1509.04772 [physics.bio-ph].

P. Boynton, A. Balatsky, I. Schuller, and M. Di Ventra, "Improving sequencing by tunneling with multiplexing and cross-correlations", *J. Comput. Electron.*, vol. 13, no. 4, pp. 794-800, 2014.

P. Boynton and M. Di Ventra, "Probing water structures in nanopores using tunneling currents", *Phys. Rev. Lett.*, vol. 111, no. 21, p. 216804, 2013.

J. Cohn, P. Boynton, J. Trivino, J. Trastoy, B. White, C. dos Santos, and J. Neumeier, "Stoichiometry, structure, and transport in the quasi-one-dimensional metal Li 0.9 Mo 6 O 17", *Phys. Rev. B*, vol. 86, no. 19, p. 195143, 2012.

ABSTRACT OF THE DISSERTATION

Nanoscale Transport of Electrons and Ions in Water

by

Paul Christopher Boynton

Doctor of Philosophy in Physics with a Specialization in Computational Science

University of California, San Diego, 2016

Professor Massimiliano Di Ventra, Chair

The following dissertation discusses the theoretical study of water on the nanoscale, often involved with essential biological molecules such as DNA and proteins. First I introduce the study of water on the nanoscale and how experimentalists approach confinement with nanopores and nanogaps. Then I discuss the theoretical method we choose for understanding this important biological medium on the molecular level, namely classical molecular dynamics. This leads into transport mechanisms that utilize water on the nanoscale, in our case electronic and ionic transport. On the scale of mere nanometers or less electronic transport in water enters the tunneling regime, requiring

the use of a quantum treatment. In addition, I discuss the importance of water in ionic transport and its known effects on biological phenomena such as ion selectivity. Water also has great influence over DNA and proteins, which are both introduced in the context of nanopore sequencing. Several techniques for nanopore sequencing are examined and the importance of protein sequencing is explained. In Chapter 2, we study the effect of volumetric constraints on the structure and electronic transport properties of distilled water in a nanopore with embedded electrodes. Combining classical molecular dynamics simulations with quantum scattering theory, we show that the structural motifs water assumes inside the pore can be probed directly by tunneling. In Chapter 3, we propose an improvement to the original sequencing by tunneling method, in which N pairs of electrodes are built in series along a synthetic nanochannel. Each current time series for each nucleobase is cross-correlated together, reducing noise in the signals. We show using random sampling of data from classical molecular dynamics, that indeed the sequencing error is significantly reduced as the number of pairs of electrodes, N, increases. In Chapter 4, we propose a new technique for *de novo* protein sequencing that involves translocating a polypeptide through a synthetic nanochannel and measuring the ionic current of each amino acid through an intersecting *perpendicular* nanochannel. We find that the distribution of ionic currents for each of the 20 proteinogenic amino acids encoded by eukaryotic genes is statistically distinct using our theoretical method.

Chapter 1

Introduction

Water is an essential biological medium, so present that we often take it for granted. Every organism we know of needs it to survive, mainly due to water's ability to dissolve organic molecules and essential salts. In addition, Water composes about 70% of a typical cell by volume, which makes it the natural medium for biological activity as well as for studying biological molecules. Yet with all of our understanding of the importance of water's bulk properties and functions we still do not fully understand its properties on the nanoscale, a scale where many essential biological processes take place. Within this dissertation I intend to study the topic of water on the nanoscale with novel approaches, particularly utilizing the transport of electrons and ions. In doing so I will shed light on some of the features of water at this scale and on the way introduce new ideas for sequencing DNA and proteins, water being core to the results and my theoretical analysis. But first I introduce below the well-studied components that lead into my own research.

1.1 Water on the Nanoscale

1.1.1 Nanopores and Nanogaps

A structure known as a nanopore, a nanoscale sized pore in a membrane, allows one to confine a molecule of interest for further study. This is particularly useful in the burgeoning field of biosensing, in which one seeks to detect and identify single biological molecules such as the nucleotides of a strand of DNA or the amino acids of a protein. Furthermore, nanopores can now be fabricated to atomic resolution with mechanically-controllable break junctions (MCBJs) [1] or controlled dielectric breakdown [2], dwarfing the focused ion beam milling or electron beam methods that first achieved nanometer resolution [3]. The nanogaps, or nanopores with one dimension of confinement instead of two, from [1] were made with gold MCBJs [4, 5] inside of a dielectric SiO_2 nanopore to achieve a nanopore device related to Fig. 2.1. In addition to having atomic resolution these nanogaps can also be as small as one atom across, resulting in a device that allows the study of molecules as small as water.

This device won't be the first capable of probing the properties of water, however. In fact, water has already been studied when confined on the nanoscale with several different experimental techniques such as neutron scattering [6] and inelastic x-ray scattering [7], as explained in Section 2.1, but these methods are difficult to employ.

1.1.2 Modeling Water

To simulate water on the nanoscale one requires information on each water molecule, yet systems have to be large enough to avoid boundary effects. To satisfy

both of these restrictions with reasonable compute times one finds classical molecular dynamics (MD). Since simulations are run at room temperature or higher, quantum effects become negligible with respect to the dynamics of system. Classical MD is Newtonian and therefore calculates the potential energy from bonds (harmonic), angles (harmonic), dihedrals or torsion angles(sinusoidal), impropers or out-of-plane bending (harmonic), Urey-Bradley cross-terms (harmonic), van Der Waals or VDW interactions (Lennard-Jones), and lastly electrostatic interactions (Coulomb), which can be calculated more efficiently using the particle mesh Ewald method. With NAMD, the MD software we use, the position, velocity, time, and energy is used in the Velocity Verlet algorithm to efficiently step forward in time on the femtosecond scale while preserving the symplectic form on phase space [8]. To keep a constant temperature in the system a Langevin thermostat is used, which adds a damping term to Newton's equations and randomly reassigns velocities as needed.

NAMD is used in Chapters 2, 3, and 4 in conjunction with the TIP3P water model, a 3 site model (for atomic position and charge) that reproduces the dielectric constant of water.

1.2 Electronic and Ionic Transport on the Nanoscale

1.2.1 Electronic Transport

Systems with electrodes separated by a dielectric gap of mere nanometers or less, as in Chapters 2 and 3, enter the tunneling regime. Since tunneling is purely a quantum effect the electronic transport must be treated quantum mechanically. We choose to use a single-particle scattering approach as illustrated in Fig. 1.1, where

electrons are injected from infinity into the left (L) and right (R) reservoirs with different equilibrium distribution functions based on the bias across the nanojunction. The nanojunction is the system that the electrons scatter off of, either reflecting back to the way they came or tunneling through to the other side.

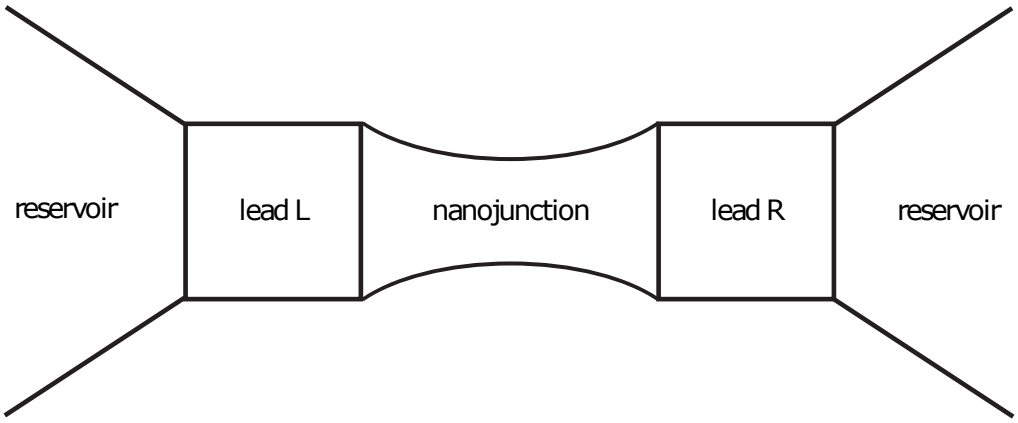

Figure 1.1: Schematic of the scattering approach used in Chapters 2 and 3. The nanojunction, the object of our study, is sandwiched between two ideal leads. Each of those leads is connected to a reservoir of electrons that extends to infinity.

In Chapter 2, we study the effect of volumetric constraints on the structure and electronic transport properties of distilled water in a nanopore with embedded electrodes. Combining classical molecular dynamics simulations with quantum scattering theory, we show that the structural motifs water assumes inside the pore can be probed directly by tunneling. In particular, we show that the current does not follow a simple exponential curve at a critical pore diameter of about 8 Å, rather it is larger than the one expected from simple tunneling through a barrier. This is due to a structural transition from bulk-like to "nanodroplet" water domains. Our results can be tested with present experimental capabilities to develop our understanding of water as a complex medium at nanometer length scales.

1.2.2 Ionic Transport

In nature, nanochannels (nanopores with much greater length than diameter) are used extensively for ionic transport due to their ability to balance or imbalance voltage differences. In a neuron the protein nanochannels are voltage-gated so that when the axon membrane potential rises enough, it triggers the Na^+ channels to open for Na^+ ions to enter the axon and further increase the membrane potential, triggering a positive feedback loop known as an action potential. The cascade gets reversed when the membrane potential gets even higher and triggers the K^+ channels to open and send K^+ ions out of the axon, eventually restoring the resting potential. This is just one example of the importance of nanochannels and ionic transport in biology.

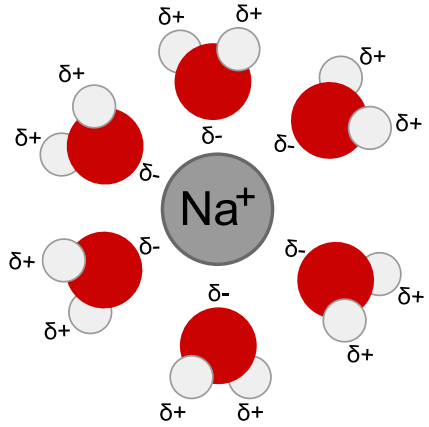

Figure 1.2: (Color online) Schematic of a sodium ion, Na^+, surrounded by a hydration layer. The positive ion attracts the negative oxygen (red), indicated by the partial charge δ-. The positive hydrogens (white) with partial charge δ+ are repelled by Na^+. Credit - Public Domain.

But what prevents negative ions from entering the membrane instead of the Na^+ ions? The reason is that these biological channels have ion selectivity due to surface charges that accept only positive charges, along with the fact that Na^+ is smaller than its competitor Cl^-. But many other ion channels are selective due to

hydration layers, which are tightly bound water molecules that form rough spherical shells around ions (see Fig. 1.2). For example, gramicidin is a neutral but polar channel that accepts most monovalent cations but rejects Cl^-, smaller than many monovalent cations. However, gramicidin only accepts ions and water in single file due to its size and Cl^- is tightly bound to its first hydration layer, making it entropically unfavorable to enter the channel compared to its cation competition. In addition, the asymmetry of water causes the first hydration layer of Cl^- to be frustrated, which increases the effective volume of this pseudoparticle [9].

1.3 Applications in Sequencing

1.3.1 DNA Sequencing

DNA consists of 4 types of monophospate nucleotides, each associated with a nucleobase. The 4 nucleobases are adenine (A), cytosine (C), guanine (G), and thymine (T), and to make a monophosphate nucleotide a nucleobase must bond with deoxyribose and a phosphate group. The phosphodiester bonds then connect pairs of nucleotides together by their sugars to make DNA's primary structure, the single-stranded DNA (ssDNA) as in 1.3. Although the double helix DNA (dsDNA), a secondary structure, is the better known form of DNA, ssDNA proves to be easier to distinguish with nanopores due to the exposed nucleobases [10].

Nanopore sequencing of DNA started with longitudinal ionic transport through the biological nanopore α-hemolysin using a longitudinal electric field to drive the DNA through the pore due to the negative charge of DNA from its phosphate backbone negative charge [11]. However, this method lacks true single base discrimination and

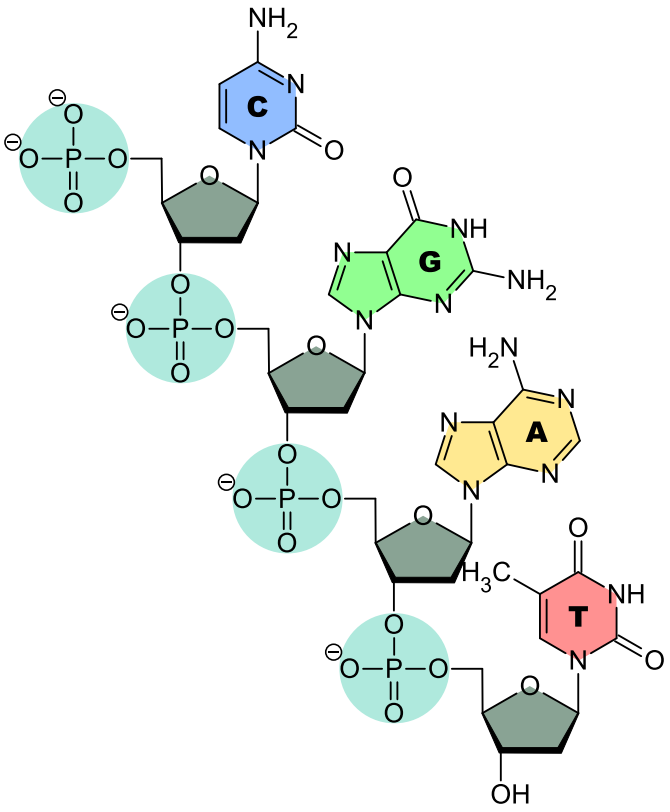

Figure 1.3: (Color online) A structural formula for nucleobases C, G, A, and T each bonded to deoxyribose and then linked together via phosphodiester bonds (one phosphate group to connect two nucleosides). Credit - user:Sponk / Wikimedia Commons / Public Domain.

so a new method for nanopore sequencing was developed. Sequencing by tunneling is a next-generation approach to read single-base information using electronic tunneling transverse to the single-stranded DNA (ssDNA) backbone while the latter is translocated through a narrow channel. The original idea considered a single pair of electrodes to read out the current and distinguish the bases [12, 13].

In Chapter 3, we propose an improvement to the original sequencing by tunneling method, in which N pairs of electrodes are built in series along a synthetic nanochannel. While the ssDNA is forced through the channel using a longitudinal field it passes by each pair of electrodes for long enough time to gather a minimum of

m tunneling current measurements, where m is determined by the level of sequencing error desired. Each current time series for each nucleobase is then cross-correlated together, from which the DNA bases can be distinguished. We show using random sampling of data from classical molecular dynamics, that indeed the sequencing error is significantly reduced as the number of pairs of electrodes, N, increases. Compared to the sequencing ability of a single pair of electrodes, cross-correlating N pairs of electrodes is *exponentially* better due to the approximate log-normal nature of the tunneling current probability distributions. We have also used the Fenton-Wilkinson approximation to analytically describe the mean and variance of the cross-correlations that are used to distinguish the DNA bases. The method we suggest is particularly useful when the measurement bandwidth is limited, allowing a smaller electrode gap residence time while still promising to consistently identify the DNA bases correctly.

1.3.2 Protein Sequencing

Proteins are built from a sequence of amino acids, of which there are 20 that are encoded by eukaryotic genes ignoring selenocysteine. These 20 amino acids differ in size, structure, and charge state greatly as seen in Fig. 1.4. However, all of these amino acids have the same core degrees of freedom that help determine secondary structures (e.g. α-helices), those being backbone dihedrals ψ and ϕ as in Fig. 1.5. When linked together by peptide bonds amino acids have a total of 3 backbone dihedrals, of which ω stays close to constant. Since bond angles and lengths are also fairly rigid, *psi* and *phi* must agree to form a stable structure. A Ramachandran diagram plots this relationship between *psi* and *phi* for a given amino acid, often identifying connected regions in the phase space with secondary structures.

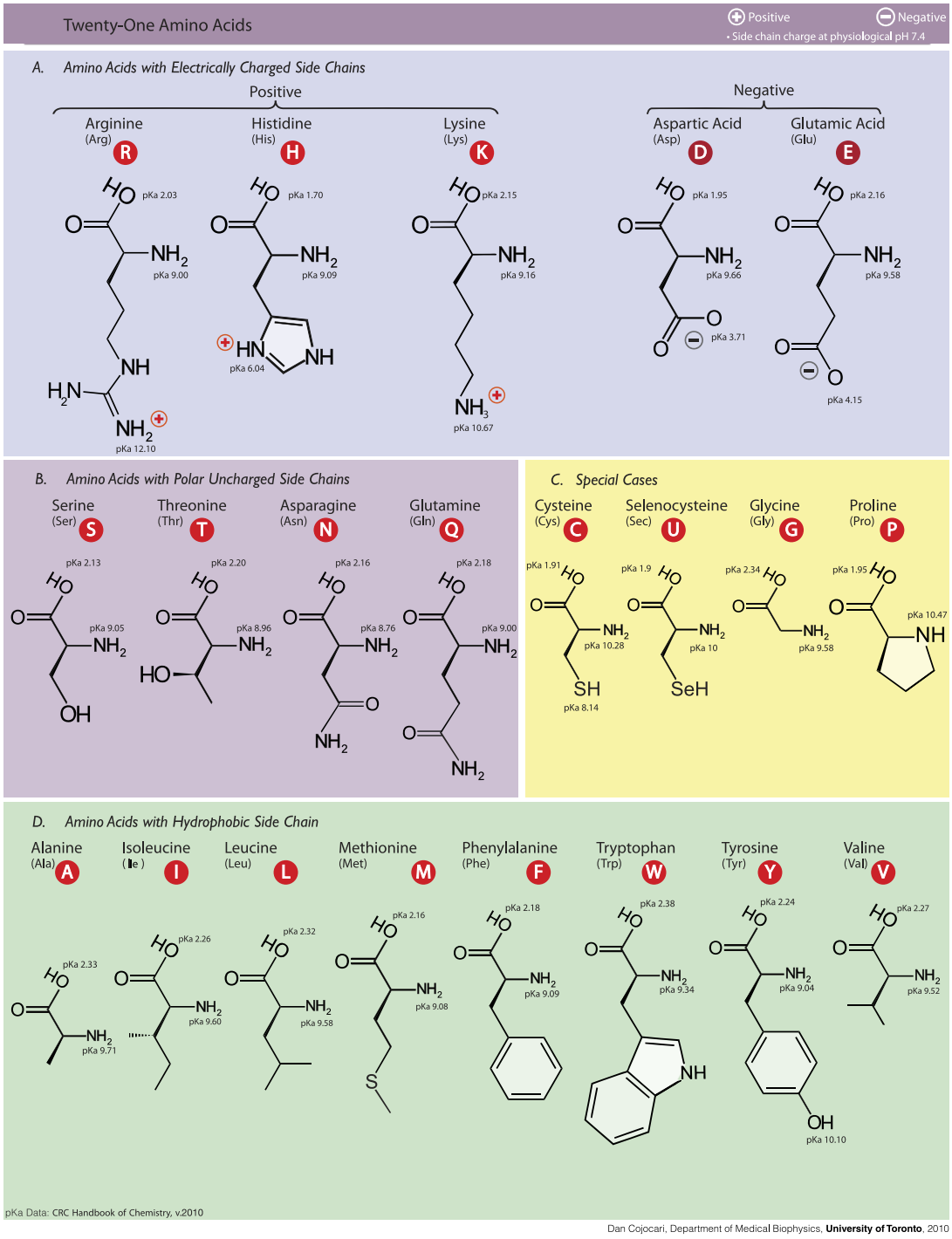

Figure 1.4: (Color online) Structure formulas for all 21 proteinogenic amino acids encoded by eukaryotic genes, including selenocysteine, organized by side chain charge at physiological pH. Credit - Dan Cojocari / CC-BY-SA-3.0.

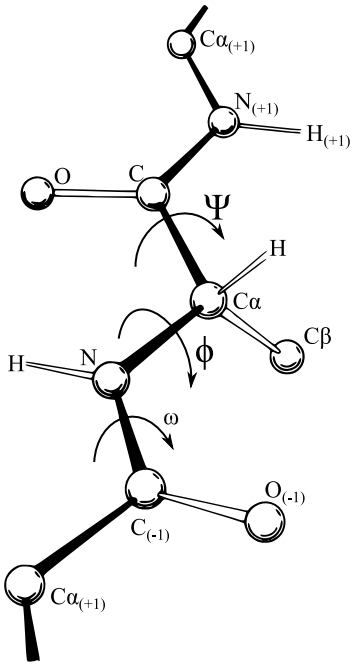

Figure 1.5: Schematic of a protein backbone with dihedral angles ψ, ϕ, and ω. Credit - Jane Shelby Richardson / CC-BY-SA-3.0, vectorised by Adam Redzikowski.

Proteins only truly obtain their biological purpose when their tertiary structure is formed, which is why protein folding and protein structure prediction are so well studied. A single defect in a protein sequence could alter the tertiary structure's binding site and destroy the purpose of the protein. To that effect, *De novo* protein sequencing is essential for understanding cellular processes that govern the function of living organisms and all post-translational events and other sequence modifications that occur after a protein has been constructed from its corresponding DNA code. By obtaining the order of the amino acids that composes a given protein one can then determine both its secondary and tertiary structures through structure prediction, which is used to create models for protein aggregation diseases such as Alzheimer's Disease. Mass spectrometry is the current technique of choice for *de novo* sequencing. However, because some amino acids have the same mass the sequence cannot be

completely determined in many cases.

In Chapter 4, we propose a new technique for *de novo* protein sequencing that involves translocating a polypeptide through a synthetic nanochannel and measuring the ionic current of each amino acid through an intersecting *perpendicular* nanochannel. To calculate the transverse ionic current blockaded by a given amino acid we use a Monte Carlo method along with Ramachandran plots to determine the available flow area, modified by the local density of ions obtained from molecular dynamics and the local flow velocity ratio derived from the Stokes equation. We find that the distribution of ionic currents for each of the 20 proteinogenic amino acids encoded by eukaryotic genes is statistically distinct, showing this technique's potential for *de novo* protein sequencing.

Chapter 2

Probing Water Structures in Nanopores using Tunneling Currents

2.1 Introduction

Liquid water is a very common and abundant substance that is considered a fundamental ingredient for life more than any other. Yet we do not fully understand many of its properties, especially when we probe it at the nanometer scale, although a lot of research has been done on this important system in this regime [14, 15, 6, 7, 16, 17]. Some of the first experimental studies of water on the nanoscale have been done using a scanning-tunneling microscope (STM) [14], in which the tunneling barrier height was found to be unusually low. This was hypothesized to be the result of the three-dimensional nature of electron tunneling in water. Some STM experiments actually studied the tunneling current as a function of distance to understand the solid/liquid

interface and found that the tunneling current oscillates with a period that agrees with the effective spacing of the Helmholtz layers [15]. Water has also been studied when encapsulated by single-walled carbon nanotubes in which, via neutron scattering, the water was observed to form a cylindrical "square-ice sheet" which enclosed a more freely moving chain of molecules [6]. These structures are related to the fact that these carbon nanotubes have cylindrical symmetry and are hydrophobic. More recently, the dynamics of water confined by hydrophilic surfaces were studied by means of inelastic X-ray scattering showing a phase change at a surface separation of 6 Å. Well above 6 Å there are two deformed surface layers that sandwich a layer of bulk-like water but below 6 Å the two surface layers combine into one layer that switches between a localized "frozen" structure and a delocalized "melted" structure [7]. On the computational side, many molecular dynamics (MD) simulations have been done to study water in a variety of environments. Of particular interest has been the study of hydrophobic channels because in this case water has been shown to escape from the channel altogether for entropic gain [16, 17]. However, these structures, and in particular the formation of water nanodroplets, are difficult to probe experimentally.

Recent interest in fast DNA sequencing approaches has been crucial to the advancement of novel techniques to probe polymers in water environments at the nanometer scale. In particular, the proposal to sequence DNA by tunneling [13, 18] has been instrumental for the development of sub-nanometer electrodes embedded into nanochannels [4, 19, 20]. These techniques open the door to investigating the properties of liquids volumetrically constrained by several materials by relating the local structure of the liquid to electrical (tunneling) currents.

In this Letter, we take advantage of these newly-developed experimental tech-

niques and propose the study of water in nanopores with embedded electrodes. We find that indeed the structural motifs water assumes inside pores of different diameters can be probed directly by tunneling. In fact, we predict that the tunneling current does not follow a simple exponential curve at a critical pore diameter of about 8 Å as simple tunneling through a barrier would produce. Instead, water domains form a specific density of states which in turn gives rise to these peculiar features. Our findings can be tested with the available experimental capabilities on similar systems [4, 19, 20].

To better understand the nature of this substance on the nanoscale, we study the effects of confinement on water's structure and electronic transport properties in silicon nitride nanopores using classical molecular dynamics (MD) combined with quantum transport calculations. Since the system is at room temperature quantum effects related to protons are negligible, which allows us to use NAMD 2.7 [8], a highly parallel classical MD application. We have chosen to work with Si_3N_4 nanopores because they are readily fabricated to have very small constrictions. Note also that the environment we consider is not hydrophobic because silicon nitride (Si_3N_4) nanostructures are known to have dangling atoms that produce polar surfaces [21].

2.2 Methods

2.2.1 Molecular Dynamics

The system is built in a manner similar to the synthetic pore in [22]. Using VMD [23], we build a β-Si_3N_4 membrane containing a double-conical pore with inner diameter ranging from 4.5 to 9.25 Å (atom center to atom center). The membrane

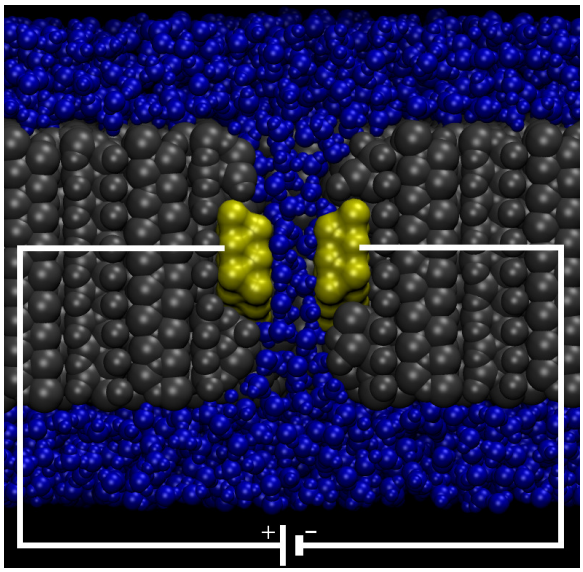

Figure 2.1: (Color online) A cross section of a nanopore for MD after equilibration. The Si_3N_4 (gray) membrane is cut into a hexagonal prism 2.6 nm thick and 9.1 nm wide (from one vertex to the opposing hexagonal vertex) to satisfy periodic boundary conditions in all three dimensions of space. We use a double-conical pore with a 10 ° slant off of cylindrical to better represent currently fabricated nanopores. The Si_3N_4 is harmonically constrained to reproduce its dielectric behavior in experiments. The system is solvated to create a water (blue) reservoir above and below the pore of combined thickness 2.6 nm. The electric field goes from left to right between the gold (yellow) electrodes.

includes two fixed embedded gold electrodes that span the small constriction at the center of the pore, similar to the pairs of electrodes introduced in [13, 18]. Above and below the membrane lie water reservoirs of about 3200 molecules combined that provide a buffer between periodic images, and provide the bulk with which the water molecules in the pore can be recycled (see Fig. 2.1). We can safely ignore any entrance effects on the structure of the confined water between the electrodes since the correlation length of bulk water at room temperature is between 1.5 and 2 Å [24] and each pore entrance lies approximately 8 Å from the region of interest. In addition, the confinement is gradual due to the double-conical shape and testing has been done

on similar nanopores filled with water and ions to show that increasing the thickness of the pore beyond 2.6 nm does not change the basic dynamics of the system [25]. We utilize the CHARMM27 force field [26, 27] for the interactions of the TIP3P water whereas we use UFF parameters for the Si_3N_4 [28]. Furthermore no ions are added to the system to permit the study of pure water structures and their effect on tunneling current. A Langevin thermostat keeps temperature set to 295 K with a damping coefficient of 1 ps^{-1} applied to the Si_3N_4 while a bias of 1 V between the embedded electrodes is achieved using the Grid-steered Molecular Dynamics feature in NAMD 2.7 [8]. More specifically, we impose a 3-dimensional potential grid on the region between the gold electrodes which is linear in the transverse axis and constant in the remaining two. Padding must be added to the box so that there are no discontinuities in the field at the edges, after which the grid is interpolated by cubic polynomials. From these polynomials the gradient is taken to obtain the electric field, which has been checked to be accurate in the affected region. Note that we do not treat the image forces created by the polarized water's proximity to an equipotential surface (both electrodes). However, our calculations show that the strength of these forces is less than the force due to the 1 V bias, although within an order of magnitude for the water molecules on the electrode surface. Therefore, we expect that since the image forces act to align the water, much like the external field does, the structural effects we find can only be enhanced by their inclusion. The entire equilibrated system evolves over 5 ns in an NVT ensemble with 1 fs time steps, yet atomic coordinates are recorded every ps.

2.2.2 Current Distributions

Position data snapshots of the gold and the surrounding water molecules are then taken to evaluate the current over time. Although each snapshot is only a static representation of the system we can safely calculate the tunneling current at each recording. This is because the time scale governing the tunneling electrons ($\sim 10^{-15}$ s) is much smaller than the time scale of the relevant dynamics of the water molecules ($\sim 10^{-12}$ s) [29]. The effect of dephasing and other inelastic effects have been estimated to be small for the distribution of currents at reasonable electron-molecular vibrations and rotational time scales [18], thus they can be neglected. To calculate the current we use a single-particle scattering approach that involves obtaining a tight-binding Hamiltonian (see, e.g., [30]) and the single-particle retarded Green's function, as detailed in [18] with no added noise. The resultant current is given by $I = \frac{2e}{h} \int_{-\infty}^{\infty} dE\, T(E)[f_r(E) - f_l(E)]$ where e is the elementary charge, h is Planck's constant, E is the energy of the scattering electron, T is the total transmission function, and f_r and f_l are the right and left electrode Fermi-Dirac distribution functions, respectively.

In our analysis we cut out the first 1000 snapshots to eliminate any transient behavior. The currents from the remaining 4001 snapshots are binned to give a current distribution for each nanopore diameter, as in Fig. 2.2. The distributions take the form of approximate Gaussian distributions, however the current axis is on a log scale. Therefore the distributions are approximately of the form

$$P(I) \sim \exp\left\{\frac{-(\log(I/\mu) + 3cs^2/2)^2}{2s^2}\right\}, \tag{2.1}$$

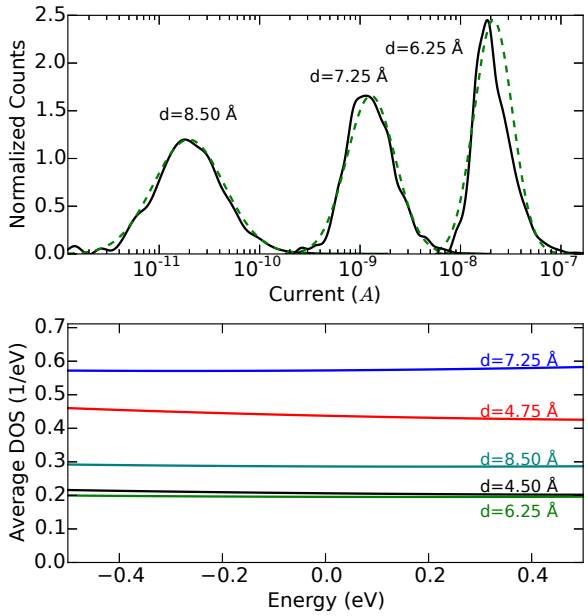

Figure 2.2: (Color online) Top: Normalized distributions of current with the current axis on a log scale. The solid lines reflect the normalized distributions of current at pore diameters of 6.25, 7.25, and 8.5 Å. The dashed lines are Gaussian fits to each distribution of $\log I$. Bottom: Time averaged DOS as a function of energy referenced at the Fermi level of gold for different pore diameters.

where $P(I)$ is the probability of realizing the current value I, μ is the average current, $c = \ln 10$, and s is the standard deviation of the distribution of $\log I$. This is equivalent to stating that the distribution of $\log I$ is approximately a normal distribution. These observations suggest that the coupling between the electrodes and the water molecules is controlling the current distributions [18].

2.3 Results and Discussion

The current averages are plotted against pore diameter in Fig. 2.3. A simulation was run for every diameter from 4.5 to 9.25 Å in 0.25 Å intervals. Below 4.5 Å water is completely excluded from the region between the electrodes and above 9.25 Å

we obtain sub-pA average currents which cannot be easily detected with present techniques. As a medium water acts to reduce the effective barrier height to about 1 eV [31, 19], much lower than the work function of gold (4.3 eV [32]). To emphasize this point we have calculated the current of a rectangular tunneling barrier in which the barrier height is the work function of gold and the barrier width is the diameter minus twice the distance between the edge of a jellium electron model with the gold density ($r_s = 3$) and the center of the closest plane of gold atoms [33]. The result of this calculation (dashed line in Fig. 2.3) gives currents that are generally smaller than those of the MD simulations (solid line) until about 8.5 Å. This is due to the simplistic choice of geometry of the barrier which becomes more influential as the pore diameter increases. The lower inset of Fig. 2.3 shows how the current standard deviation follows the same trend as the current average and generally decreases with increased pore diameter. Even so, the standard deviations and averages still differ and resemble each other because of the nature of the distributions of current in Eq. (2.1) and the range of current values spanning one to two orders of magnitude as seen in Fig. 2.2.

The first feature to notice in Fig. 2.3 is the deviation from the line of exponential dependence that includes diameters 7.25 to 8.25 Å. Since the current axis is on a log scale, this deviation appears deceptively small. However the MD current values can be several times larger than the currents from a regression in the domain of the deviation. We now show that this increase in the tunneling current is the result of *structural changes* in the water.

To study the structure of water we plotted time averaged density profiles of oxygen for several pores as detailed in Fig. 2.4. The number of water molecules in

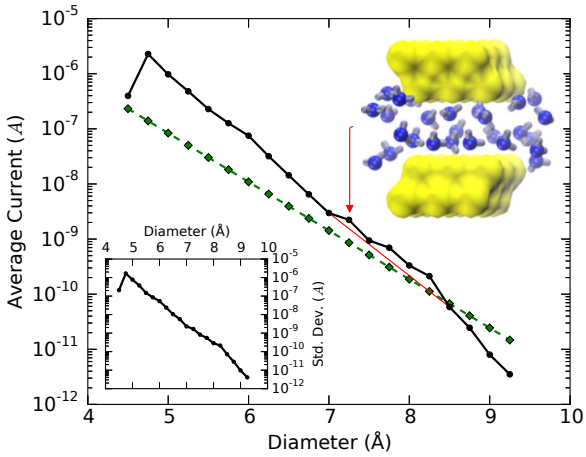

Figure 2.3: (Color online) Average current plotted against pore diameter with the current axis on a log scale. The solid line reflects the time averaged current from the MD simulations while the dashed line reflects the current of a rectangular barrier (see text). The red line connects the points at 7 and 8.5 Å and is there only to highlight the bump in the current. The lower inset demonstrates the standard deviation of the current from MD against pore diameter again with the current axis on a log scale. The upper inset, created with VMD [23], shows a snapshot of the gold (yellow) electrodes and the surrounding water (blue) at a pore diameter of 7.25 Å.

each snapshot can go from about 10 in the case of the 4.5 Å pore diameter to about 45 in the case of 9.25 Å. The density maps are plotted at the midpoint of each respective z-slice of space along the (x, y) plane with the magenta outlines representing the pore-electrode boundary at the midpoint of each z-slice. At the center of the pore the boundary is dominated by the flat electrodes yielding near rectangular confinement, whereas the density above and below the electrodes shows the circular structures created in a cylindrical confinement.

We first notice that at 6.25 Å only one layer of water can fit between the electrodes, although there is enough space for large fluctuations as seen by the blurred density. At about 7.25 Å we see the formation of two layers of water packed together tightly. In fact, these layers start forming "nanodroplets" to reduce energy (see inset in Fig. 2.3). Lastly at 8.5 Å the bilayer of water is smeared all over the confined space

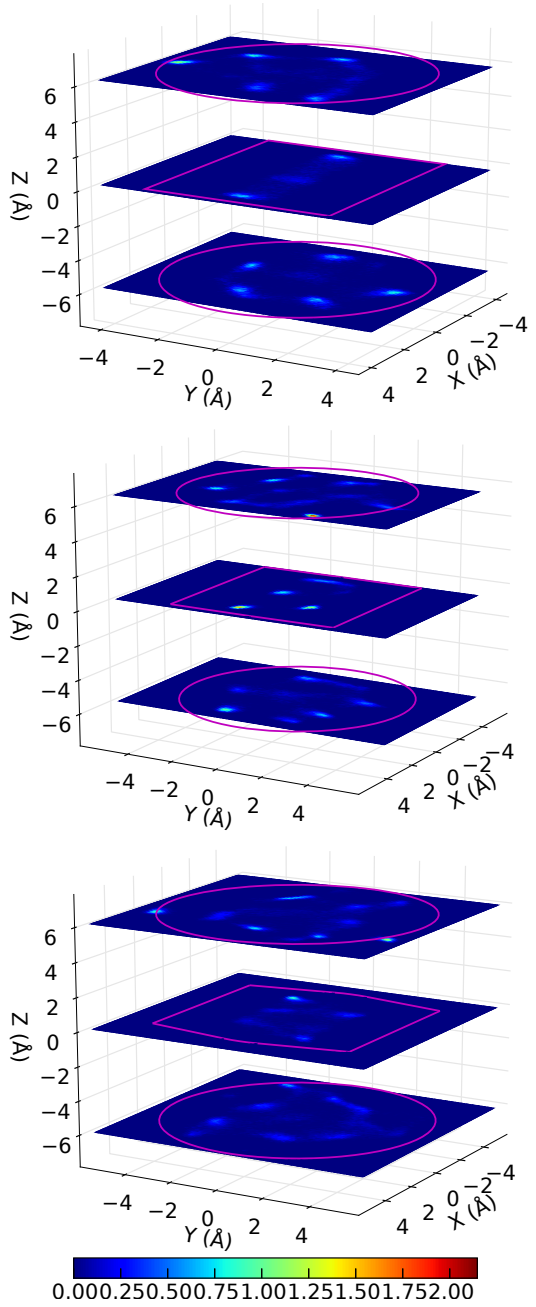

Figure 2.4: (Color online) Time averaged water density plots for the (Top: 6.25, Center: 7.25, Bottom: 8.5) Å diameter pore. The oxygen density (number/Å^3) is displayed via color. The y-axis is the transverse electrode axis, the z-axis is the longitudinal axis, and the x-axis is the remaining transverse axis. The z-axis has been divided into five parts with the bottom, center, and top fifths of oxygen being shown. The magenta outlines represent the pore-electrode boundary.

implying large fluctuations again. These structural tendencies coordinated with length reinforce the conjecture that the deviation of the current from a simple exponential form is correlated to a sort of structural criticality. However, we have checked that these critical structures do not resemble a solid form of water. Instead thermal fluctuations cause water molecules to be exchanged with the bulk so that several molecules may explore the available space between the electrodes for an energetically favorable position.

In order to have a better understanding of the effect of these structural motifs we have computed the density of states (DOS) at the different diameters. In fact, the DOS roughly follows the same trend as that of the current (see bottom panel of Fig. 2.2): at about 7.25 Å the DOS is larger than at any other diameter. This implies that certain structural forms of water introduce more states for the tunneling electrons to utilize, effectively reducing the barrier between the electrodes, thus increasing the current. In the case of 6.25 Å the DOS is the smallest of those shown in Fig. 2.2, meaning that the fluctuating single layer of water that we see in Fig. 2.4 introduces less states for the tunneling electrons to utilize.

The last feature to notice in Fig. 2.3 is the abrupt change in current from 4.5 to 4.75 Å. This corresponds to the first occasion in which water molecules can enter the space between the electrodes. At a pore diameter of 4.5 Å the DOS remains low due to the exclusion of water, but increases dramatically when water molecules are confined between the electrodes at a pore diameter of 4.75 Å (see bottom panel of Fig. 2.2). Although the distance between the electrodes slightly increases, the tunneling current actually increases because of the introduction of water molecules and therefore an increased DOS.

2.4 Conclusions

In conclusion, we have shown with a combination of MD and quantum transport simulations the potential for probing the structure of confined water in nanopores with tunneling currents. We found that the distributions of the log of the current are normal, suggesting that the coupling between the electrodes and the water molecules governs the form of the distributions [18]. We also find a highly non-linear dependence of the log of the current as a function of pore diameter. This non-linearity is due to the introduction of states for electrons to tunnel through when water molecules form nanodroplets between the electrodes. Because the effective tunneling barrier is reduced when electrons tunnel through water compared to vacuum [31, 19], we record currents in the range of pA to μA. These values as well as the recent demonstration of nanopores with embedded electrodes make our predictions within reach of experimental verification.

2.5 Acknowledgements

This work was supported by the National Institute of Health. We thank Chun-Keung Loong for useful discussions.

Chapter 2, in full, is a reprint of material as it appears in P. Boynton and M. Di Ventra, "Probing water structures in nanopores using tunneling currents", *Phys. Rev. Lett.*, vol. 111, no. 21, p. 216804, 2013. The dissertation author was the primary investigator and author of this publication.

Chapter 3

Improving Sequencing by Tunneling with Multiplexing and Cross-correlations

3.1 Introduction

A cheap and fast method to sequence DNA would revolutionize the way health care is conducted [10]. With such a method, medicine would be catered to the individual based on genetic implications, an approach that goes under the name of personalized or precision medicine. The research behind DNA sequencing is rich and plentiful, with many techniques that have much potential. Two of the most successful techniques currently used, single molecule real time sequencing (SMRT) [34] and ion torrent semiconductor sequencing (ITS) [35], need on the order of 10 hours, including full preparation time, for one run, which sequences 1 Gb and 100 Mb, respectively [36]. Both techniques take advantage of massively-parallel sequencing to achieve these

benchmarks.

However, most of the current sequencing techniques, SMRT included, require fluorescent dyes to distinguish the DNA bases [36]. In other words, these techniques cannot greatly improve in speed and are inherently costly, both for the sample preparation, equipment and to operate. On the other hand, ITS does not utilize fluorescent dyes but instead depends on the detection of hydrogen ions released once a deoxyribonucleotide triphosphate (dNTP) forms a covalent bond with a complementary nucleotide [35]. This means that the overall costs are smaller in comparison but the technique nevertheless suffers from small read lengths of about 200 base pairs per run [36], implying the technique would be difficult (or too costly) to apply to *de novo* sequencing.

Quite recently a new approach has been suggested that envisions the sequencing of single-stranded DNA (ssDNA) with electronic currents transverse to the DNA backbone as it passes through a nanochannel [12, 13]. A schematic is shown in Fig. 3.1. This approach has been recently demonstrated experimentally by sequencing micro-RNA and short DNA oligomers [37].

When the electrodes are fabricated so that the gap only allows a single base to fit at a time [4], one can truly obtain single-base discrimination without the need of amplification or chemicals. Because of the speed of electronic-based detection, one can achieve sequencing rates of 1.2 Mb/hour with 0.1% error per base without accounting for any parallelism or preparation time. This rate can be achieved with only 10 kHz sampling rate [38], given that about 30 measurements are needed per base (derived using data from [18]). An increase of sampling rate to 1 MHz would achieve a sequencing rate of 120 Mb/hour with the same error. Finally, increasing the error

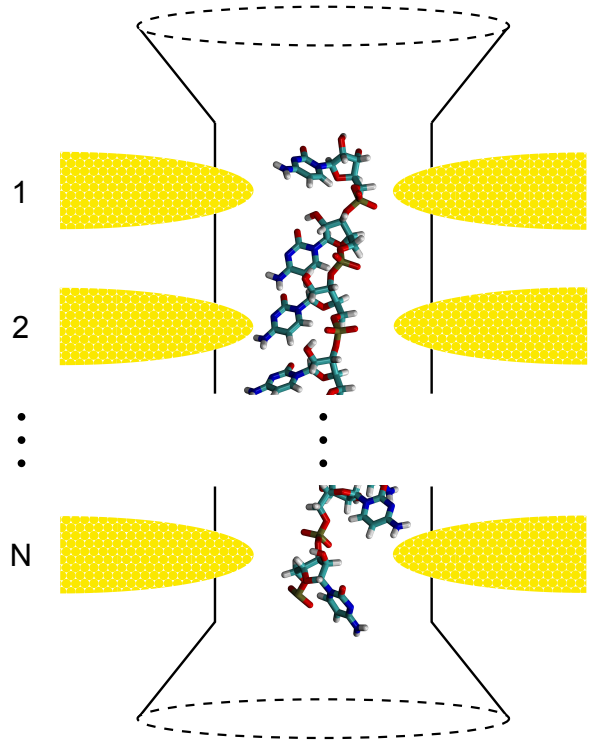

Figure 3.1: (Color online) Schematic for the multiplexed transverse electronic sequencing device. The solid-state nanopore is outlined in black with the dashed lines representing the conical entrance/exit that leads to the cylindrical nanochannel. Within the nanochannel is the ssDNA strand to be sequenced and several pairs of embedded gold electrodes labeled 1, 2, ... N to indicate the existence of a series of N pairs of electrodes. Each pair of electrodes would be attached to a voltage source so that the DNA bases align with the field and the tunneling current flows stronger. In addition, a single pair of biased electrodes would be placed diametrically opposite above and below the pore to push/pull the negatively charged ssDNA strand through the nanochannel.

by an order of magnitude would only slightly decrease the sequencing rate [13]. Since this nanopore method does not require the ssDNA strand to be of a certain length to function, the read length depends solely on the sequencing device's bandwidth and its ability to keep the ssDNA strand untangled and consistently translocating through the pore. In addition, as a label-free method, the technique benefits from a modest preparation time and reduced operating costs.

On the other hand, due to the speed of translocation of the ssDNA and the

linear width of a single nucleotide, roughly 6.3 Å [39], the current through each nucleotide has to be measured in a short period of time with limited bandwidth. Experiments have found the translocation speed to be difficult to control [38, 40], yet the gate modulation of nanopore surface charges promises to reduce this speed and add an element of control to the instantaneous velocity of the ss-DNA strand [41]. With few current measurements per base, it becomes hard to identify the sequence of the ss-DNA strand without substantial errors. Therefore, if bandwidth is an issue, we suggest the use of a nanochannel containing several pairs of electrodes in series like in a multiplexing configuration, as shown in Fig. 3.1. We show that the signals from each pair of electrodes can be cross-correlated to significantly reduce noise and consequently reduce errors in base identification. To prove this point we have analyzed the cross-correlations of many ssDNA translocation realizations, finding that with a limited bandwidth already two pairs of electrodes far surpass the sequencing capability offered by a single pair. The approach we propose expands upon the recent work by Ahmed *et al.* [42], where the multiple electrode current readout was considered for the case of a multilayer graphene nanopore [43]. Here, we use the molecular dynamics simulations to characterize the noise along with a different cross-correlation analysis to estimate the signal to noise improvements on the multiple contact readout of solid-state nanochannels.

In experiments, the signal from electrons tunneling through a single nucleotide of ssDNA switches between a high average current state to a low average background current state in a pulse-like manner [38, 40, 44]. Short episodes of background current occur because of the changing adsorption between the DNA base and the electrodes while long episodes are explained by the absence of a DNA base. Using current

thresholds and the time spent in each background current episode one can mark the beginning and end of each nucleotide in the time series. With this method the jth nucleotide that travels through the first electrode pair can be matched with the jth nucleotide that travels through the following electrode pairs for cross-correlation. After the current time series from the ith electrode pair for the jth nucleotide is isolated, the short episodes of background current can be removed to leave only the pulses of current indicative of tunneling through the jth nucleotide of ssDNA. We define the resultant signal as $\mathcal{I}_i^j$.

3.2 Molecular Dynamics Methods

To simulate this process, we first use a combination of molecular dynamics (MD) performed with NAMD2 [8] and quantum transport calculations to obtain a current time series from a single electrode pair for each of the four bases: adenine (A), cytosine (C), guanine (G), and thymine (T). The contributions from neighboring nucleotides to the current have been found to be negligible provided the electrode cross-section is on the order of 1 nm [12]. The MD results we use here have been taken from previous work in [18] where the simulation proceeds as follows. A double-conical Si_3N_4 nanopore with embedded gold electrodes in the center is built with a minimum diameter of 1.4 nm and a maximum diameter of 2.5 nm (similar to Fig. 3.1 with just one electrode pair). The inner diameter is such that the homogeneous ssDNA can just pass through so that the electrode spacing can be at a minimum to enhance the signal. The ssDNA is placed parallel to the longitudinal axis so that the first base has past the entrance of the pore. The pore-DNA system is solvated in a TIP3P water sphere and constrained with periodic boundary conditions in an NVT ensemble with

a 1 M solution of K^+ and Cl^-. The system is evolved in time with 1 fs steps and kept at room temperature with Langevin damping. To drive the ssDNA through the pore within a feasible simulation time a global longitudinal electric field of 6 kcal/(mol Å e) is applied. When a base of ssDNA sits in between the electrodes the longitudinal pulling field is turned off and a transverse field of the same magnitude is turned on to calculate the electronic transport. This is an approximation to the transverse field being much larger than the longitudinal field, which is the optimum operating regime for the present sequencing device as the bases are better aligned with the transverse field [13].

The current is calculated with a single-particle elastic scattering approach using a tight-binding Hamiltonian [30]. Coordinate snapshots of the molecular dynamics are taken every ps, with which a tight-binding Hamiltonian is created for the region between the gold electrodes. The Fermi level is taken to be that of bulk gold. To obtain the tunneling current through the ssDNA, we use the single-particle retarded Green's function,

$$G_{\mathrm{DNA}}(E) = \frac{1}{ES_{\mathrm{DNA}} - H_{\mathrm{DNA}} - \Sigma_t - \Sigma_b}, \tag{3.1}$$

where E is the energy, S_{DNA} and H_{DNA} are the overlap and Hamiltonian matrices, respectively, of the electronic junction, and Σ_t and Σ_b are the top and bottom electrode self-energies, respectively, for the interaction with the junction contents. The Green's function for gold needed to calculate Σ_t and Σ_b is approximated as in [45]. The transmission function is obtained from the Green's function and the self-energies in the usual way (see, e.g., [30]). The current is then given by

$$I = \frac{2e}{h} \int_{-\infty}^{\infty} dE\, T(E)[f_t(E) - f_b(E)], \tag{3.2}$$

where e is the elementary charge, h is Planck's constant, E is the energy of the scattering electron, T is the total transmission function, and f_t and f_b are the top and bottom electrode Fermi-Dirac distribution functions, respectively [30]. This process is carried out for every snapshot to obtain a time series for each of the four bases.

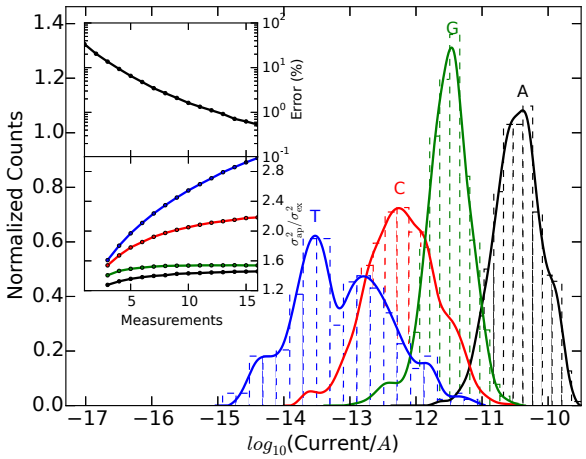

Figure 3.2: (Color online) Normalized current distributions from [18] for the four bases, A, C, G, T, with one pair of electrodes, where the solid lines are cubic spline mirror-symmetric interpolations of the dashed line histograms. The distributions describe the probability of the base-10 log of the current due to the multi-scale nature of tunneling currents. The upper inset plots the sequencing error percentage per base on a log scale against the number of measurements per base, m. The lower inset plots the Fenton-Wilkinson approximated variance (see Eq. (3.12)) for $\sigma^2_{N=2,m}$ divided by the exact variance of $\log(g^j_{N=2,m})$ against m for $j = A, C, G, T$, where the color of the line corresponds to the base whose distribution has the same color.

The points in the time series are, to a good approximation, independent since the time for electrons to tunnel ($\sim 10^{-15}$ s) is much smaller than the time between each snapshot recording (10^{-12} s). This coincides with experiments where we expect each point in $\mathcal{I}^j_i$ to be effectively independent since the time scale governing the molecular disorder that modulates the current (the fastest being water at $\sim 10^{-12}$ s [29]) is much smaller than the typical time scale of measurement ($>10^{-6}$ s or a kHz sampling rate [38, 40, 44]). As a result, we do not expect the cross correlations to cut out these fast

noise time scales in the current, rather the slowly propagating modes.

A probability distribution for the current values is created for each of the four bases by binning each respective time series as seen in Fig. 3.2. From these probability distributions we construct a set of time series, $\{I_i^j\}$, that resemble the signals generated by a ssDNA passing through a nanochannel with N pairs of electrodes, or $\{\mathcal{I}_i^j\}$. Each I_i^j is the tunneling current time series from the ith electrode pair and the jth nucleotide in the ssDNA that is centered around $t = 0$ for convenience. Given that the spacing between opposing electrodes is roughly equivalent from electrode pair to electrode pair along the nanochannel, the pore-electrode environment would be nearly identical in each case.

3.3 Cross-correlations

Due to the independence of $\mathcal{I}_i^j$ and I_i^j we can use a Monte Carlo method in which numbers are generated from a uniform distribution and then matched to a current value in the cumulative distribution function (cdf) for the jth nucleotide to create the set of $\{I_i^j\}$. In addition, we can use a cyclic cross-correlation to maintain a constant overlap length for any set of time shifts. This is achieved by creating a periodic summation for each I_i^j defined as

$$\tilde{I}_i^j(t) = \sum_{k=-\infty}^{\infty} I_i^j(t - kT_i^j), \tag{3.3}$$

where T_i^j is the length of time I_i^j elapses. We then cross-correlate the N time series for each jth nucleotide together using

$$g_N^j(\tau_1, \ldots, \tau_{s-1}, \tau_{s+1}, \ldots, \tau_N) = \frac{1}{T_s^j} \int_{-\infty}^{\infty} dt\, I_s^j(t) \prod_{i \neq s}^{N} \tilde{I}_i^j(t + \tau_i), \tag{3.4}$$

to obtain a single function.

The function g_N^j is the N-point cross-correlation, while τ_i is the time shift of the ith electrode pair. I_s^j is the time series for the jth nucleotide with the smallest length of time, T_s^j. We choose all but I_s^j to be periodically extended so that no overlapping current values are included more than once within any $\tilde{I}_i^j$. By dividing by T_s^j the cross-correlation values are normalized to be independent of the time overlap.

Due to the nature of the probability distributions for the tunneling currents (see Fig. 3.2), g_N^j covers several orders of magnitude and thus is best portrayed as $\log(g_N^j)$, where the log is taken as base 10. We then bin cross-correlation values over the set of $\{\tau_i\}$ such that each distinct point $\log(g_N^j(\tau_1, \ldots, \tau_{s-1}, \tau_{s+1}, \ldots, \tau_N))$ with $\tau_i \in (-T_i^j/2,\, T_i^j/2]$ is a dimension of the histogram (i.e., including only one period for each τ_i). On the basis of how we have constructed the set $\{I_i^j\}$ using the properties of statistical independence, we can treat each point in g_N^j, and consequently $\log(g_N^j)$, as following the same probability distribution. As a result, the joint probability distribution is symmetric over the exchange of any two dimensions. However, because of the built-in correlation between each point of the cross-correlation g_N^j, the joint probability distribution is not purely isotropic and none of the dimensions may be traced out.

For ease of computation we build each I_i^j to have equal length ($T_i^j = T$) and uniform spacing (Δt) implying that the number of measurements taken at each

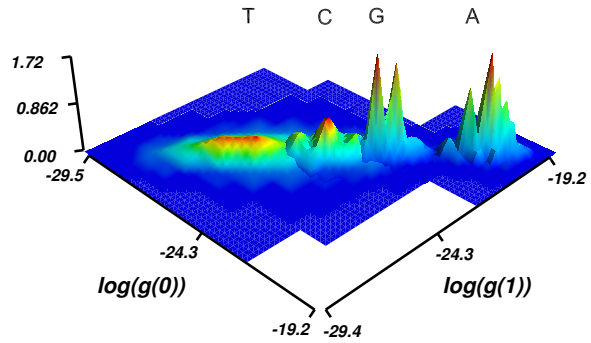

Figure 3.3: (Color online) Normalized joint distributions, $z = P^j_{2,2}(\log(g^j_{2,2}))$, for $j = A, C, G, T$. Since $g^j_{2,2}$ only has $d = 2$ distinct points there are only 2 independent dimensions in the joint distributions. These joint distributions are linear interpolations of the original histograms. The color is only used to further illustrate changes in the z-axis and does not represent the same z values across different distributions.

electrode pair, $m = T/\Delta t$, is the same for each nucleotide. Since the order of the nucleotides does not affect the outcome we just need to compute g^j_N for $j = A, C, G, T$ to understand how cross-correlating the time series from all electrode pairs together affects the distinguishability of the four DNA bases. However, to gain this understanding we must construct the set $\{I^j_i\}$ for N electrode pairs with a certain m value and compute $g^{A,C,G,T}_N$ many times so that we have a large pool of cross-correlations to interpret and histogram. In this case g^j_N would have $d = m^{N-1}$ distinct points, meaning that the joint probability distribution for $\log(g^j_N)$ would be d-dimensional. For reference purposes we add the number of measurements per electrode pair, m, as an index to the cross-correlation function, now $g^j_{N,m}$, and define $g^j_{N,m}(k)$, $k \in [0,\, d-1]$ as the kth point of the cross-correlation function, essentially flattening the set $\{\tau_i\}$ to one index k. After creating the histogram for $\log(g^j_{N,m})$ we linearly interpolate it to obtain the continuous joint probability distribution $P^j_{N,m}(\log(g^j_{N,m}))$, as seen in Fig. 3.3.

With $P^j_{N,m}$ for $j = A, C, G, T$ determined with a given number of pairs of elec-

trodes, N, and measurements per pair, m, we can now compute the distinguishability of the DNA bases. To do this we calculate the average probability of incorrectly determining the identity of a DNA base given a set of tunneling current time series, $\{I_i^j\}$, from the corresponding nucleotide. This can be expressed by the following equation as

$$\begin{aligned} e_{N,m}^X &= \left\langle \frac{\sum_{j\neq X} \tilde{P}_{N,m}^j(g_{N,m}^X)}{\sum_{j=A,C,G,T} \tilde{P}_{N,m}^j(g_{N,m}^X)} \right\rangle_{g_{N,m}^X} \\ &= \left\langle \frac{\sum_{j\neq X} P_{N,m}^j(\log(g_{N,m}^X))}{\sum_{j=A,C,G,T} P_{N,m}^j(\log(g_{N,m}^X))} \right\rangle_{g_{N,m}^X}, \end{aligned} \tag{3.5}$$

where $e_{N,m}^X$ is the error probability of choosing base X correctly with N pairs of electrodes and m measurements per pair while $\tilde{P}_{N,m}^j$ is the probability distribution for $g_{N,m}^j$ instead of $\log(g_{N,m}^j)$. The average is an ensemble average taken over all possible cross-correlation functions for base X. Then we average $e_{N,m}^X$ over all of the DNA bases, $X = A, C, G, T$, to obtain the average error probability per base to sequence DNA:

$$E_{N,m} = \frac{1}{4} \sum_{X=A,C,G,T} e_{N,m}^X. \tag{3.6}$$

3.4 Results and Discussion

With a collection of error probabilities for different values of m and N we can now evaluate the efficacy of this multiplexing technique. We have calculated $E_{N,m}$ for $N = 2, m = 2 - 9$ and $N = 3, m = 2 - 3$, as illustrated in Fig. 3.4. For both $N = 2$ and $N = 3$, $E_{N,m}$ decreases linearly with increasing m on a logarithmic scale, meaning

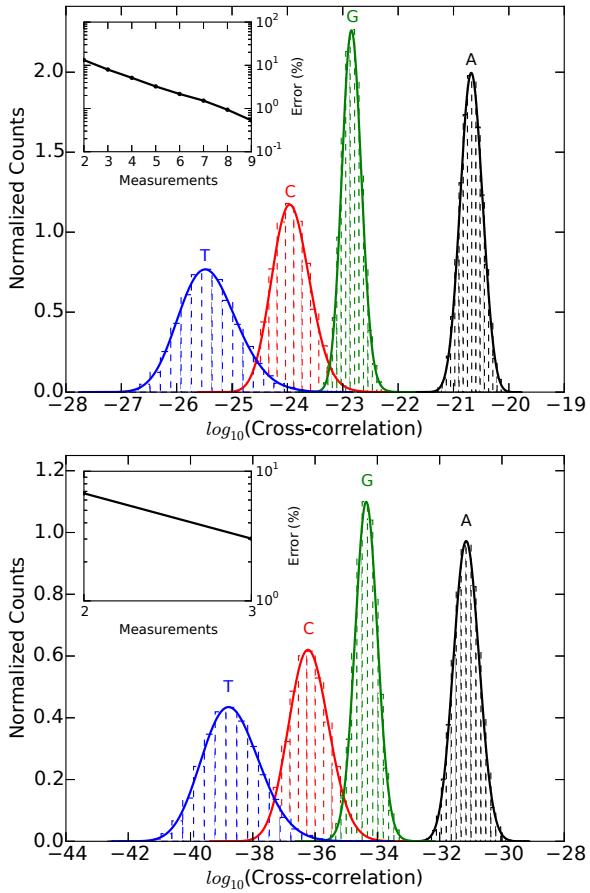

Figure 3.4: (Color online) Normalized distributions for $\log(g^j_{2,9}(0))$ (top) and $\log(g^j_{3,3}(0))$ (bottom) for $j = A, C, G, T$, where the solid lines are cubic spline mirror-symmetric interpolations of the dashed line histograms. The insets plot the sequencing error percentage per base for $N = 2$ (top) and $N = 3$ (bottom) on a log scale against the number of measurements per base per electrode pair, m.

$E_{N,m} \sim \beta e^{-am}$ where β and a are positive constants. Due to limited error data for $N = 3$, we compared means and variances to confirm this general trend. Compared to the sequencing error with a single pair of electrodes ($E_{N=1,m}$), which is also linear with m on a log scale (see the upper inset of Fig. 3.2), $E_{N=2,m}$ and $E_{N=3,m}$ have nearly double and triple, respectively, the linear rate of decline. Because of the exponential

relationship with m, we can generously claim

$$E_{N=2,m} \sim \beta(E_{N=1,m}/\beta)^2, \tag{3.7}$$

and

$$E_{N=3,m} \sim \beta(E_{N=1,m}/\beta)^3. \tag{3.8}$$

Therefore the improvement in identification errors is significant. In fact, more generally we can assume

$$E_{N,m} \sim \beta(E_{1,m}/\beta)^N. \tag{3.9}$$

This result can be easily justified. If the cross-correlation of the N current signals for each base j, $\{I_i^j\}$, from an N electrode pair system did not lose any of the information contained in the original signals, then Eq. (3.9) would not be generous at all but instead nearly exact. However, a cross-correlation of two different signals certainly results in a loss of information, which manifests itself in the sequencing error by decreasing the exponent N by some factor α representing the fraction of information that was preserved. In other words, the original $\tilde{N}$ signals contain $\tilde{N}\tilde{m}$ points of information, but when cross-correlated what remains is some fraction of that, $\alpha\tilde{N}\tilde{m}$, which results in a more accurate relation between $E_{N,m}$ and $E_{1,m}$,

$$E_{\tilde{N},\tilde{m}} \sim E_{N=1,m=\alpha\tilde{N}\tilde{m}} \sim \beta(E_{N=1,m=\tilde{m}}/\beta)^{\alpha\tilde{N}}. \tag{3.10}$$

By calculating the slope of each line, $\log E_{N,m}$ against m for $N = 2, 3$, with a linear regression we obtain $\alpha = 0.83$ for $N = 2$ and $\alpha = 1.00$ for $N = 3$. This suggests that α saturates to 1 as N increases since with a higher N comes a better chance to

reconstruct the original signals from the cross-correlation.

On inspection of Eq. (3.5), one should notice that $e^X_{N,m}$ only depends on the probabilities, $P^j_{N,m}(\log(g^X_{N,m}))$ with $j = A, C, G, T$, and not explicitly on N or m. As a result, for the error to decrease as it does for $N = 2, 3$ the joint probability distributions for all 4 bases, $P^j_{N,m}$ where $j = A, C, G, T$, must grow farther and farther apart as N or m is increased to reduce their overlap. This is indeed the case and we can study the degree to which the distributions are separated by analyzing the moments of the distributions. Since analyzing the form of the joint distributions, as in Fig. 3.3, becomes too difficult as the number of dimensions, $d = m^{N-1}$, is increased, we settle with analyzing the probability distributions for a single point of the cross-correlation function (e.g., Fig. 3.4).

Because the distributions in Fig. 3.4 have only one independent variable, $\log(g^j_{2,9}(0))$ for the top and $\log(g^j_{3,3}(0))$ for the bottom, they are fairly smooth due to the integration over all of the other dimensions of the joint distribution. The distributions in Fig. 3.4 are well approximated by normal distributions, which makes the distributions for $g^j_{2,9}(0)$, $g^j_{3,3}(0)$, and generally any other single point of $g^j_{N,m}$ for any N and m, approximately log-normal by definition.

A log-normal random variable, Y, is best characterized by the mean, $\hat{\mu}$, and variance, $\hat{\sigma}^2$, of $\ln Y$, which follows a normal distribution. $\log Y$ is related to $\ln Y$ with a mean of $\mu = \hat{\mu}/c$ and a variance of $\sigma^2 = \hat{\sigma}^2/c^2$, where $c = \ln 10$. We can also approximate the original distributions for $\log I^j_i$ in Fig. 3.2 as normal, making the distributions for I^j_i approximately log-normal as well.

If we then examine the discrete form of Eq. (3.4) we find that the unshifted

point of the cross-correlation function, labeled $g^j_{N,m}(p)$, can be written as

$$g^j_{N,m}(p) = \frac{1}{m}\sum_{k=0}^{m-1}\prod_{i=1}^{N}\bar{I}^j_i(k), \tag{3.11}$$

where $\bar{I}^j_i$ is the discrete form of I^j_i indexed by measurement number. Any other point in $g^j_{N,m}$ is a similar sum of products except that the set of discrete currents has been shifted. The product of any number of log-normal random variables is also log-normal, with its mean and variance parameters defined as the addition of the mean and variance parameters of the random variables that went into the product. Since, for a given base j and any index k, every pair of electrodes' current value, $\bar{I}^j_i(k)$, follows the same probability distribution, the mean and variance parameters for $\prod_{i=1}^{N}\bar{I}^j_i(k)$ are simply $N\hat{\mu}_1$ and $N\hat{\sigma}^2_1$, respectively. Here, $\hat{\mu}_1$ is the mean of the natural log of the tunneling current with 1 pair of electrodes while $\hat{\sigma}^2_1$ is the variance whereas μ_1 and σ^2_1 would be the mean and variance of the base 10 log of the tunneling current, as in Fig. 3.2. Recalling the properties of independence built-in to the set of $\{I^j_i\}$, we know that each product in the summation is independent. Therefore we can use the Fenton-Wilkinson approximation, [46], to obtain the mean and variance of $\log(g^j_{N,m}(0))$ (exactly depicted in Fig. 3.4) from μ_1 and σ^2_1,

$$\begin{aligned}\sigma^2_{N,m} = \frac{\hat{\sigma}^2_{N,m}}{c^2} &= \frac{\ln[1+(e^{N\hat{\sigma}^2_1}-1)/m]}{c^2}\\ &= \frac{\ln[1+(e^{Nc^2\sigma^2_1}-1)/m]}{c^2},\end{aligned} \tag{3.12}$$

$$\begin{aligned}\mu_{N,m} = \frac{\hat{\mu}_{N,m}}{c} &= \frac{N\hat{\mu}_1 + N\hat{\sigma}^2_1/2 - \hat{\sigma}^2_{N,m}/2}{c}\\ &= N\mu_1 + cN\sigma^2_1/2 - c\sigma^2_{N,m}/2,\end{aligned} \tag{3.13}$$

where $\sigma^2_{N,m}$ and $\mu_{N,m}$ are the variance and mean of $\log(g^j_{N,m}(0))$ while $\hat{\sigma}^2_{N,m}$ and $\hat{\mu}_{N,m}$ are the variance and mean of $\ln(g^j_{N,m}(0))$, for a certain value of j. The Fenton-Wilkinson approximation assumes that the sum of log-normal random variables is also log-normal, which is not exact, and then derives the mean and variance parameters by moment matching [46].

$\mu_{N,m}$ changes dramatically with N, but not much with m. Therefore as m is increased with a fixed N, it is mostly the change in $\sigma^2_{N,m}$ that is responsible for the reduced overlap between the cross-correlation distributions and consequently the reduced sequencing error, $E_{N,m}$. While the mean of $\log(g^j_{N=2,m})$ coincides almost exactly with $\mu_{N=2,m}$, the variance of $\log(g^j_{N=2,m})$ can differ from $\sigma^2_{N=2,m}$. In the lower inset of Fig. 3.2 we plot $\sigma^2_{N=2,m}$ divided by the exact variance of $\log(g^j_{N=2,m})$ against m for $j = A, C, G, T$ to evaluate the performance of the Fenton-Wilkinson approximation. We can see that all four lines seem to be asymptotically approaching some maximum correction factor. The variance for adenine and guanine is fairly well represented by the approximation, explained by the fact that the $\log(\text{Current}/A)$ distributions for those two bases are closest to resembling normal distributions. Thymine's $\log(\text{Current}/A)$ distribution appears to have a bimodal component, which explains why the Fenton-Wilkinson approximation badly represents the variance of $\log(g^T_{N=2,m})$. Nevertheless, the approximation can be used as an analytical upper bound on the exact variance of $\log(g^j_{N=2,m})$ for $j = A, C, G, T$. This variance is an indicator for the sequencing error but it is not sufficient to determine the error alone since the joint distributions are needed.

3.5 Conclusions

An enhancement to the sequencing by tunneling method is proposed, in which N pairs of electrodes are built in series along a synthetic nanochannel. The ssDNA is forced through the channel using a longitudinal field, as in the original method [12, 13, 18], and potentially controlled with gate modulation of nanochannel surface charges [41]. In this manner the strand of ssDNA passes by each pair of electrodes for long enough to gather a minimum of m tunneling current measurements, where m is determined by the level of sequencing error desired. Each current time series for each base, I_i^j, is then cross-correlated together using a cyclic method to balance the resultant function. With these cross-correlations, one may identify the DNA base by referring to cross-correlation probability distributions that would be obtained from a calibration run.

We have shown that indeed the sequencing error is significantly reduced as the number of pairs of electrodes, N, is increased. Compared to the sequencing ability of a single pair of electrodes, cross-correlating N pairs of electrodes is *exponentially* better due to the approximately log-normal nature of the original tunneling current probability distributions. We have also used the Fenton-Wilkinson approximation to analytically describe the mean and variance of the cross-correlations that are used to distinguish the DNA bases. When bandwidth is limited, this sequencing method is useful to allow a smaller electrode gap residence time while still promising to consistently identify the DNA bases correctly.

3.6 Acknowledgements

This work was supported in part by the National Institutes of Health, US DOE, and ERC-DM-321031. A. V. Balatsky acknowledges useful conversations with T. Ahmed, J. Haraldsen, T. Kawai, and M. Taniguchi.

Chapter 3, in full, is a reprint of material as it appears in P. Boynton, A. Balatsky, I. Schuller, and M. Di Ventra, "Improving sequencing by tunneling with multiplexing and cross-correlations", *J. Comput. Electron.*, vol. 13, no. 4, pp. 794-800, 2014. The dissertation author was the primary investigator and author of this publication.

Chapter 4

Sequencing Proteins with Transverse Ionic Transport in Nanochannels

4.1 Introduction

Living organisms depend on proteins to carry out the genetic code and perform many vital cellular tasks like metabolism [47]. To understand how a protein works one must understand its structure. Proteins are special because of how versatile they are in binding to other molecules, and the structure of these binding sites often indicate the precise use of a protein.

The first step in understanding protein structure is knowing the sequence of a protein, meaning the order of the amino acids that compose it. There are 20 amino acids that are used as building blocks by eukaryotic genes to make proteins, all of which have the same chain of atoms as a backbone. What distinguishes each amino

acid is its side chain, which can span from a single hydrogen in the case of glycine (GLY) to containing an indole functional group in the case of tryptophan [47]. For a protein to function these amino acids fold up into secondary and tertiary structures that expose features like binding sites, which can be predicted based on the protein sequence. Ongoing research attempts to understand protein aggregation diseases such as Alzheimer's Disease [48] by performing simulations of structure formation, which would not be possible without the knowledge of the components of the peptides and proteins involved. In addition, protein sequences allow the synthesization of other proteins, which is necessary to compensate for diseases like Diabetes Type I in which the body does not produce the necessary peptide hormone insulin [49, 50].

The most common method for *de novo* protein or peptide sequencing (namely sequencing a protein for the first time) is mass spectrometry, a technique that involves fractionating the peptide into many smaller peptides and then obtaining the mass-to-charge ratio of each new peptide from the mass spectrometer. The problem with this technique is that fractionation is often carried out with gel electrophoresis, which is inherently slow [51]. In addition, fractionation must be repeated many times to obtain small enough peptides so that one can discern the composite amino acids from just the total mass-to-charge ratio [52]. Also, *de novo* sequencing is sometimes impossible with this technique since some amino acids have the same mass and charge (*e.g.*, leucine and isoleucine).

Edman degradation is another common method for *de novo* protein or peptide sequencing that utilizes repeated chemical washing and N-terminal cleaving to identify the sequence of amino acids one at a time [53]. However, Edman degradation suffers from the same issue of fractionation as mass spectrometry since devices can only

reliably sequence peptides up to about 30 amino acids [54]. Nonetheless, the end result of identification via chromatography of each singled out chemically modified amino acid is reliable, albeit slow, but does require the use of many reagents.

The advent of nanopore DNA sequencing [10, 55] has brought several modern techniques to protein detection: longitudinal ionic transport [56, 57] and transverse electronic transport [58]. In the case of ionic transport through a single nanopore, detection of the protein folding state is achieved experimentally and modeled with exclusion volumes by [56]. Of course, protein sequencing with such a technique is a more difficult task and has not been achieved as of yet [57]. In fact, longitudinal ionic transport detects a current blockade which is the convolution of several blockade events from different amino acids [55].

Transverse electronic transport, a technique in which amino acids are detected by a pair of electrodes transverse to peptide translocation, has been shown to be successful in identifying single amino acids and even in differentiating between tyrosine and phosphotyrosine [58], a post-translational modification. However, only 12 of the 20 amino acids were able to be detected by this technique with two different electrode gap distances (0.55 nm and 0.7 nm) since the tunneling current is highly dependent on this gap distance and an amino acid's ability to enter the gap. In other words, a single gap cannot be used for all amino acids.

This brings us to our proposed technique, sequencing proteins with *transverse ionic transport*. Like the two aforementioned techniques, this method is inspired by a DNA sequencing method [59, 60] and does not require reagents or fractionation since these devices do not place a limit on the length of the polypeptide [55], meaning these nanopore techniques have the potential to be much faster. The structure of

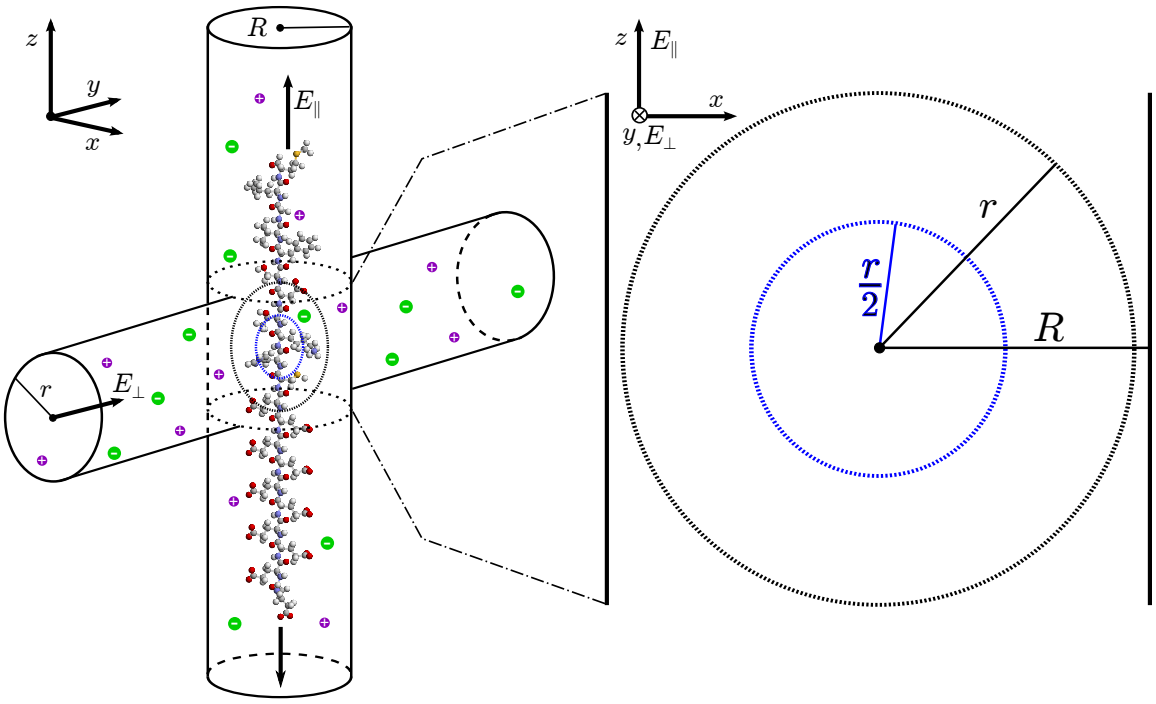

Figure 4.1: (Color online) A schematic of the transverse ionic transport sequencing method. Two nanochannels intersect: the vertical or longitudinal channel along z with radius R and the horizontal or transverse channel along y with radius r. The polypeptide translocates along the longitudinal channel crossing the transverse channel that contains ions, purple K^+ and green Cl^-, that flow along the transverse channel due to an electric field, $E_\perp$, in the $+y$ direction. In this case the polypeptide consists of neurokinin A starting at the C-terminus at the top of the figure attached to one cysteine (CYS) followed by 10 glutamic acids (GLUs), a negatively charged amino acid, where the last GLU makes up the N-terminus (see later in text for more on this structure). This negatively charged polypeptide is driven towards $-z$ by an electric field, $E_\parallel$, in the $+z$ direction. The dotted lines represent the top and bottom extremities of the intersection of the transverse channel, which are expanded to the right along with the thick dashed lines representing the area-limiting cross section (outer black line) and the Monte Carlo radial limit (inner blue line) that lie in the xz-plane. For visibility purposes the polypeptide is enlarged by a factor of 3 in both of its dimensions from the actual scale that we used in simulations while the ion radius is enlarged by a factor of 1.5.

this proposed device is the same as in [59, 60], with a longitudinal nanochannel for polypeptide translocation and an intersecting transverse nanochannel for ionic transport driven by an electric field, $E_\perp$, as in Fig. 4.1. However, the longitudinal

nanochannel must be larger than in [59] to accomodate the various sizes of the amino acids and instead of 4 DNA bases we need to distinguish 20 amino acids. Therefore, the molecular dynamics (MD) simulation method utilized in [59] is time prohibitive for our purposes so we resort to a hard sphere model to account for the electrostatic properties of each amino acid, which requires only one MD run per amino acid to execute. Afterwards we use Monte Carlo sampling to calculate ionic current distributions based on external azimuthal rotations (ϕ') and dihedral angle (ϕ and ψ) distributions, or Ramachandran plots. We show that the distribution of ionic currents for each of the 20 proteinogenic amino acids encoded by eukaryotic genes is indeed statistically distinct, and propose a protocol for *de novo* protein sequencing based on this technique.

4.2 Theoretical Approach

Let us then consider the configuration of crossed nanochannels we have in mind. Although not necessary for our conclusions, we assume for simplicity the nanochannels to have circular cross sections. We will discuss the suggested experimental preparation later in the manuscript.

The polypeptide of interest unfolds inside a nanochannel pulled with a longitudinal force, while it blocks the ionic current flowing in a transverse channel, as schematically shown in Fig. 4.1. We take this longitudinal force to be much less than the transverse force that drives the ions through the transverse nanochannel so that the amino acid resides in the region of nanochannel intersection long enough to obtain the necessary measurements of ionic current for identification. As a result we can assume that the longitudinal ionic flow is negligible when compared to the transverse ionic flow.

It is well understood that the hydration layers surrounding each amino acid have different binding energies [61, 62], which certainly affect the ionic transport transverse to each amino acid. In addition, the amino acid may attract or repel ions due to its solvated charge or polarity state [63, 64]. In order to understand the aqueous environment of each amino acid and determine its effect on the ionic transport, we run MD simulations for each amino acid. We consider the system at normal human body temperature, 310 K, and the solvated system is large enough to make quantum effects negligible. This allows us to use classical molecular dynamics and employ the highly-parallel NAMD2 [8] to run all of our simulations.

The MD setup starts with a single amino acid isolated from a straight (dihedral angles $\psi = \phi = 180°$) peptide chain, as in Fig. 4.1 with proline (PRO) as an exception, which is positioned so that the z-axis is the longitudinal axis. The rest of the MD methods can be found in Appendix A.

The water padding is large enough in this system to examine proximal radial distribution functions (pRDFs) from the amino acid's surface for K^+ and Cl^- up to the point where the concentrations level out to the bulk values. We use the radius from the surface of the amino acid because the features in the concentration will be more prominent as opposed to using the radius from the origin, since the amino acids have irregular shapes. Similarly calculated pRDFs on DNA have been shown to be fairly accurate for reconstructing the surrounding solute even when combining all surface atoms' pRDFs into one [65, 66], as is done in our calculations.

To obtain the pRDFs, we count the number of ions (for K^+ and Cl^-) in 0.5 Å thick shells starting from the surface of each amino acid, which is defined by the intersection of the composing atoms' van der Waals (vdW) spheres. We then calculate

the volume of each shell by subtracting the inner volume of the intersecting spheres from the outer volume, using a grid approximation with 0.1 Å sides for each volume calculation. With the number of ions and the volume of the corresponding shell we calculate the local concentration of K^+ and Cl^- as a function of $r_>$, taken to be the perpendicular distance from the vdW surface to the radial midpoint of the shell, from the first shell at $r_> = 0.25$ Å to the last at $r_> = 44.75$ Å, which is below the 4.8 nm upper bound of water padding.

As can be seen in Fig. 4.2, the concentrations reach a sufficiently steady bulk value at varying radii, with the maximum bulk $r_>$ determined to be approximately 15 Å. Therefore we can focus on the part of the plots pertaining to $r_> \leq 15$ Å to determine the solvation properties of each amino acid. As an example of our numerical procedure, we have chosen to feature the amino acid GLU in Fig. 4.2A, which has a negatively charged side chain at physiological pH (7.4), lysine (LYS) in Fig. 4.2B, which has a positively charged side chain at the same pH, and methionine (MET) in Fig. 4.2C, whose side chain is hydrophobic at this pH. These three amino acids are of similar size, which allows us to better compare the effects of charge states on transverse ionic current. We can immediately notice that the part of the pRDFs that we care about is quite different for each featured amino acid. GLU in Fig. 4.2A has a higher concentration of K^+ due to its negativity while LYS in Fig. 4.2B has a higher concentration of Cl^- due to its positivity. Then there is the hydrophobic MET in Fig. 4.2C, which appropriately repels both K^+ and Cl^- without much preference.

In the setting of an external electric field driving transverse ionic flow around an amino acid within a peptide, the potential barrier that ions must overcome in transport is influenced mostly by the electric potential in the neighborhood of the

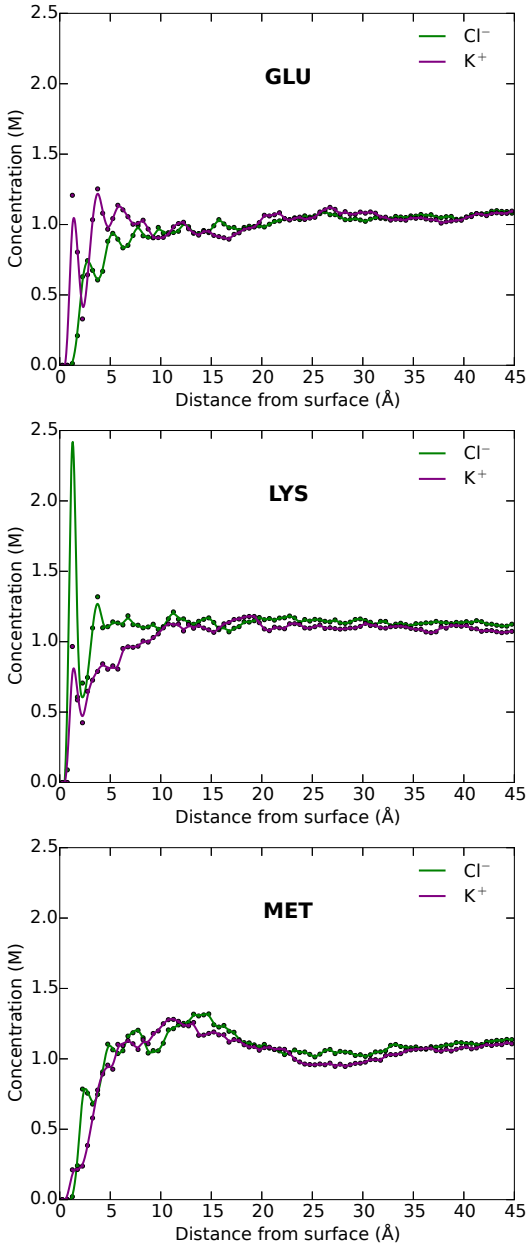

Figure 4.2: (Color online) Plots of ionic concentration against distance from each amino acid's vdW surface, $r_>$, for amino acids GLU, LYS, and MET. K^+ is represented by the purple line and Cl^- is represented by the green line.

area-limiting cross section perpendicular to the ionic flow, imaged in Fig. 4.1 as the black thick-dashed line. This is partly due to the short interference time between the flowing ions and the circumvented amino acid. In our theoretical approach, we treat the equilibrium ionic concentrations as indicators of this electric potential to

develop a hard sphere model with which we can calculate the distribution of ionic current for each amino acid. By calculating an effective radius, r_{eff}, that is applied to every atom in the amino acid beyond its vdW radius, we can sample many amino acid orientations using a Monte Carlo approach to determine all of the ionic current distributions. We theorize that most of the variation in the transverse ionic transport will come from the exclusionary effects of the amino acid with respect to the direction of ionic flow, meaning that a large pool of orientations must be sampled to obtain an accurate view of these distributions.

In order to obtain the effective radius for each amino acid, we start with the definition of the average transverse ionic current of an ionic species i, K^+ or Cl^- in our case, with valency $\tilde{z}_i$ flowing around an amino acid.

$$\begin{aligned}
\langle I_i \rangle &= q\tilde{z}_i \int_{r_i}^{r_f} \int_{\theta_i}^{\theta_f} \langle \tilde{g}_i \tilde{v}_i(r, \theta'(\hat{r}'), \phi'(\hat{r}')) \rangle_{\hat{r}'} r \mathrm{d}\theta \mathrm{d}r \\
&= C \int_{r_i}^{r_f} \langle \tilde{g}_i \tilde{v}_i \rangle_{\hat{r}'}(r) r \mathrm{d}r \\
&= C' \int_{r_i}^{r_f} \langle \tilde{g}_i \tilde{v}_i \rangle_{\hat{r}'}(r) \frac{\tilde{A}(r)}{2r} \mathrm{d}r \\
&\approx C'' \int_{0}^{r_\mathrm{b}} g_i(r_>) v_i(r_>) \frac{A(r_>)}{2(r_> + r_\circ)} \mathrm{d}r_> \\
&= C'' \int_{r_{\text{eff}}}^{r_\mathrm{b}} g_{i,\mathrm{b}} v_{i,\mathrm{b}} \frac{A(r_>)}{2(r_> + r_\circ)} \mathrm{d}r_>
\end{aligned} \tag{4.1}$$

Here, the identifying transverse ionic current, I_i, through the aforementioned area-limiting cross section perpendicular to the ionic flow, is averaged over all rotational orientations equally with $\hat{r}'$ representing the unit $\vec{r'}$ vector of the amino acid while q is the electron charge and C, C', and C'' are constants. $[\theta_i, \theta_f]$ is the window of θ where the amino acid under study has non-negligible influence compared to neighboring amino acids. r_i is the radius where $\tilde{g}_i$, the local number density as a function of

spherical coordinates, is first nonzero. r_f is the radius where the influence of the amino acid is no longer felt in the concentration and thus we need not continue the integral for the purpose of the effective radius, r_{eff}, calculation. $\tilde{v}_i$ is the transverse velocity through the cross section as a function of standard spherical coordinates while $\tilde{A}(r)$ is the surface area of the sphere of radius r. In addition, $r_>$ is the perpendicular distance from the vdW surface to the radial midpoint of a shell of thickness $\mathrm{d}r_>$ and surface area A, $r_\circ$ is the average radius from the origin to the vdW surface, r_b is a value of $r_>$ where the pRDF, g_i (plotted in Fig. 4.2), has become sufficiently steady around the bulk density, $g_{i,\mathrm{b}}$, so as to represent a shell in the bulk, v_i is the flow velocity of the ion species i as a function of $r_>$, while $v_{i,\mathrm{b}}$ is the maximum of v_i, which occurs in the bulk by construction.

The first approximation that we make is that all of the rotational orientations are uniformly likely, when in reality θ' is fairly constant due to the stiffness of the peptide bond and given how small the diameter of the pore is in comparison to the length. However, when we average over ϕ' we fully explore the number density around the shell, so averaging over θ' does not introduce any new data but adds more weight to the side chain as opposed to the ends of the backbone. This counteracts the simplification we make in our MD runs where we use isolated amino acids and include the number density at the ends of the backbone, which would normally be expelled by the nearest neighbor amino acids. Also, the internal dihedrals are assumed fixed since they do not fluctuate much under the imposed longitudinal electric field (see their implementation in the current distribution calculations). Lastly, when we change variables from r to $r_>$ we have to approximate r as $r_> + r_\circ$, which is a minor approximation when considering that all of the other functions in the integral have

well-defined transformations. We can now use the following simplified equation to calculate the effective radius for our hard sphere model for every amino acid and ion species combination:

$$\int_{r_{\text{eff}}}^{r_{\text{b}}} \frac{A(r_>)}{2(r_> + r_\circ)} \mathrm{d}r_> = \int_0^{r_{\text{b}}} \frac{g_i(r_>)}{g_{i,\text{b}}} \frac{v_i(r_>)}{v_{i,\text{b}}} \frac{A(r_>)}{2(r_> + r_\circ)} \mathrm{d}r_>. \tag{4.2}$$

However, this equation requires the ratio of the transverse flow velocity compared to the bulk, and due to the small length scales we can use the Stokes equation, similar to [67]. The details of this calculation can be found in Appendix A. From these calculations we find that $r_{\text{b}} = (R - r_\circ)/2$ and then from our pRDF plots (see Fig. 4.2) we learn that the bulk concentrations start at approximately $r_{\text{b}} \geq 15$ Å. Therefore for our model to work we have to take $R \geq 30 + \max\{r_\circ\} = 34.16$ Å, where the max is over all amino acids, and then in the interest of minimizing the bulk ionic current we choose $R = 35$ Å. We also set the transverse nanochannel radius to the same value for simplicity.

The insets of Fig. 4.3 show the results of our calculations for $v_i/v_{i,\text{b}}$; the top graph represents Cl^- around LYS while the bottom graph shows K^+ around LYS. The other amino acids have similar parabolic forms for $v_i/v_{i,\text{b}}$, but differing r_{b} because of differing $r_\circ$. With $v_i/v_{i,\text{b}}$ calculated for every amino acid we can return to Eq. (4.2) to calculate our effective radii for our hard sphere model. This calculation is shown graphically in Fig. 4.3, where the straight magenta line is the argument (including $\mathrm{d}r_>$ as $\Delta r_> = 0.5$ Å) of the left-hand side of Eq. (4.2), which is the average cross-sectional area that the shell of thickness $\Delta r_>$ at $r_>$ occupies in the plane of interest ($y = 0$). The blue line represents the argument of the right hand side of Eq. (4.2), again including $\mathrm{d}r_>$ as $\Delta r_> = 0.5$ Å without the modulation of the velocity ratio, leaving

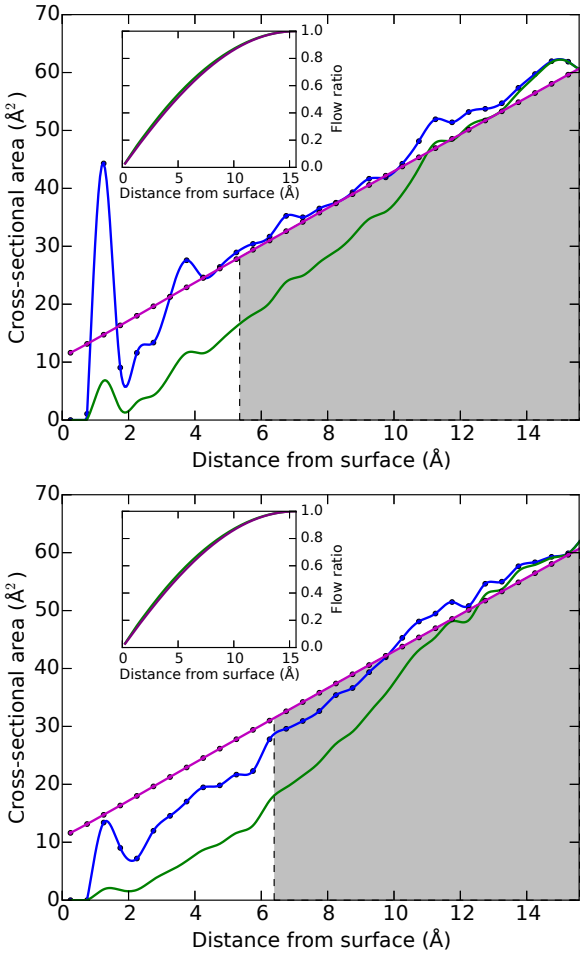

Figure 4.3: (Color online) The top graph represents area plots of Cl^- around LYS while the bottom graph shows area plots of K^+ around LYS. The straight magenta line is the average cross-sectional area that the shell of thickness $\Delta r_> = 0.5$ Å at $r_>$, the distance from the vdW surface of LYS, occupies in the plane $y = 0$. The blue line represents the average cross-sectional area that the ionic solution, with the number of ions from the shell of thickness $\Delta r_> = 0.5$ Å at $r_>$, would occupy in the plane $y = 0$ if those ions were reorganized to have bulk concentration, $g_{i,\mathrm{b}}$. The smooth green curve is the blue curve modulated by the ratio of the velocity with its maximum, $v_i/v_{i,\mathrm{b}}$, which is plotted in the inset of each graph. The area under the smooth green curve is equal to the shaded gray area under the straight magenta curve while the dashed vertical line marks the effective radius for ion species i specifically for LYS.

the average cross-sectional area that the ionic solution, with the number of ions from the shell of thickness $\Delta r_>$ at $r_>$, would occupy in the plane of interest ($y = 0$) if

those ions were reorganized to have concentration $g_{i,\mathrm{b}}$. Finally the smooth green curve is the blue curve modulated by $v_i/v_{i,\mathrm{b}}$. The area under the smooth green curve is equal to the shaded gray area under the straight magenta curve, with the dashed vertical line marking not only where the shaded gray area ends on the left but also the effective radius for ion species i for the given amino acid. Because of the influence of the velocity, the fluctuations in concentration farther from the amino acid have more effect than closely bound spikes. For example, the Cl^- ion atmosphere located 1 Å from the surface of LYS has less effect on the effective radius compared to the next spike in concentration further out from the amino acid, as seen in the green curve. The fact that LYS is positively charged still shows in the effective radii though, with the attractive Cl^- ions having a 5.38 Å addition to the vdW surface compared to 6.43 Å for the repulsive K^+ ions. The rest of the effective radii can be found in Supplementary Table A.1.

This brings us to our Monte Carlo calculation of the transverse ionic current around each amino acid. Now that we have r_{eff} for each ion species that we add to the vdW radius of every atom in our amino acid, we can compare the available cross-sectional area through the $y = 0$ plane and apply the same bulk concentration and estimated bulk velocity, $g_{\mathrm{b}} = 1$ M and v_{b}, to all amino acids to obtain the ionic current values. We do not need to evaluate the available area in the entire cross section though since we only need to calculate up to the largest radius determined by r_{eff} for all amino acids. Therefore we use a radius of $R/2$ from the origin (see Fig. 4.1 where we are now limited to $r = R$) as the circular boundary for all of the amino acids since this circle encloses all of the extended amino acid surfaces in any applicable rotational configuration while also being enclosed by the bulk boundary defined by

$r_> = r_\mathrm{b}$ where the velocity begins to decline from v_b. We also approximate $[\theta_i, \theta_f]$ as $[\pi/4, 3\pi/4]$ by comparing the backbone ends' vdW radius to half of the distance in z between amino acids (half of ideally ~ 3.8 Å [68]). In this manner we can ignore portions of the cross section that would clearly be dominated by neighboring amino acids for the purposes of understanding each amino acid's transverse ionic transport signature.

As previously mentioned, the current becomes sensitive to rotational conformations and dihedral angles in this portion of the calculation. Therefore, instead of assuming uniformity in θ' and straight dihedral angles like we did for the effective radius, we fix θ' to 0 due to the rigidity of the peptide bond and we use Ramachandran plots, [69, 70], to sample realistic values for ϕ and ψ, dihedral angles, which encompass the internal degrees of freedom for a chain of amino acids [70, 71]. That leaves the azimuthal angle, ϕ', which we leave as uniformly distributed since as a whole the peptide does not have an azimuthal preference, except if the peptide is very short in which case the transverse electric field that is only applied to a few amino acids can affect the entire chain. We then apply Monte Carlo to a lone amino acid, the details of which can be found in Appendix A. The reason we use a lone amino acid, the same one from our MD simulations, for calculating the ionic current distributions is that the first step to understanding the viability of this technique is distinguishing each amino acid separately via transverse ionic current. Since most of the exclusion due to the amino acid comes from the region of small z, where the uniqueness of the amino acid is demonstrated, the exclusion from one amino acid in a chain can be derived from our single amino acid distributions. As a result we do not treat the effect of neighboring PRO, which alters the dihedral angles so as to straighten the

polypeptide chain. However, changing an amino acid's dihedrals slightly does not change the ionic current distributions much since most of the variation in the current comes from azimuthal rotation of the amino acid.

Lastly, we must calculate the bulk velocity, v_b, that we will use in the simple equation for the transverse ionic current, $I_i = q\tilde{z}_i g_\mathrm{b} v_\mathrm{b} \langle A_i \rangle$ and $I = \sum_i I_i$, where $\langle A_i \rangle$ is the average area outside of the effective surface from Monte Carlo. This calculation can be found in Appendix A, resulting in $v_\mathrm{b} = 77.23$ m/s.

4.3 Results and Discussion

With a set of ionic currents for each amino acid determined from Monte Carlo utilizing our hard sphere model, we histogram each set of currents and use cubic spline interpolation to arrive at Fig. 4.4. The ionic currents tend to form multimodal (most often bimodal) distributions that are best described as a mixture of several normal distributions. The first and last peaks of each distribution tend to be the highest due to the variation in ϕ'. This is because the ionic current as a function of ϕ' is roughly sinusoidal with a period of π and ϕ' is uniformly distributed, which means the near minimum and near maximum values of the ionic current are chosen the most. Also due to the size of the nanochannels, the ionic current ranges in the tens of nA, which is well within the range of modern measurement devices that can resolve pA currents [72, 60]. Beyond that, this ionic current only represents up to $R/2$ of the whole cross section. By using the parabolic $\hat{v}$ from the bulk region we calculate the contribution from the rest of the cross section, $r_> > r_\mathrm{b}$ but still within the θ limitations, as 69.86 nA after correcting the velocity for experiment. This value is comparable to the ionic current values from Fig. 4.4, meaning the distinctive component of the ionic current

will not be dwarfed by the bulk in an experimental setting.

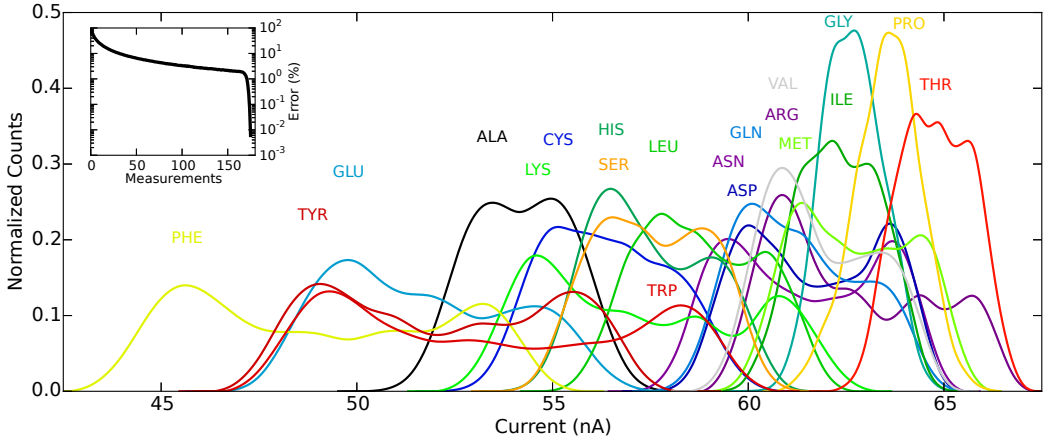

Figure 4.4: (Color online) The transverse ionic current distributions for all 20 proteinogenic amino acids encoded by eukaryotic genes (identified with their standard three-letter abbreviations). The distributions have been normalized to the current values in nA. The inset plots the average error percentage over all 20 amino acids of identifying an amino acid correctly using M current measurements from that amino acid where the error percentage is on a log scale.

Although a fair number of amino acids do not deviate much from their vdW size identity, namely PRO remaining on the smaller side (large current) and phenylalanine on the larger side (small current), many more (*e.g.*, alanine) have shifted due to their interaction with the ions. However, the vdW volume does remain strongly relevant in the standard deviation of the distributions, where the larger amino acids (arginine, phenylalanine, tryptophan, tyrosine) find more variation in ionic current as the dihedrals or ϕ' are altered.

At a glance there is significant overlap between all of the distributions, yet the graph seems crowded mostly because of the sheer amount of plots to compare. We quantify the distinguishability of the ionic current distributions by calculating the error in selecting the correct amino acid, X, given M measurements from X. Based

on the maximum likelihood decision rule [73], the error is defined by

$$e_m^X = 1 - \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{19}\sum_{Y\neq X}^{\{Y\}} H\left(\prod_{m=1}^{M} P^X(I_{m,j}^X) - \prod_{m=1}^{M} P^Y(I_{m,j}^X)\right)\right], \tag{4.3}$$

where J is the total number of realizations of the error calculation, $\{Y\}$ is the set of all 20 amino acids, H is the Heaviside step function, and $P^Y(I_{m,j}^X)$ is the probability of $I_{m,j}^X$, the jth realization of the mth ionic current measurement sampled from the current distribution for X, in Y's ionic current distribution. Here, we assume that each measurement of ionic current is approximately independent. Next we average over X to obtain $\langle e_m^X\rangle_X$ and then multiply by 100 to get the error percentage, which is plotted in the inset of Fig. 4.4. The error drops at a moderate rate with increasing M, but significantly drops off for $M > 160$ when the likelihood of at least one measurement giving zero probability to incorrect amino acids becomes very likely, making the product of those incorrect probabilities zero. For instance, at $M = 175$ the error percentage is practically 0%, and certainly less than 0.1%, a reasonable level of error. With a measurement frequency of 100 kHz, [72], and a best case scenario of 175 measurements per residue without any lapses in between, the sequencing rate becomes 571 residues per second.

4.4 Sequencing Protocol

To build a nanofluidic device with intersecting channels as we suggest one may employ focused ion beam milling, as achieved in [60] with two 10 nm diameter intersecting nanochannels. Our model requires two 7 nm diameter intersecting nanochannels, which is certainly achievable given that [74] has shown non-intersecting

sub-5 nm nanochannels from the focused ion beam milling technique. Although we have predicted that all 20 amino acids are statistically distinct within the framework of circular channels, other cross sections like rectangles or ellipses for the transverse channel allow fewer amino acids to blockade the ionic transport but still provide enough space for ions to flow past the translocating polypeptide. This results in improved residue selectivity and therefore decreased error as well as reduced post-processing time for deconvolution of the amino acid signals, which is necessary if more than one amino acid resides in the nanochannel intersection. Since the source of the distinguishability of the amino acids is their structural and electronic uniqueness we can assume that using a rectangular or elliptical transverse cross section with enough space along x for ionic flow would also result in 20 statistically distinct amino acids.

Once the sequencing device is built with transverse electrodes to control ionic flow, the protein or polypeptide of choice must be unfolded to translocate it through the longitudinal nanochannel. By using a high enough pulling force, around 250 pN [75, 68] that we also apply to our model, the polypeptide will unfold as well as translocate through the nanochannel. As opposed to chemical denaturing, force unfolding results in more confined and reliable Ramachandran plots [75, 68], which directly translates to more reliable ionic current distributions. After the polypeptide is unfolded the pulling force can be adjusted according to one's ionic current measurement frequency and desired rate of error. For example, a desired 0.1% or less of error requires $M = 175$ and with a sequencing rate of 100 kHz as before, the maximum pulling speed would be 217 nm/s assuming an amino acid length of 3.8 Å. As a result, the maximum applicable pulling force would be ~ 180 pN [75].

The next issue is then how this polypeptide is pulled through the nanochannel.

As we have discussed, amino acids have varied charge states in solution. Therefore, to utilize an electric field for pulling (see Fig. 4.1) one has to attach charges to the polypeptide. These charges must be attached at the end of the chain so that one does not interfere with the ionic transport signatures of each amino acid. The best way to achieve this is by using a combination of solid phase peptide synthesis (SPPS), which excels at synthesizing smaller peptides [76], and native chemical ligation (NCL) [77] to attach a sequence of charged amino acids to the N-terminus of the polypeptide under study. We choose GLU as our charged amino acid because of how easily differentiable it is from the other amino acids (see Fig. 4.4) and how easy it is to produce. Using Fmoc, 9-fluorenylmethyloxycarbonyl or the chemical group that protects the N-terminus from reactions until desired, SPPS starting with N,N-bis(2-mercaptoethyl)-amide (BMEA) [78] one creates a sequence of GLU with a length that will give the polypeptide chain plus GLU sequence a large enough charge to pull with an electric field. Fmoc SPPS is also used to attach a CYS residue to the N-terminus of the unknown polypeptide with a polyethylene glycol (PEG) support [79]. Then one uses NCL to take advantage of the transthioesterification reaction to form a native amide bond between the N-terminal CYS residue and the thioester precursor BMEA [78].

Another option is to use optical tweezers [80, 81] to target a terminal amino acid to pull the whole polypeptide. This approach has been utilized for longitudinal nanopore DNA sequencing [82, 83], resulting in more control over translocation due to the high tunability of optical tweezers. Advances in optical tweezers further allow a single beam to trap multiple targets [84], potentially with computer-generated holograms [85], which would allow even more control over the entire polypeptide.

4.5 Conclusions

We have proposed a novel *de novo* protein sequencing method in which an unfolded protein confined to a nanochannel is probed by transverse ionic transport through an intersecting nanochannel. This method promises to offer improved discrimination between amino acids by utilizing the 3-dimensional structure and electronic properties of each amino acid, as compared to techniques like mass spectrometry that can only probe total mass and charge [52]. We developed a hard sphere model for transverse ionic transport that employs the average equilibrium ionic concentrations surrounding all 20 amino acids derived from MD and ionic flow ratios determined by the Stokes equation. With this hard sphere model we were able to calculate distributions of ionic current for each amino acid based on Monte Carlo sampling of internal and external rotational conformations. All 20 amino acids were found to be statistically distinct and a sequencing error rate per residue of less than 0.1% was obtained with $M = 175$ measurements per amino acid, implying a best case scenario of 571 residues per second with a measurement frequency of 100 kHz [72].

This approach is certainly experimentally achievable since 10 nm diameter intersecting nanochannels have been demonstrated for the purpose of DNA sequencing [60] and polypeptides can be pulled through the nanochannel with optical tweezers or by adding charged residues to the polypeptide terminus and employing an electric field. Protein sequencing is very important since DNA sequencing cannot predict post-translational modifications and the ability to identify the sequence of a protein leads to the ability to understand its structure, which is the key to understanding many crippling diseases like Alzheimer's [48]. We therefore hope our work will motivate the experimental realization of the proposed protein sequencing protocol.

4.6 Acknowledgements

We thank Jonas Pedersen and Antti-Pekka Jauho for useful discussions on error calculation and for bringing the Maximum Likelihood method to our attention.

Chapter 4, in full, is a reprint of material as it appears in P. Boynton and M. Di Ventra, "Sequencing proteins with transverse ionic transport in nanochannels", 2015 (under peer-review). arXiv:1509.04772 [physics.bio-ph]. The dissertation author was the primary investigator and author of this publication.

Appendix A

Supplementary Information for Chapter 4

A.1 Methods

A.1.1 Molecular Dynamics

The amino acid is centered along the z-axis according to the geometric center in z of its terminal N and C atoms, while the molecule is centered in the xy-plane according to the geometric center in x and y of its terminal N atom and a nearest neighboring amino acid's terminal N atom. To fix the rotation angle between amino acids, as a convention, the terminal N atoms always have $y = 0$, as is the case in Fig. 4.1.

Since PRO has more rigid dihedral angles, we need to center it with the help of two neighboring GLYs, which have flexible dihedral angles, on each side of a single PRO. The nearest neighbor GLYs are configured to have dihedral angles that

compensate for those of PRO while the farthest neighbor GLYs are configured to be straight, so that PRO and the two straight GLYs are directed along the longitudinal axis while the two straight GLYs are aligned in the xy-plane. As a result, we can center and then isolate PRO by using the usual centering method on the geometric average of the two straight GLYs, staying consistent with the choice of angles for the rest of the amino acids.

Once the amino acid is isolated, we solvate the system into a right hexagonal prism with regular hexagonal xy-planes having a height of 11 nm and an apothem of 5.9 nm to be used in NAMD2 with periodic boundary conditions in all three dimensions of space. This configuration gives every atom from every amino acid at least 4.8 nm of water padding in the unit cell, or in other words at least 9.6 nm of water between any atom and the closest atom in any neighboring periodic image. We then passivate and ionize the system to about 1 M of KCl, a typical biological solute. The size of the ions will certainly change the average local concentrations near the amino acid, which may then affect the ionic transport. We utilize the CHARMM22 with CMAP force field [86, 87] for all of the amino acid, TIP3P water, and ion interactions. Each amino acid was held fixed throughout the run so that it would not diffuse around and the surrounding solution could equilibrate and be analyzed consistently. After equilibrating at 0 K and progressively ramping up the temperature to 310 K, the system is allowed to evolve in an NPT ensemble first for 1 ns followed by an NVT ensemble for 5 ns, all with 1 fs time steps and 1 ps coordinate recordings. The temperature is held fixed using a Langevin thermostat with a damping coefficient of 5 ps^{-1}. The first ns of the NVT production run is discarded as transient, leaving 4 ns of run time, or 4000 coordinate snapshots, to analyze radial concentration profiles.

A.1.2 Velocity Calculation

Due to the small length scales of the intersecting portion of our nanochannel system, we can use the Stokes equation,

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\mu\frac{\mathrm{d}}{\mathrm{d}x}\hat{v}_i(x)\right) + q\tilde{z}_i\hat{g}_i(x)E_\perp = 0, \tag{A.1}$$

where μ is the dynamic viscosity of the fluid, $E_\perp$ is the external electric field applied in the y direction, $\hat{v}_i$ is the transverse velocity through the cross section, $\hat{g}_i$ is the local number density, and ion-ion interactions are ignored. Here the flow velocity is independent of y due to the fact that the transverse nanochannel length is much larger than the diameter of the longitudinal nanochannel, $2R$, and that $2R$ is comparable to the diameter of the transverse nanochannel (as depicted in Fig. 4.1). Independence from z is similarly due to the longitudinal nanochannel's length being much larger than $2R$ but we must also choose R to be large enough for ions to diffuse along z after they enter the longitudinal channel at $y = \pm R$. This will make any variation along z have negligible impact on the end result of an average v_i over all rotational conformations. In our case we simply use the dynamic viscosity of water ($\mu = 7.5 \times 10^{-4}\,\mathrm{Pa}\cdot\mathrm{s}$), even though the viscosity of water with ions will vary slightly [67], and a reasonable value of $E_\perp = 5 \times 10^8\,\mathrm{N/C}$ taken from [67]. However, we must transform this equation into one that depends on $r_>$ to obtain v_i. Since v_i and g_i are averages over all rotational orientations, the problem is condensed to the region $x > 0$ and $[\theta_i, \theta_f]$. With θ_i and θ_f close enough to $\pi/2$, we can approximate $r_>$ as $x - x_\circ$, where $x_\circ$ is some constant, since at least for the amino acid backbone the contour lines of $r_>$ resemble those of x.

With these approximations we obtain

$$\frac{\mathrm{d}}{\mathrm{d}r_>}\left(\mu\frac{\mathrm{d}}{\mathrm{d}r_>}v_i(r_>)\right)+q\tilde{z}_i g_i(r_>)E_\perp=0. \tag{A.2}$$

This will give us the rough form of $v_i/v_{i,\mathrm{b}}$ between the following boundary conditions:

$$\begin{aligned} &v_i(r_>=0)=0\\ &v_i(r_>=\bar{r}_>=\frac{R-r_\circ}{2})\geq v_i(r_>). \end{aligned} \tag{A.3}$$

$\bar{r}_> = (R-r_\circ)/2$ is approximately halfway between the vdW surface and the longitudinal nanochannel surface and is also our upper bound on $r_>$ as the domain of v_i. To obtain v_i for the entire range necessary for Eq. (4.2), we require that $r_\mathrm{b} = \bar{r}_> = (R-r_\circ)/2$ since $\bar{r}_>$ must be in the bulk as well. From our pRDF plots (see Fig. 4.2) we learn that the bulk concentrations start at approximately $r_> = 15$ Å, meaning that $r_\mathrm{b} \geq 15$ Å. Therefore for our model to work we have to take $R \geq 30 + \max\{r_\circ\} = 34.16$ Å, where the max is over all amino acids, and then in the interest of minimizing the bulk ionic current we choose $R = 35$ Å.

A.1.3 Monte Carlo

The Ramachandran plots that we use in our Monte Carlo calculations account for a 250 pN longitudinal force that is applied to the polypeptide chain (ubiquitin and polyglycine in [69]) to pull it through the nanochannel. The pulling force acts to limit the phase space available to the dihedral angle pair (ϕ, ψ), making the configurations that are close to straight ($\psi = \phi = 180°$) much more appealing [69]. PRO and GLY have significantly different plots from the rest of the amino acids due to how the side

chain of PRO bonds with its own amine nitrogen, part of the amino acid backbone, leading to restricted dihedrals while GLY has a hydrogen instead of a side chain leading to more freedom in the dihedrals. This way, while the rest of the amino acids are described by the Ramachandran plot of ubiquitin, which contains all 20 of the proteinogenic amino acids encoded by eukaryotic genes and well represents 18 of them, we describe PRO with the (ϕ, ψ) plot from the isolated PRO values within ubiquitin and GLY with the Ramachandran plot of the polyglycine analog of ubiquitin [69]. With these Ramachandran plots we use Monte Carlo sampling to obtain (ϕ, ψ) pairs that we then implement on a lone amino acid, where the number of realizations is dependent on the size of the domain of the Ramachandran plot (1408 realizations for ubiquitin). We also rotate the amino acid in $\phi' \in [0, 2\pi)$ by all multiples of $\pi/12$. Then the amino acid is projected onto the $y = 0$ plane and using Monte Carlo (1000 realizations) we calculate the area outside of the effective surface, called A_i, yet within either $\pi/2$ sector of radius $R/2$ centered around $z = 0$, where the ion i will fit according to its vdW radius.

A.1.4 Maximum Velocity

Since v_{b} is the maximum velocity between the amino acid and the longitudinal nanochannel surface as aforementioned, we can use Eq. (A.1) to obtain the max of $\hat{v}$, which is equivalent to v_{b}. In this case we focus on the velocity within the bulk region, namely from the midpoint between the amino acid and the channel surface, $x = x_{\mathrm{mid}} = (R + r_{\circ})/2$, to the channel surface, $x = R$. We employ the following

boundary conditions,

$$\begin{aligned} &\hat{v}(x = R) = 0, \\ &\hat{v}(x = x_{\text{mid}} = \frac{R + r_\circ}{2}) \geq \hat{v}(x), \end{aligned} \tag{A.4}$$

which are very similar to Eq. (A.3). By assuming a constant bulk concentration, g_{b}, over this region we quickly come to a parabolic solution to Eq. (A.1) as well as determining $v_{\text{b}} = q\tilde{z}g_{\text{b}}E_{\perp}(R - r_\circ)^2/4\mu$, where $\tilde{z}_i$ has been simplified to $\tilde{z}$ since both ion species have the same valency. Then we have $v_{\text{b}} = 154.47$ m/s by choosing a reasonable $r_\circ = 4$ Å, a necessity in making v_{b} independent of the amino acid under study, which is more likely in experiment. In fact, the absolute value of the velocity determined from the Stokes equation is known to differ from experiment [67], as opposed to the velocity ratio that we have utilized thus far. However, these differences appear to be systematic [67], and can be solved by dividing v_{b} in half, resulting in the corrected $v_{\text{b}} = 77.23$ m/s.

A.2 Effective Radii

Table A.1: Effective radii in Å

Amino Acid	r_{eff} for Cl^-	r_{eff} for K^+
ALA	8.28	7.68
ARG	3.50	4.49
ASN	5.23	5.55
ASP	6.15	4.69
CYS	7.15	6.85
GLN	4.85	5.28
GLU	8.03	7.44
GLY	6.18	5.75
HIS	6.30	6.08
ILE	5.25	4.78
LEU	5.97	5.74
LYS	5.38	6.43
MET	4.92	4.73
PHE	7.87	7.84
PRO	5.02	4.43
SER	6.85	6.99
THR	4.15	4.57
TRP	6.59	6.74
TYR	6.93	7.29
VAL	5.37	5.27

A.3 Acknowledgements

Appendix A, in full, is a reprint of material as it appears in P. Boynton and M. Di Ventra, "Sequencing proteins with transverse ionic transport in nanochannels", 2015 (under peer-review). arXiv:1509.04772 [physics.bio-ph]. The dissertation author was the primary investigator and author of this publication.

Bibliography

[1] M. Tsutsui, S. Rahong, Y. Iizumi, T. Okazaki, M. Taniguchi, and T. Kawai, "Single-molecule sensing electrode embedded in-plane nanopore," *Sci. Rep.*, vol. 1, 2011.

[2] H. Kwok, K. Briggs, and V. Tabard-Cossa, "Nanopore fabrication by controlled dielectric breakdown," *PLOS ONE*, vol. 9, no. 3, p. e92880, 2014.

[3] A. Storm, J. Chen, X. Ling, H. Zandbergen, and C. Dekker, "Fabrication of solid-state nanopores with single-nanometre precision," *Nat. Mater.*, vol. 2, no. 8, pp. 537–540, 2003.

[4] M. Tsutsui, K. Shoji, M. Taniguchi, and T. Kawai, "Formation and self-breaking mechanism of stable atom-sized junctions," *Nano Lett.*, vol. 8, no. 1, pp. 345–349, 2008.

[5] M. Tsutsui and M. Taniguchi, "Single molecule electronics and devices," *Sensors*, vol. 12, no. 6, pp. 7259–7298, 2012.

[6] A. I. Kolesnikov, J.-M. Zanotti, C.-K. Loong, P. Thiyagarajan, A. P. Moravsky, R. O. Loutfy, and C. J. Burnham, "Anomalously soft dynamics of water in a nanotube: A revelation of nanoscale confinement," *Phys. Rev. Lett.*, vol. 93, p. 035503, 2004.

[7] R. H. Coridan, N. W. Schmidt, G. H. Lai, P. Abbamonte, and G. C. L. Wong, "Dynamics of confined water reconstructed from inelastic x-ray scattering measurements of bulk response functions," *Phys. Rev. E*, vol. 85, p. 031501, 2012.

[8] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, "Scalable molecular dynamics with NAMD," *J. Comput. Chem.*, vol. 26, pp. 1781–1802, 2005.

[9] Y. Yang, M. Berrondo, D. Henderson, and D. Busath, "The importance of water molecules in ion channel simulations," *J. Phys. Condens. Matter*, vol. 16, no. 22, p. S2145, 2004.

[10] M. Zwolak and M. Di Ventra, "*Colloquium*: Physical approaches to DNA sequencing and detection," *Rev. Mod. Phys.*, vol. 80, pp. 141–165, 2008.

[11] J. J. Kasianowicz, E. Brandin, D. Branton, and D. W. Deamer, "Characterization of individual polynucleotide molecules using a membrane channel," *Proc. Natl. Acad. Sci.*, vol. 93, no. 24, pp. 13770–13773, 1996.

[12] M. Zwolak and M. Di Ventra, "Electronic signature of DNA nucleotides via transverse transport," *Nano Lett.*, vol. 5, no. 3, pp. 421–424, 2005.

[13] J. Lagerqvist, M. Zwolak, and M. Di Ventra, "Fast DNA sequencing via transverse electronic transport," *Nano Lett.*, vol. 6, pp. 779–782, 2006.

[14] J. R. Hahn, Y. A. Hong, and H. Kang, "Electron tunneling across an interfacial water layer inside an STM junction: Tunneling distance, barrier height and water polarization effect," *Appl. Phys. A*, vol. 66, pp. S467–S472, 1998.

[15] M. Hugelmann and W. Schindler, "Tunnel barrier height oscillations at the solid/liquid interface," *Surf. Sci.*, vol. 541, no. 1-3, pp. L643–L648, 2003.

[16] G. Hummer, J. C. Rasaiah, and J. P. Noworyta, "Water conduction through the hydrophobic channel of a carbon nanotube," *Nature*, vol. 414, pp. 188–190, 2001.

[17] J. Russo, S. Melchionna, F. De Luca, and C. Casieri, "Water confined in nanopores: Spontaneous formation of microcavities," *Phys. Rev. B*, vol. 76, p. 195403, 2007.

[18] M. Krems, M. Zwolak, Y. V. Pershin, and M. Di Ventra, "Effect of noise on DNA sequencing via transverse electronic transport," *Biophys. J.*, vol. 97, pp. 1990–1996, 2009.

[19] T. Albrecht, "Electrochemical tunnelling sensors and their potential applications," *Nat. Commun.*, vol. 3, p. 829, 2012.

[20] K. Healy, V. Ray, L. J. Willis, N. Peterman, J. Bartel, and M. Drndi, "Fabrication and characterization of nanopores with insulated transverse nanoelectrodes for DNA sensing in salt solution," *Electrophoresis*, vol. 33, no. 23, pp. 3488–3496, 2012.

[21] L. Zhang, C. Mo, T. Wang, and C. Xie, "Strong polarity and bond characterization of nanostructured silicon nitride solids," *MRS Proc.*, vol. 286, p. 107, 1992.

[22] A. Aksimentiev and J. Comer, *Bionanotechnology Tutorial*, 2011.

[23] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *J. Mol. Graph.*, vol. 14, pp. 33–38, 1996.

[24] E. B. Moore and V. Molinero, "Growing correlation length in supercooled water," *J. Chem. Phys.*, vol. 130, no. 24, p. 244505, 2009.

[25] M. Krems, Y. V. Pershin, and M. Di Ventra, "Ionic memcapacitive effects in nanopores," *Nano Lett.*, vol. 10, no. 7, pp. 2674–2678, 2010.

[26] N. Foloppe and A. D. MacKerell, Jr., "All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data," *J. Comput. Chem.*, vol. 21, no. 2, pp. 86–104, 2000.

[27] A. D. MacKerell, Jr. and N. K. Banavali, "All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution," *J. Comput. Chem.*, vol. 21, no. 2, pp. 105–120, 2000.

[28] J. A. Wendel and W. A. Goddard, "The Hessian biased force field for silicon nitride ceramics: Predictions of thermodynamic and mechanical properties for α and βSi_3N_4," *J. Chem. Phys.*, vol. 97, p. 5048, 1992.

[29] S. Mitra, R. Mukhopadhyay, I. Tsukushi, and S. Ikeda, "Dynamics of water in confined space (porous alumina): QENS study," *J. Phys. Condens. Matter*, vol. 13, p. 8455, 2001.

[30] M. Di Ventra, *Electrical Transport in Nanoscale Systems*. Cambridge University Press, 2008.

[31] N. Prokopuk, K.-a. Son, and C. Waltz, "Electron tunneling through fluid solvents," *J. Phys. Chem. C*, vol. 111, pp. 6533–6537, 2007.

[32] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*. Brooks/Cole, 1976.

[33] N. D. Lang and M. Di Ventra, "Comment on "First-principles treatments of electron transport properties for nanoscale junctions"," *Phys. Rev. B*, vol. 68, p. 157301, 2003.

[34] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[35] N. Rusk, "Torrents of sequence," *Nat. Methods*, vol. 8, no. 1, pp. 44–44, 2010.

[36] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC Genom.*, vol. 13, no. 1, p. 341, 2012.

[37] T. Ohshiro, K. Matsubara, M. Tsutsui, M. Furuhashi, M. Taniguchi, and T. Kawai, "Single-molecule electrical random resequencing of DNA and RNA," *Sci. Rep.*, vol. 2, 2012.

[38] M. Tsutsui, K. Matsubara, T. Ohshiro, M. Furuhashi, M. Taniguchi, and T. Kawai, "Electrical detection of single methylcytosines in a DNA oligomer," *J. Am. Chem. Soc.*, vol. 133, no. 23, pp. 9124–9128, 2011.

[39] M. Murphy, I. Rasnik, W. Cheng, T. M. Lohman, and T. Ha, "Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy," *Biophys. J.*, vol. 86, no. 4, pp. 2530–2537, 2004.

[40] M. Tsutsui, M. Taniguchi, K. Yokota, and T. Kawai, "Identifying single nucleotides by tunnelling current," *Nat. Nanotechnol.*, vol. 5, pp. 286–290, 2010.

[41] Y. He, M. Tsutsui, C. Fan, M. Taniguchi, and T. Kawai, "Controlling DNA translocation through gate modulation of nanopore wall surface charges," *ACS Nano*, vol. 5, no. 7, pp. 5509–5518, 2011.

[42] T. Ahmed, J. T. Haraldsen, J. J. Rehr, M. Di Ventra, I. Schuller, and A. V. Balatsky, "Correlation dynamics and enhanced signals for the identification of serial biomolecules and DNA bases," *Nanotechnology*, vol. 25, no. 12, p. 125705, 2014.

[43] S. Garaj, W. Hubbard, A. Reina, J. Kong, D. Branton, and J. Golovchenko, "Graphene as a subnanometre trans-electrode membrane," *Nature*, vol. 467, pp. 190–193, 2010.

[44] S. Huang, J. He, S. Chang, P. Zhang, F. Liang, S. Li, M. Tuchband, A. Fuhrmann, R. Ros, and S. Lindsay, "Identifying single bases in a DNA oligomer with electron tunnelling," *Nat. Nanotechnol.*, vol. 5, pp. 868–873, 2010.

[45] A. Pecchia, M. Gheorghe, A. Di Carlo, P. Lugli, T. A. Niehaus, T. Frauenheim, and R. Scholz, "Role of thermal vibrations in molecular wire conduction," *Phys. Rev. B*, vol. 68, p. 235321, 2003.

[46] L. Fenton, "The sum of log-normal probability distributions in scatter transmission systems," *Commun. Syst. IRE Trans.*, vol. 8, no. 1, pp. 57–67, 1960.

[47] D. L. Nelson, A. L. Lehninger, and M. M. Cox, *Lehninger Principles of Biochemistry*. Macmillan, 2008.

[48] N. W. Kelley, V. Vishal, G. A. Krafft, and V. S. Pande, "Simulating oligomerization at experimental concentrations and long timescales: A Markov state model approach," *J. Chem. Phys.*, vol. 129, no. 21, p. 214707, 2008.

[49] P. G. Katsoyannis and A. Tometsko, "Insulin synthesis by recombination of A and B chains: A highly efficient method," *Proc. Natl. Acad. Sci.*, vol. 55, no. 6, p. 1554, 1966.

[50] I. S. Johnson, "Human insulin from recombinant DNA technology," *Science*, vol. 219, no. 4585, pp. 632–637, 1983.

[51] P. Schmitt-Kopplin and M. Frommberger, "Capillary electrophoresis–mass spectrometry: 15 years of developments and applications," *Electrophoresis*, vol. 24, no. 22-23, pp. 3837–3867, 2003.

[52] K. G. Standing, "Peptide and protein de novo sequencing by mass spectrometry," *Curr. Opin. Struct. Biol.*, vol. 13, no. 5, pp. 595–601, 2003.

[53] P. Edman, "Method for determination of the amino acid sequence in peptides," *Acta Chem. Scand.*, vol. 4, no. 7, 1950.

[54] R. A. Laursen, "Solid-phase Edman degradation," *Eur. J. Biochem.*, vol. 20, no. 1, pp. 89–102, 1971.

[55] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss, "The potential and challenges of nanopore sequencing," *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1146–1153, 2008.

[56] D. S. Talaga and J. Li, "Single-molecule protein unfolding in solid state nanopores," *J. Am. Chem. Soc.*, vol. 131, no. 26, pp. 9287–9297, 2009.

[57] L. Movileanu, "Interrogating single proteins through nanopores: Challenges and opportunities," *Trends Biotechnol.*, vol. 27, no. 6, pp. 333–341, 2009.

[58] T. Ohshiro, M. Tsutsui, K. Yokota, M. Furuhashi, M. Taniguchi, and T. Kawai, "Detection of post-translational modifications in single peptides using electron tunnelling currents," *Nat. Nanotechnol.*, 2014.

[59] J. Wilson and M. Di Ventra, "Single-base DNA discrimination via transverse ionic transport," *Nanotechnology*, vol. 24, no. 41, p. 415101, 2013.

[60] L. D. Menard, C. E. Mair, M. E. Woodson, J. P. Alarie, and J. M. Ramsey, "A device for performing lateral conductance measurements on individual double-stranded DNA molecules," *ACS Nano*, vol. 6, no. 10, pp. 9087–9094, 2012.

[61] R. Wolfenden, L. Andersson, P. Cullis, and C. Southgate, "Affinities of amino acid side chains for solvent water," *Biochemistry*, vol. 20, no. 4, pp. 849–855, 1981.

[62] J. Chang, A. M. Lenhoff, and S. I. Sandler, "Solvation free energy of amino acids and side-chain analogues," *J. Phys. Chem. B*, vol. 111, no. 8, pp. 2098–2106, 2007.

[63] M. M. Kish, G. Ohanessian, and C. Wesdemiotis, "The Na+ affinities of α-amino acids: Side-chain substituent effects," *Int. J. Mass Spectrom.*, vol. 227, no. 3, pp. 509–524, 2003.

[64] L. Rulíšek and Z. Havlas, "Theoretical studies of metal ion selectivity. 1. DFT calculations of interaction energies of amino acid side chains with selected transition metal ions (Co2+, Ni2+, Cu2+, Zn2+, Cd2+, and Hg2+)," *J. Am. Chem. Soc.*, vol. 122, no. 42, pp. 10428–10439, 2000.

[65] W. R. Rudnicki and B. M. Pettitt, "Modeling the DNA-solvent interface," *Biopolymers*, vol. 41, no. 1, pp. 107–119, 1997.

[66] M. Feig and B. M. Pettitt, "Sodium and chlorine ions as part of the DNA solvation shell," *Biophys. J.*, vol. 77, no. 4, pp. 1769–1781, 1999.

[67] R. Qiao and N. Aluru, "Ion concentrations and velocity profiles in nanochannel electroosmotic flows," *J. Chem. Phys.*, vol. 118, no. 10, pp. 4692–4701, 2003.

[68] G. Stirnemann, S.-g. Kang, R. Zhou, and B. J. Berne, "How force unfolding differs from chemical denaturation," *Proc. Natl. Acad. Sci.*, vol. 111, no. 9, pp. 3413–3418, 2014.

[69] G. Stirnemann, D. Giganti, J. M. Fernandez, and B. Berne, "Elasticity, structure, and relaxation of extended proteins under force," *Proc. Natl. Acad. Sci.*, vol. 110, no. 10, pp. 3847–3852, 2013.

[70] G. Ramachandran and V. Sasisekharan, "Conformation of polypeptides and proteins," *Adv. Protein Chem.*, vol. 23, p. 283, 1968.

[71] J. S. Richardson, "The anatomy and taxonomy of protein structure," *Adv. Protein Chem.*, vol. 34, pp. 167–339, 1981.

[72] R. Gao, Y.-L. Ying, B.-Y. Yan, and Y.-T. Long, "An integrated current measurement system for nanopore analysis," *Chin. Sci. Bull.*, vol. 59, no. 35, pp. 4968–4973, 2014.

[73] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.

[74] L. D. Menard and J. M. Ramsey, "Fabrication of sub-5 nm nanochannels in insulating substrates using focused ion beam milling," *Nano Lett.*, vol. 11, no. 2, pp. 512–517, 2010.

[75] M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez, "Mechanical and chemical unfolding of a single protein: A comparison," *Proc. Natl. Acad. Sci.*, vol. 96, no. 7, pp. 3694–3699, 1999.

[76] R. B. Merrifield, "Solid phase peptide synthesis. I. The synthesis of a tetrapeptide," *J. Am. Chem. Soc.*, vol. 85, no. 14, pp. 2149–2154, 1963.

[77] P. E. Dawson, T. W. Muir, I. Clark-Lewis, and S. Kent, "Synthesis of proteins by native chemical ligation," *Science*, vol. 266, no. 5186, pp. 776–779, 1994.

[78] W. Hou, X. Zhang, F. Li, and C.-F. Liu, "Peptidyl N,N-bis(2-mercaptoethyl)-amides as thioester precursors for native chemical ligation," *Org. Lett.*, vol. 13, no. 3, pp. 386–389, 2010.

[79] M. Roberts, M. Bentley, and J. Harris, "Chemistry for peptide and protein PEGylation," *Adv. Drug Deliv. Rev.*, vol. 64, pp. 116–127, 2012.

[80] A. Ashkin, J. Dziedzic, J. Bjorkholm, and S. Chu, "Observation of a single-beam gradient force optical trap for dielectric particles," *Opt. Lett.*, vol. 11, no. 5, pp. 288–290, 1986.

[81] D. G. Grier, "A revolution in optical manipulation," *Nature*, vol. 424, no. 6950, pp. 810–816, 2003.

[82] U. F. Keyser, B. N. Koeleman, S. Van Dorp, D. Krapf, R. M. Smeets, S. G. Lemay, N. H. Dekker, and C. Dekker, "Direct force measurements on DNA in a solid-state nanopore," *Nat. Phys.*, vol. 2, no. 7, pp. 473–477, 2006.

[83] E. H. Trepagnier, A. Radenovic, D. Sivak, P. Geissler, and J. Liphardt, "Controlling DNA capture and propagation through artificial nanopores," *Nano Lett.*, vol. 7, no. 9, pp. 2824–2830, 2007.

[84] F. Arai, K. Yoshikawa, T. Sakami, and T. Fukuda, "Synchronized laser micromanipulation of multiple targets along each trajectory by single laser," *Appl. Phys. Lett.*, vol. 85, no. 19, pp. 4301–4303, 2004.

[85] D. G. Grier and Y. Roichman, "Holographic optical trapping," *Appl. Opt.*, vol. 45, no. 5, pp. 880–887, 2006.

[86] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, 2009.

[87] A. D. MacKerell, M. Feig, and C. L. Brooks, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," *J. Comput. Chem.*, vol. 25, no. 11, pp. 1400–1415, 2004.