UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ERROR ANALYSIS AND PARAMETER ESTIMATION FOR NANOPORE BASED MOLECULAR DETECTION

A thesis submitted in partial satisfaction of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

by

Christopher R. O'Donnell

March 2018

The Thesis of Christopher R. O'Donnell
is approved:

Professor Ricardo Sanfelice, Chair

Professor Gabriel H. Elkaim

William B. Dunbar, Ph.D

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

Error analysis and parameter estimation for nanopore based molecular detection

by

Christopher R. O'Donnell

Nanopores are powerful tools for measuring and probing single molecules. A nanopore is a nanometer-sized opening in a membrane that separates two chambers filled with buffered ionic solution. By applying a voltage and measuring the ionic current through the nanopore, it is possible to detect the presence of individual DNA, RNA and proteins as they pass through the pore, and even read the sequence of individual nucleobases that make up a single strand of DNA. However, the speed with which molecules translocate and the size of the sensing region have presented challenges for using nanopores to sequence DNA. Most nanopore-based DNA sequencing research focuses on using biological nanopores paired with an enzyme to slow down the passage of DNA through the pore, but recent advances in solid-state fabrication technology have made it possible to create artificial solid-state nanopores in insulating membranes, typically made of silicon. These pores can be made in a larger range of sizes, are more durable, and are more amenable to large scale fabrication than their biological counterparts. In order to control the rate of molecular translocation through solid-state nanopores, researchers are developing a two-pore architecture, which utilizes time-varying voltage patterns to enable rereading of individual molecules to gain confidence in feature sensing. This thesis presents a numerical study that provides an error analysis of an idealized nanopore sequencing method in which ionic current measurements are used to sequence intact single-stranded DNA in the pore while an enzyme controls DNA motion. This analysis presents examples of systematic and random errors associated with this method of sequencing and demonstrates the necessity of rereading

sequences at least 140 times to achieve 99.99% accuracy. Two different methods of parameter estimation are then presented that overcome the problem of contamination of the measured ionic current by capacitive elements in the system and facilitate active control with the two-pore architecture.

To my wife Meghan

you are my complementary strand

without which I would be lost

Acknowledgments

First I want to thank Bill Bigley and Professor Pat Mantey without whose intervention I would not have come back to UCSC for my graduate work.

I am eternally grateful to my advisor Bill Dunbar for everything he has done for me. I am fortunate to have him as an advisor, mentor, colleague and friend. I would not be where I am today without his guidance and his courage to venture out and start his own company, and bring me along for the ride.

Professors Don Wiberg and Hongyun Wang were instrumental in helping me understand the complexities and appreciate the beauty of statistical modeling and they both profoundly contributed to my research.

I am grateful to the many people who helped me over the course of my studies with academic insight, encouragement and support including Shea Ellerson, Raj Maitra, Darrel Deo, Aimen Al-Refai, Daniel Garalde, and Robin Abu-Shumays.

Finally I want to thank my family for their continuous love and support and for nodding along politely as I explained my research even though most of the time they have no idea what I'm talking about.

Chapter 1

Introduction

Nanopores provide a simple method of serially measuring and manipulating individual DNA and DNA-protein complexes, one molecule at a time [84]. A generic nanopore device consists of a nanometer-sized opening in a membrane that separates two chambers containing a buffered electrolytic solution. An electrode is placed in each chamber and a patch-clamp amplifier applies a voltage across the membrane creating an ionic current through the nanopore (Figure 1.1). Charged particles such as DNA are placed on one side of the nanopore and are electrophoretically driven through the pore by the transmembrane electric field induced by the applied voltage. When a molecule passes through the nanopore, the resistance of the pore increases causing a drop in the measured ionic current. The depth and duration of the current drop reveal information about the size and charge of the molecule passing through the pore [73]. It is this phenomenon that enables nanopores to be used as single molecule sensors. The speed and direction with which molecules pass through the nanopore depends on the amplitude and the polarity of the voltage, as well as the charge of the molecule, and this creates the opportunity to use methods from control theory to manipulate and maximize the utility of nanopores. There are two main classes of nanopore devices: biolog-

ical and solid-state. As their names suggest, biological nanopore devices utilize biomolecules (lipids and proteins) to form membranes and nanometer-sized openings while solid-state nanopore devices rely on mechanical means to etch and drill nanopores in solid substrates such as silicon or graphene.

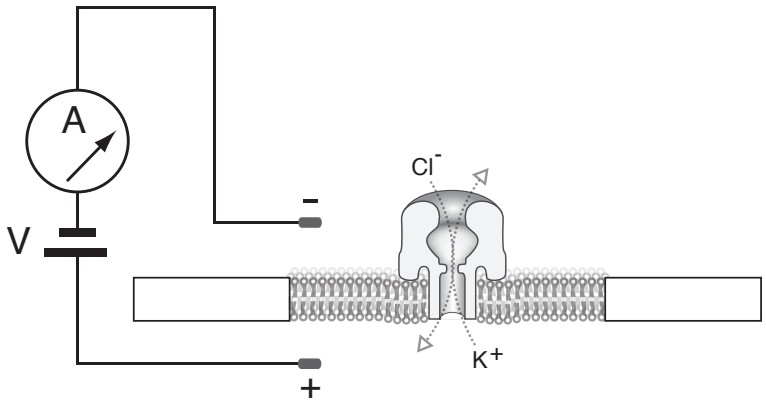

Figure 1.1: Schematic of a biological nanopore device. A single α-HL nanopore is inserted in a lipid bilayer that separates two chambers containing a buffered solution of KCl. A transmembrane voltage is applied across the bilayer and the induced ionic current through the nanopore is measured by electrodes in series with a patch-clamp amplifier.

1.1 Biological nanopores

Biological nanopores offer several advantages for single-molecule DNA analysis. First, large numbers of biological nanopores can be produced with an atomic level of precision and a remarkable heterogeneity in terms of size and composition. Second, detailed information about the molecular structure of biological nanopores is available via X-ray crystallography. Finally, biological nanopores have the ability to be genetically modified, using established techniques such as mutagenesis, to tailor the physical and chemical properties of the pore to fit a given application [84].

Many different proteins have been investigated as candidates for biological nanopores, but the most commonly used protein is α-hemolysin (αHL). Secreted by the bacterium *Staphylococcus aureus*, αHL is a cytotoxin that acts as the primary virulence factor in *S. aureas* pneumonia by spontaneously inserting into a foreign cell's lipid bilayer, disrupting the cell's electrochemical gradient [65]. αHL has a mushroom-shaped structure consisting of a spherical vestibule with an internal diameter of approximately 3.6 nm and a stem consisting of an approximately 5 nm long and 2.2 nm wide β-barrel [74] (Figure 1.2A). The inner diameter of the nanopore reduces down to approximately 1.4 nm where the stem meets the vestibule creating a limiting aperture that restricts the size of the molecules that can pass through the pore. The limiting aperture is wide enough to allow single-stranded DNA (ssDNA), RNA and unfolded protein chains to translocate through the nanopore, while double-stranded DNA (dsDNA) can enter the vestibule but not pass through the pore [52].

Although αHL is the most widely used protein for nanopore experimentation, the pore does not have the sensitivity required to detect the sequence of a translocating DNA strand [26, 62, 63]. In an attempt to overcome this problem, researchers at the University of Oxford (led by Hagan Bayley) first engineering αHL to have a single DNA oligonucleotide attached to the inside of the vestibule enabling the pore to identify single-base mismatches in translocating DNA [31]. They further improved the sensitivity of αHL by covalently attaching a molecular adapter in the lower stem [13] and modifying the amino acid side chains that affect the recognition sites within the β-barrel [76]. Even with these improvements, αHL has a fundamental structural flaw in the length of the stem where sensing of molecules takes place. Approximately 10-12 nucleotides at a time can fit inside the β-barrel that makes up the stem, and all of these nucleotides contribute to the

measured ionic current, which masks the unique current signatures of individual nucleotides [76].

Researchers continue to look for alternative proteins to use as biological nanopores that improve on the sensitivity of αHL and found one in the form of *Mycobacterium smegmatis* porin A (MspA). MspA is a channel-forming protein that constitutes the major diffusion pathway for hydrophilic molecules in the bacterium *M. smegmatis* [75]. Like αHL, MspA has a limiting aperture of $\sim$1.2 nm which makes it wide enough for ssDNA to translocate through the pore but too narrow for dsDNA (Figure 1.2A). Unlike αHL, MspA has a funnel-shaped geometry with a sensing region at the bottom of the pore that is only $\sim$0.5 nm long. Researchers at the University of Washington (led by Jens Gundlach) have genetically modified MspA so that only 3-4 nucleotides contribute to the measured ionic current yielding a 3.5-fold enhancement in nucleotide separation efficiency as compared to wild-type αHL [16]. This improvement along with the possibility of further genetic modification to improve the sensitivity of the pore has made MspA one of the most promising biological nanopores for DNA sequencing to date.

Even though the sensitivity of biological nanopores has improved, the speed with which ssDNA moves through the pore is too fast to perform sequencing directly in realtime. Under typical experimental conditions, intact ssDNA passes through a nanopore with an average rate that approaches $\sim$1 nt/μs while the hardware used to measure the small ionic currents requires a rate of $\geq$1 nt/ms to achieve single base recognition [7]. Researchers have devised several methods to overcome this issue and slow down the translocation rate of ssDNA. The most promising of these strategies involves the use of enzymes to regulate the motion of DNA through the pore [4, 57, 14, 43]. Coupling an enzyme motor with a nanopore is an attractive strategy because the enzyme-DNA complex forms in

bulk solution, which allows it to be electrophoretically captured in the nanopore. Once the enzyme-DNA complex is captured, the enzyme processively steps the DNA molecule through the nanopore in a relatively slow and controlled manner providing ample time for the small changes in the ionic current to be measured.

One of the most successful uses of this method was developed by researchers at the University of California, Santa Cruz (led by Mark Akeson) utilizing the bacteriophage phi29 DNA polymerase in conjunction with an αHL nanopore [43]. Akeson and his group devised a method of using a blocking oligo which allows phi29 to attach itself to a DNA strand in bulk phase but prevents the enzyme from processing the strand until the enzyme-DNA complex is captured in a nanopore [12]. Once captured in the pore, the blocking oligo is stripped off by the force of the applied transmembrane voltage and phi29 begins base-by-base ratcheting of the DNA through the nanopore at a rate of $\sim$2.5 nt/s, which is suitable for sequencing [12]. Akeson and his group further improved their sequencing method by coupling phi29 with the modified MspA nanopore developed by Gundlach and his research group at the University of Washington (Figure 1.2C). They were able to show that individual ssDNA molecules traversing through the short and narrow constriction of MspA under the control of phi29 produced distinct sequence specific current levels [49]. While these results are very promising, this method still faces several significant hurdles. First, the measured ionic current levels are a function of the 3-4 nucleotides residing in the sensing region of the nanopore, which necessitates a significant computationally effort to deconvolve the signal into a specific sequence. Second, the multi-nucleotide signal levels combined with the nondeterministic ratcheting motion of phi29 makes it difficult to accurately distinguish the number of bases in homopolymer regions of DNA. Finally, phi29 is an imperfect molecular motor that can randomly slip and skip several bases at

a time or temporarily ratchet DNA in the opposite direction [12, 49].

While biological nanopores are powerful tools for single-molecule analysis they do have some inherent disadvantages. The lipid bilayer that supports biological nanopores is mechanically unstable and most lipids used in academic research rely on the spontaneous formation of a bilayer. Once the bilayer membrane is formed, a single protein pore must spontaneously insert into the bilayer through a process that involves feedback control and perfusion to minimize the chance of a second insertion. Biological nanopores are also very sensitive to experimental conditions such as pH, temperature and salt concentration requiring researchers to tailor experiments around the chemistry of the pore as well as that of the molecule being investigated. Finally, development of a commercial product using biological nanopores requires the difficult task of integrating a biological system into large-scale arrays.

1.2 Solid-state nanopores

Some of the disadvantages of biological nanopores have been addressed by the use of solid-state nanopores. Solid-state nanopores are nanometer-sized openings formed in a solid substrate by drilling or etching [15]. In comparison to biological nanopores, solid-state nanopores are mechanically and chemically more stable [33], offer the ability to tune the size and shape of the nanopore with subnanometer precision [77], and the ability to fabricate high-density arrays of nanopores [36]. The most common substrate used for forming solid-state nanopores is silicon (typically silicon nitride (SiN) or silicon oxide (SiO_2)) due to its high chemical resistance and low mechanical stress [41], but recently researchers looking to improve solid-state nanopore performance have also used aluminum oxide (Al_2O_3 [87] and graphene [69, 21] (Figure 1.2B).

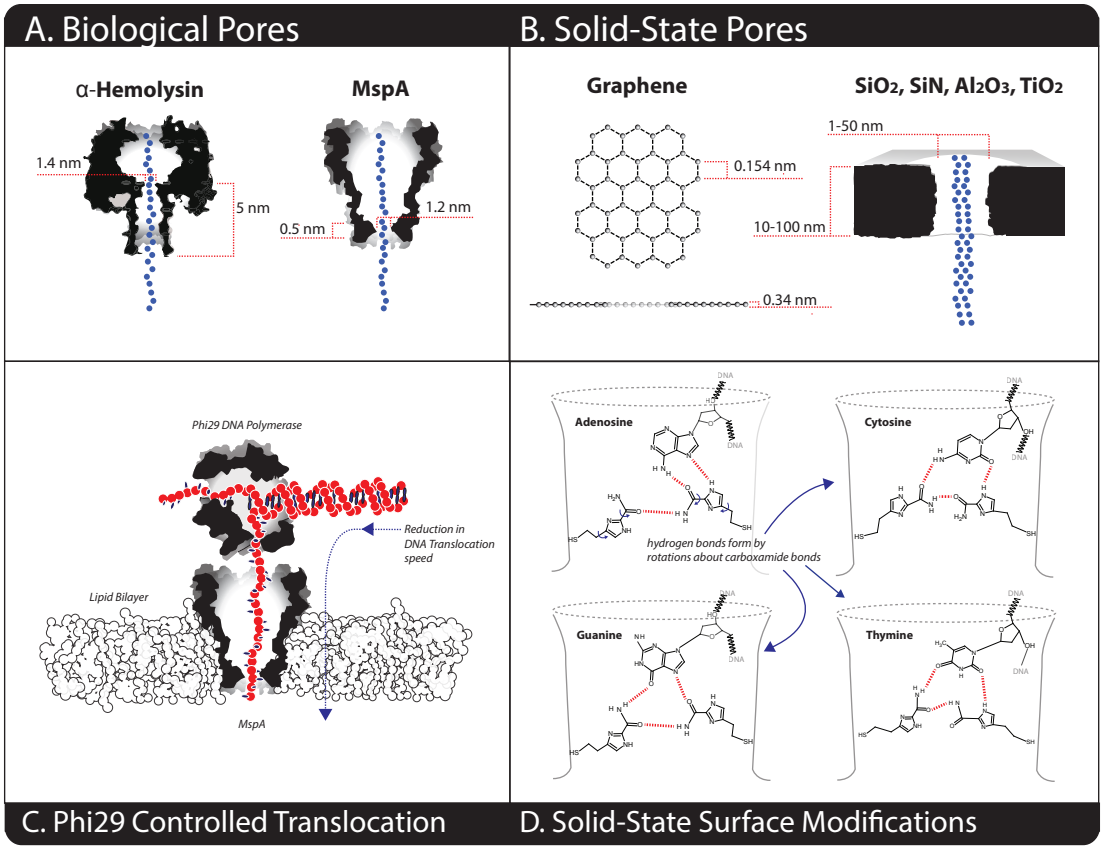

Figure 1.2: Varying types and geometries of nanopores (A, B) and leading configurations for sequencing (C, D). (A) Relative dimensions (approximate) of biological pores α-hemolysin and MspA. (B) Dimensions of graphene with single-atom thickness, and range of pore diameters in varying solid-state substrates. (C) The mechanism of phi29 polymerase mediated DNA translocation developed in [12] and implemented on the MspA nanopore in [49]. The motorï£¡s ability to both polymerize (at ∼40 nt/s) and unzip (at ∼2.5 nt/s) the strand is utilized to register DNA motion progress and position sensing, with the unzipping direction shown in this illustration. Though dwell times at each position are not constant, but exponentially distributed, the rates meet the requirements for DNA speed reduction [7]. (D) Functionalized electrode readers of nucleobases via 4(5)-substituted 1-H-imidazole-2-carboxamide. Different 180° rotations occur over the specified bonds on the carboxamidemolecules to allow hydrogen bonding with different nucleobases, causing detectable variations in electron tunneling signals between two electrodes attached to the two carboxamides.

The size and shape of solid-state nanopores can vary depending on the substrate and fabrication technique used. Researchers at Harvard University (led

by Jene Golovchenko) developed a novel technique called ion beam sculpting that uses a focused ion beam (FIB) to mill tiny holes in SiN membranes with nanometer precision [41]. This technique has been widely used to create pores with nanometer dimensions and provided a starting point for DNA translocation measurements [15]. Another group of researchers at Delft University (led by Cees Dekker) adapted a technique from silicon microfabrication that uses electron-beam lithography and reactive-ion etching to create relatively large (∼50 nm) holes in silicon oxide (SiO_2). A transmission electron microscope (TEM) is then used to soften the SiO_2 allowing it to slowly deform and shrink the holes down to as small as 2 nm [77]. This technique enables direct visual feedback through the use of the TEM and provides a way to fine-tune solid-state nanopores with subnanometer precision [15]. Several other researchers have used the focused electron beam of a TEM to directly drill sub-10 nm nanopores [92, 20, 2, 87]. Other promising fabrication techniques including atomic-layer deposition [86], sputtering and evaporation [89], and chemical etching [1] have also been developed. The advancement of these techniques over recent years has enabled researchers to fabricate nanopores as small as 1 nm, making them smaller than the limiting apertures of αHL or MspA [15]. However, unlike biological nanopores, the exact internal dimensions of solid-state pores are unknown, which means that two solid-state nanopores with the same size opening can behave differently when conducting ionic current [71].

The sensitivity of solid-state as well as biologic nanopores is predicated on the size of the sensing region, which for solid-state nanopores is a function of the diameter of the pore opening and the thickness of the membrane. With this in mind, it is no surprise that there is great interest in graphene as a potential solid-state DNA sequencing platform. Graphene is a two-dimensional sheet of carbon atoms with a thickness of only one atomic layer (∼0.34 nm) [53]. The thickness

of a single layer of graphene is comparable to the spacing between nucleotides in ssDNA (0.32-0.52 nm), which means a graphene nanopore could theoretically detect translocating DNA with single-nucleotide precision [84]. Initial experiments with graphene nanopores have shown the detection of dsDNA [21, 50, 69] as well as ssDNA [61]. However, the measured ionic current through bare graphene nanopores is noisier than that of comparable SiN nanopores and it is difficult to fabricate the pores without defects [50]. Attempts to ameliorate these issues have been made by using atomic-layer deposition to coat graphene with TiO_2 [50], but further research and experimentation is needed to realized the full potential of graphene nanopores.

Another promising technique for detecting individual nucleotides with solid-state nanopores is the use of a tunneling current. Researchers at Osaka University were the first to embedded nanoelectrodes within a solid-state nanopore and use a transverse tunneling current across the pore to identify single nucleotides [80] and slow the translocation rate of DNA through the pore [79]. Researchers at Arizona State University (led by Stuart Lindsay), expanded on this technique by functionalizing embedded electrodes with a benzamide-based molecule that hydrogen-bonds to DNA bases in different orientations [42]. Translocating nucleobases interact with the molecule attached to the tips of the electrodes and create a transient tunneling current across the nano-gap between the electrodes with a distinct signal for each base due to the specific bond orientation [32] (Figure 1.2D). This method leads the field of solid-state nanopore sequencing with a detection signal-to-noise ratio considerably higher than what appears to be possible with ionic-current based sensing through biological nanopores.

Even though researchers have developed novel techniques for improving the sensitivity of solid-state nanopores, the significant challenge of controlling the rate

of molecular translocation through the pore still exists. Researchers at IBM have proposed extending the concept of embedded electrodes to create a solid-state nanopore consisting of a metal-insultator-metal sandwich that acts as a transistor [60]. This “DNA transistor" has been shown in simulations to be capable of controlling the translocation of DNA through a nanopore in a base-by-base ratchet fashion [47], similar to phi29 with biological nanopores, but this method has yet to be demonstrated in actual experiments. It has been postulated that phi29 or other processive enzymes could be coupled with solid-state nanopores to help with rate control. However, it is unclear if the imprecise and generally unknown surface structures of solid-state nanopores would be conducive to preserving enzymatic function. Researchers have been able to form hybrid biological-solid-state nanopores by inserting an αHL pore tethered to a strand of DNA into a solid-state nanopore [27], which may provide a suitable platform to couple with phi29.

1.3 Two-pore architecture

In lieu of the rate control methods described above, researchers at Two Pore Guys, Inc. (2PG) are developing a two-pore solid-state nanopore architecture that is designed to slow down the molecule passing through the pores for dramatically improved sensing. The two-pore architecture consists of two solid-state nanopores, positioned sufficiently close together as to allow the co-capture of a single strand of DNA, and three electrically isolated fluidic chambers that enable independent voltages to be applied across each nanopore (Figure 1.3).

To precisely control the motion of DNA with a two-pore architecture, a strand must first be captured in both pores. This can be accomplished by applying a transmembrane voltage across both nanopores to capture a DNA strand first in one pore and, as it translocates, the leading end of the strand is captured again in

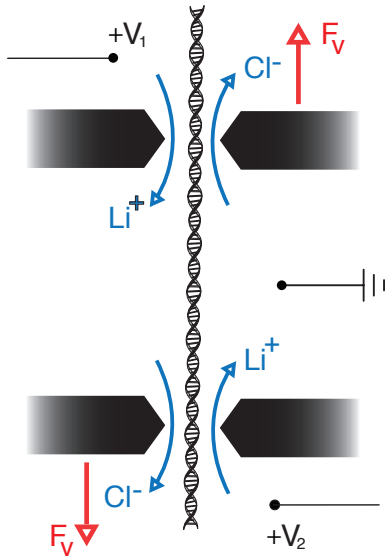

Figure 1.3: Schematic of the two-pore architecture. Two solid-state pores are separated by a common chamber creating three electrically isolated fluidic compartments. A working electrode is placed in the chambers above each pore and a common ground electrode is place in the middle chamber. Two separate voltages are applied across each membrane enabling the capture and control of molecules in both nanopores.

the second pore. Upon capture in both pores, it is theorized that a DNA strand can be held between the nanopores by applying equal and opposite voltage forces across the two membranes or slowly pulled through the pores by creating a small force differential between the opposing voltages. Once a single strand of DNA has been captured in both pores, the features of the molecule could be reread repeatedly by alternating the polarity of the command voltages.

In order to gain statistical confidence about features of a molecule that have been sensed with a nanopore, many different sensing reads of those features must be performed and combined to reach a consensus. Multiple reads of a given molecule can be achieved by creating many copies of that molecule through amplification techniques. However, these techniques are costly and have error rates

of their own resulting in a population of not-quite-perfect copies of the original molecule [38]. A much more sound strategy is to reread the exact same molecule multiple times to generate a sufficient number of reads to form a statistically high confidence consensus.

The two-pore solid-state nanopore architecture under development is capable of rereading a single molecule multiple times, but this operation requires switching the polarity of the command voltages on microsecond time scales. A challenge with time-varying voltages is that the capacitive elements of the nanopore system induce a transient response when a step change in voltage occurs. This capacitive effect contaminates the measured ionic current and limits the time-resolution for detecting DNA or DNA-protein dynamics making it impossible to measure the population of molecular responses that are faster than the transient settling time (up to 30% in [90]).

The work presented in this thesis provides motivation for developing the two-pore architecture as a method to improve the utility of nanopores as single molecule sensors and details companion tools that are fundamental for achieving that goal. The first study presented examines the necessity for performing multiple reads of a single molecule and how to combine those reads into a statistically high confidence consensus. Two different methods of parameter estimation are then presented that overcome the problem of contamination of the measured ionic current by capacitive elements in the system and facilitate active control.

Chapter 2

Error Analysis of Idealized Nanopore Sequencing

The following chapter describes excerpts from work published in *Electrophoresis* ([55]) for which I was first author.

2.1 Introduction

High-throughput sequencing technologies can generate genome-scale sequence data with high accuracy, making it possible to identify genomic markers for a growing list of common diseases, including cancers [91]. The leading commercial platforms (Roche, Illumina, Life Technologies) can generate 1-100s of gigabases per instrument run, with run times on the order of hours to days. For technologies that achieve at most 1% raw error rates, however, read lengths are short, generally tens to hundreds of base pairs. Such short-read sequencing necessitates massive data storage requirements and complex bioinformatics algorithms for genome alignment and assembly, and complicates studies involving linkage analysis. The short reads also require the devices to have a high degree of parallelization, so that

there is sufficient coverage of the sequenced DNA to achieve desired error thresholds. Still another drawback is the need for an amplification step using enzymes that have less then 100% fidelity. In particular, it is common that starting material is amplified to create a library for sequencing, which then undergoes a second amplification reaction to create a clonal colony as in Illumnia's on-chip bridged amplification reaction [51]. Such intensive sample preparation may also require unattainable amounts of starting material. Despite these issues, the short-read and massively parallel devices control the market principally because they provide the highest throughput and sufficiently low error rates.

Single-molecule sequencing (SMS) devices have alleviated the sample preparation requirements of massively-parallel devices by eliminating the need for template amplification [78]. The SMS from Helicos Biosciences (HeliScope) preserves the high-throughput feature ($\sim$3 Gb/day), but reads remain short ($<$ 60 bp) and errors are higher (3-5%), diminishing the value of simpler sample preparation [91]. The SMS from Pacific Biosciences (PacBio RS) boosts read lengths to 10 kb, but throughput is reduced ($<$ 0.1 Gb/run) and error rates are considerably higher (15%). Errors can be reduced with this technology by using circular template DNA, but at the price of shorter read lengths [78]. Despite the high error rate, the long-read feature of the PacBio RS technology makes it useful to use in concert with short-read and low-error platforms, specifically for whole-genome sequencing in which the longer reads provide alignment scaffolds for the short read contigs (though DNA mapping technologies [39] are competing for this market).

The ideal sequencing platform would require minimal sample preparation and zero amplification, would be modular and scalable to ensure sufficient throughput for any given application, and would have sufficiently long reads *and* low errors to permit robust detection of any feature, including rare variants [37] and structural

variants such as repetitive regions [91]. No single platform currently possesses all of these assets. Nanopores have been pursued as a candidate SMS platform in university research labs [7], and a subset of the resulting intellectual property has been commercialized, most notably by the company Oxford Nanopore Technologies (ONT) [59]. ONT's sequencing platform, the MinION, utilizes chemically modified biological nanopore channels, promising minimal sample preparation and read lengths up to hundreds of kb [34]. Sample preparation requires no amplification, nor labeling of nucleotides; instead, individual DNA strands are captured by electrophoresis into each nanopore channel from a bulk-phase chamber, and the impeded channel current is used to sense the nucleotides that pass through the limiting constriction of the channel. This work considers a model method in which intact ssDNA is threaded through a biological pore for sequencing [12, 49], as opposed to an alternative approach in which mononucleotides are sensed in concert with exonuclease-catalyzed ssDNA hydrolysis above the pore [13]. Unfortunately, intact ssDNA passes too fast through the pore when the rates of electrophoresis are unimpeded ($\sim$1 Mb/s), when compared to the ionic current measurement bandwidth ($\sim$1 kb/s) [7]. To keep ssDNA motion within measurement bandwidths, a leading nanopore sequencing method uses a DNA polymerase enzyme perched on top of the pore to control the rate of each DNA molecule through the pore [12]. In this configuration, the sensitivity of biological pores for identifying the sequence of intact ssDNA has improved, with the occluded current through the MspA pore a function of 4 nucleotides positioned at the narrowest constriction of the channel [49], and ONT claiming modified pores that are sensitive to 3 nucleotides at a time.

We consider an idealized nanopore sequencer in which an enzyme controls ssDNA motion through the pore, and the ionic current amplitude is a function of

one or more nucleotides. When more than one nucleotide affects the current (e.g., a triple), systematic errors may make it impossible to resolve certain sequences, regardless of depth of coverage; we quantitatively consider examples where this is the case. Notably, such errors would also persist when using nanopores regardless of the method of DNA control (i.e., with enzymes or by any other method). Absent these systematic errors, we consider next random errors introduced by the use of an enzyme to control ssDNA motion through the pore. Specifically, the enzyme is idealized by modeling ssDNA motion as moving in single nucleotide steps with durations from an exponential distribution of known rate (we ignore backtracking which has been experimentally observed [12, 49]). When homopolymer regions move through the pore with no change in current amplitude, the number of nucleotides associated with each detectable amplitude level must be inferred, and this introduces random insertion or deletion errors that can be reduced only by rereading the same sequence multiple times. We derive an analytic expression for the rate of error decay as a function of the number of reads, and examine the resulting error rate trends for known sequences (16.6 kb Human Mitochondrial DNA [68], 4.6 Mb Escherichia coli K-12 [5]). We then simulate nanopore signals to incorporate the effects of added measurement noise and the consequent low-pass filtering required to reduce noise for robust amplitude detection. Using a novel amplitude-level detection and duration binning method for base calling, consensus sequences generated in the noiseless case are shown to match the analytic trends exactly, and increasing noise is shown to increase the error rate.

2.2 Methods

2.2.1 Simulated nanopore signals

All simulations were performed using the MATLAB software package. The nanopore sensor was modeled as having single nucleotide sensitivity producing distinct ionic current amplitudes at $(3, 2, 1, 0)$ pA for the nucleobases (A, G, C, T). We varied this amplitude-to-base assignment and observed no measurable difference in the computed error rates for the sequences considered (Section 2.5). The passage of DNA through the nanopore was modeled as unidirectional with the lifetime of each nucleotide in the sensor from an exponential distribution of known rate. Simulated data was produced by first generating an ideal pulse-train signal at 10 MHz for a chosen DNA sequence, where the dwell time for each nucleobase was randomly selected from the exponential distribution with mean 1 ms. White noise was added to the idealized signal, which was then low-pass Bessel filtered at 100 kHz and downsampled to 500 kHz. White noise variance, which we label as '1X noise', was chosen to produce a 2:1 signal-to-noise ratio (S/N) when the Bessel filter was set to 5 kHz bandwidth to emulate conditions comparable to those observed experimentally [12, 49]. Analysis of signals with 2X this noise, and without noise, was also performed. At 1X noise, the mean enzyme rate was also varied to examine its influence on error rate performance (Section 2.5).

2.2.2 Base-calling algorithm, alignment and consensus

Noise on each simulated signal was reduced by applying a running mean filter followed by a Savitzky-Golay filter of order 2. To identify ionic current levels, a custom step detection algorithm was employed using a gradient threshold to detect transitions between levels and amplitude thresholds to classify levels by

nucleobase. The number of nucleotides assigned to each current level was determined using a binning method to sort each level by its duration. The optimal sizes of the bins were chosen to maximize the sum of the probabilities that each current level is assigned the correct number of bases (Section 2.5). Error calculations were performed by comparing the predicted sequence to a known reference sequence. The current levels of the two sequences were globally aligned using the MATLAB function 'nwalign' with affine gap penalties. The numbers of nucleotides at each aligned level were compared and errors in the predicted sequence were classified as insertions, deletions, or substitutions. Insertions and deletions were counted on a per nucleotide basis, whereas substitutions were counted in terms of the number of current levels with misidentified amplitudes.

Multi-read consensus sequences were generated by first performing a progressive multiple alignment of the ionic current levels of the reads using the MATLAB function 'multialign' with the option 'TerminalGapAdjust' set to true. The multiple alignment was used to generate a consensus sequence of current levels using the MATLAB function 'seqconsensus' with the option 'Gaps' set to 'all'. Nucleotides were then assigned to the consensus sequence current levels using the optimal binning method, where the duration of the consensus levels were determined by computing the mean duration for each level. To ensure that the correct current levels were included in the calculation, each predicted sequence used to generate the consensus was globally aligned with the consensus sequence and only the durations of the aligned current levels were used for computing the mean duration times. Error analysis for the multi-read consensus sequences was performed in the same manner as for the single-read predicted sequences.

2.3 The potential for systematic channel errors

In the simplest case that the current amplitude is a function of only one nucleotide in the channel, a necessary and sufficient condition for recovering DNA sequence is that each letter generate a distinct amplitude that can be detected above experimental noise. When more than one nucleotide affects the current amplitude, it is less clear what sequence can be recovered. Consider the case where three nucleotides affect the current. If 64 distinct and detectable amplitude levels are generated for every triple-letter combination of the four nucleotides (i.e., 4^3), then there is no ambiguity in the identified sequence. On the other hand, if there are less than 64 detectable levels, there may or may not be ambiguity. Below, we present a generalized case in which there are an infinite number of sequences that can not be recovered, regardless of how many times the sequence is read.

Assume the current is a function of three nucleotides, and suppose the four triples in the set $\{CCC, CCA, CAC, ACC\}$ generate the same amplitude. Then, for any $n \geq 1$, there is a set of length-n subsequences $Z_1 \cdots Z_n$ constructed from A and C that cannot be distinguished from each other within the sequence $CCZ_1 \cdots Z_nCC$. As a specific example, within the sequence $\cdots TCCCACCACCG \cdots$, the subsequence $CACCA$ cannot be differentiated from $CCCCC$, $ACCAC$, or any other 5-letter combination of C and A in which As are separated by two or more Cs. A proof of the generalized statement, and comparable statements for cases when the amplitude is a function of two or four nucleotides, are provided in the Appendix A. Experimentally building a map from letters to amplitude is required to determine if such channel errors are present for a given nanopore. Practically, two sequences would be considered to have the "same amplitude" if the magnitude of the difference between the two amplitude levels has a S/N of less than 1.5 after applying the low-pass filters designed for signal-to-sequence conver-

sion (S/N 1.5 is a minimum threshold for idealizing the signal by Markov-based methods [83, 64], with S/N 2 or larger required for simpler methods [67], Appendix A). While additional low-pass filtering can always boost S/N and therefore improve discrimination between amplitude levels, too much filtering will result in excessive deletions of fast events. Thus, the effective filter bandwidth designed for optimal signal-to-sequence conversion will tradeoff S/N for detection time resolution. For the remainder of this work, we idealize the sequencing problem and assume there are no systematic channel errors. Specifically, the current amplitude is assumed to be a function of one nucleotide at a time (i.e., the channel is single-nucleotide sensitive).

2.4 Errors due to nondeterministic sensing times

By using an enzyme to control the motion of ssDNA through the nanopore [12, 49], the strand is temporarily immobilized for sensing before moving in single-nucleotide steps. A challenge for base calling is that the duration in each immobilized position is nondeterministic. For an ideal enzyme, we can model the duration as following an exponential distribution of known mean dwell time τ. We consider two complications with nondeterministic sensing times. First, without some signal that the enzyme has moved to the next position on the DNA, inferring the length of each detected subsequence is challenging, and in particular one expects errors to grow with the length of homopolymer regions. Second, when experimental noise requires the use of low-pass filtering to permit robust detection of each sequence-specific amplitude level, a fraction of sensing times are too fast for detection and result in an increase in deletions. We consider first the errors that are intrinsic to inferring the length of the sequence that corresponds to each detected amplitude level, and then the errors induced by adding noise to the idealized sensing signal.

The sensing problem is idealized by assuming that the current measurements are noiseless and sensitive to one nucleotide at a time. Specifically, each base (A, G, C, T) generates the current amplitude $(3, 2, 1, 0)$ pA, and each level is detectable regardless of duration (i.e., no durations are too fast since no noise filtering is required). A challenge is that we assume the enzyme does not provide an explicit tracking mechanism within the ionic current signal, which is consistent with the literature [12, 49], and so no new information can be extracted from the signal unless a sequence-specific amplitude shift is detected. This makes it difficult to identify the length of the sequence that corresponds to each detectable level. Mathematically, let τ_i be the duration during which the i-th nucleotide along the DNA is at the sensing position that determines the amplitude level. Each τ_i is an exponentially distributed random variable with mean τ. In our model problem, the transition from the i-th nucleotide to the $(i+1)$-th nucleotide being at the sensing position is detectable only if these two nucleotides are different (and thus yield different amplitude levels). Let s_j be the duration of the j-th segment along the time series of 4 distinct amplitude levels. If the sequence was entirely non-repeating, $\tau_i = s_i$ for all $i = 1, \ldots, n_t$, with n_t the length of the sequence. To quantify the challenge of inferring sequence length in general, consider the example of the sequence $TCCCAGG$ moving through the nanopore sensor starting from the right end. Sensing G first, we measure amplitude 2 pA for the duration $s_1 = \tau_1 + \tau_2$. Next, we measure A at amplitude 3 pA for the duration $s_2 = \tau_3$. Next, we measure C at amplitude 1 pA for the duration $s_3 = \tau_4 + \tau_5 + \tau_6$. Finally, sensing T we measure amplitude 0 pA for duration $s_4 = \tau_7$. The length of the sequence at each detected level must be inferred. For a single pass through the sequence, if s_4 gets a large sample value from the exponential distribution, it would appear that more than one T is present; likewise, if s_3 is made up of

three faster-than-average durations, it would appear that less than three Cs are present. Clearly, such random errors can be reduced only by repeatedly taking measurements of s_j for each level, and generating a consensus (average) time for that level from which the sequence length estimate is made.

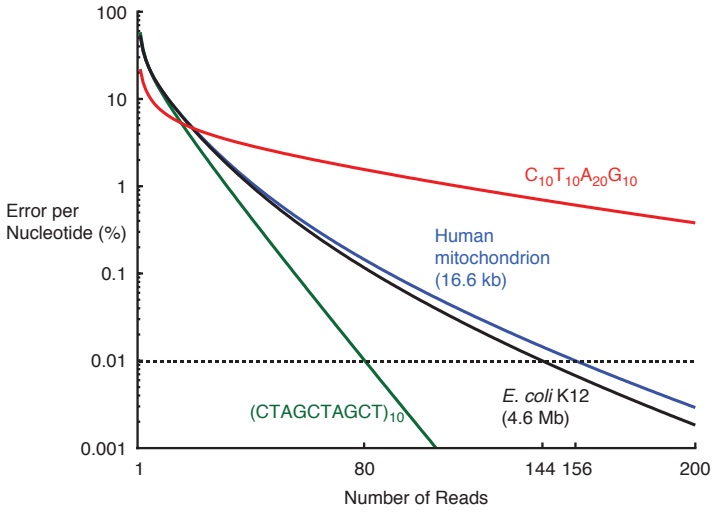

Figure 2.1: Analytic error rates for enzyme-controlled ssDNA nanopore sequencing. The idealization assumes a noiseless single-nucleotide sensor, but with no mechanism for tracking single-nucleotide displacements through homopolymer regions. Durations of each nucleotide in the sensor are from a single exponential distribution of known rate, consistent with an ideal enzyme controlling DNA motion through the sensor. The analytic error rates are computed for Human Mitochondrion [68], and for E. coli K-12 [5], for a 50 mer non-repeating sequence (green) and a 50 mer with length 10 and 20 mer homopolymer regions (red). Error reduction is accomplished only by rereading the same sequence and averaging the duration at each resolvable sequence-specific amplitude level.

We derive a time-binning strategy that estimates the length k of each sequence from the measured duration s_j at each nucleotide-specific amplitude level. Since each $s_j = \sum_{i=1}^{k} \tau_{i+i_0}$ is the sum of k independent samples of an exponentially distributed random variable, each s_j has a Gamma distribution. By rereading the sequence n times, denoting the measured set of durations $\{s_j^1, ..., s_j^n\}$, the estimate for sequence length (k_{est}) for each detected level is computed using the variable

$x = \left(\frac{1}{n}\sum_{l=1}^{n} s_j^l\right)/\tau$ with the simple equation

$$k_{\text{est}}(n) = \begin{cases} 1, & \text{if } x \leq b_1 \\ 2, & \text{if } b_1 < x \leq b_2 \\ 3, & \text{if } b_2 < x \leq b_3 \\ \vdots \end{cases} \tag{2.1}$$

with optimized bin values $(b_1, b_2, b_3, ...) = (1.472, 2.483, 3.488, ...)$ chosen to minimize the error rate (Appendix A). If the average s_j is between 2.483τ and 3.488τ for a detected level at 1 pA, for example, the estimated length is $k_{\text{est}}(n) = 3$ producing the sequence estimate CCC. Since the random variable $(x \cdot n)$ has a gamma distribution with shape parameter $(k \cdot n)$ and scale parameter 1, the error rate per nucleotide for a k-repeat based on measurements from n reads is $\text{Err}(k, n) = \frac{1}{k}\sum_j |j - k|\Pr(k_{est}(n) = j)$. This error rate has an analytical expression that can be computed in MATLAB using the incomplete gamma function (Appendix A). From this expression, the per-nucleotide error rate $g(n)$ is computed for any given sequence as a function of the number of reads n, and is given by the equation

$$g(n) = \frac{1}{n_t}\sum_{k=1}^{m} q_k \cdot \text{Err}(k, n) \tag{2.2}$$

where n_t is the length of the sequence, q_k is the total number of nucleotides belonging to length-k repeats in the sequence, and m is the longest repeat length present in the given sequence.

We computed error rates using equation (2.2) for four different sequences, including the 16.6 kb Human Mitochondrial DNA sequence [68], and for the 4.6 Mb *Escherichia coli* K-12 sequence [5] (Figure 2.1). The other two sequences are 50 nu-

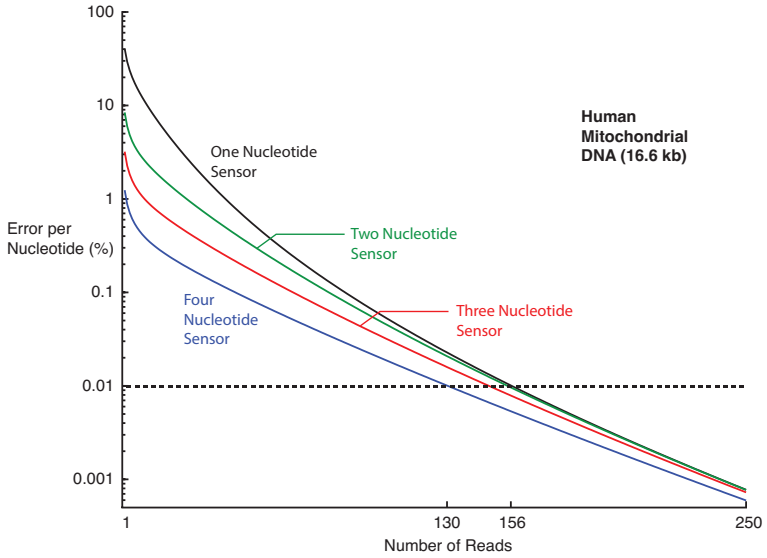

Figure 2.2: Effect of nanopore sensor footprint on the analytic error rates. The analytic error rates decrease with an increase in the size of the sensor footprint, but approach the single nucleotide sensor error rate as the number of reads increases.

cleotides in length and are used to show the influence of homopolymer length. Not surprisingly, the non-repeating sequence has the fastest rate of error decay and, as expected, longer stretches of homopolymer regions require a greater number of reads to reduce the error. With no mechanism for tracking the motion progress through homopolymer regions of ssDNA, rereading the same strand is required to reduce errors to acceptable levels. The figure suggests that to achieve the Q40 standard (99.99% confidence) requires reading known sequences over ∼150 times. Achieving multiple reads could be accomplished by single-pass reading of many copies in parallel in a multi-channel array, or by rereading the same strand at each pore [12]. When considering nanopore sensors that are a function of more than one nucleotide, the analytic error rate performance improves, but only if there are no systematic channel errors (Figure 2.2). Notably, the single-read error improves from 40.5% per nucleotide for a single-nucleotide sensor to 1.24% per nucleotide for a four-nucleotide sensor, for the 16.6 kb Human Mitochondrial DNA. The im-

provement is a byproduct of being able to detect the length of homopolymers that are the same length or shorter than the sensor footprint. The improvement is less dramatic, however, when higher accuracy is needed (the four-nucleotide nanopore sensor requires 130 reads for Q40 accuracy, Figure 2.2).

2.5 Influence of measurement noise and enzyme rate on base-calling errors

To consider next the effect of measurement noise on base-calling performance, we simulate ionic current signals. The mapping of bases (A, G, C, T) to the amplitude $(3, 2, 1, 0)$ pA was again used, with the sequence of the first 50 nucleotides of the Mitochondrial DNA sequence [68] used to generate each signal. For each signal, durations for each base were randomly drawn from an exponential distribution with mean $\tau = 1$ ms. A gradient-based algorithm was developed for level detection, and the time-binning strategy in equation (2.1) was used to assign the number of bases for each detected level. The reference sequence was used to compute the errors for each estimated sequence and for multi-read consensus sequences. To emulate experimental noise, a sufficient amount of white noise is added to the unfiltered ideal signal to produce $\sim$0.5 pA root-mean-square after low-pass Bessel filtering at 5 kHz bandwidth. This noise we label as '1X noise' and results in S/N of 2 between adjacent amplitude levels at 5 kHz bandwidth, which is sufficient for detection by standard methods [67] and by our gradient-based algorithm. The simulation makes use of a model of the nanopore instrument that has been experimentally validated [22], specifically by including the low-pass Bessel filter used in the current sensing amplifier. The Bessel filter is set at 100 kHz bandwidth and additional filtering is performed for robust level detection,

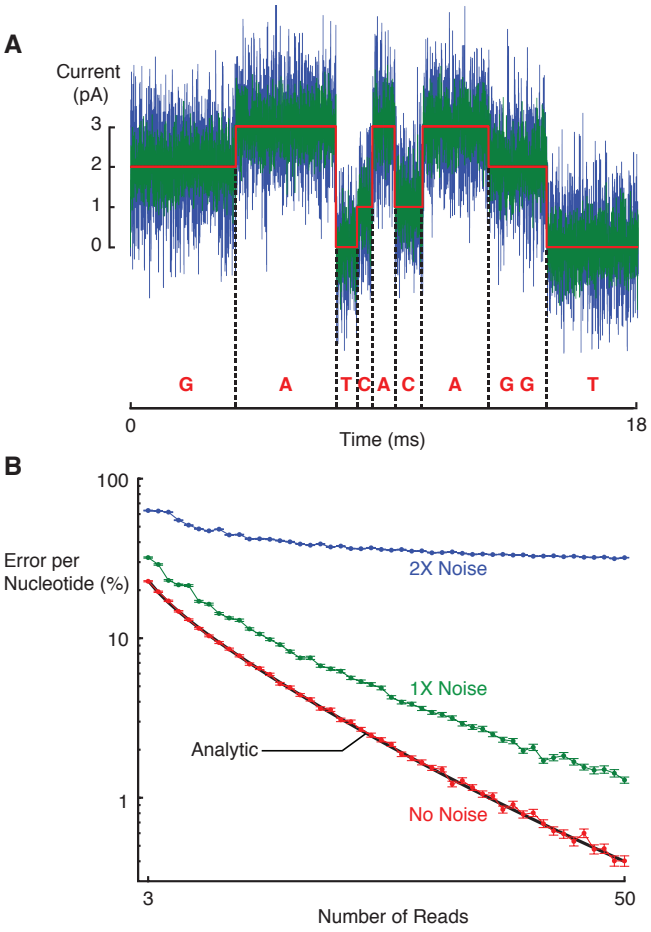

Figure 2.3: Base-calling logic applied to simulated nanopore signals shows error rate performance that matches the analytic error rate in the absence of noise, and increasing error rates with measurement noise. (A) Example signal traces for the first 10 nucleotides in the sequence, with no noise (red), 1X noise (green) and 2X noise (blue). The randomness of level durations shows the need for multiple reads to identify sequence lengths with confidence. (B) Mean error rates as a function of number of reads for the first 50 nucleotides of the Human Mitochondrial DNA sequence [68]. Data points are the mean error per nucleotide from 900 independent multi-read consensus sequences, with each consensus computed using the reported number of reads and with each read being drawn from a set of 10,000 simulated signals. Error bars are the standard error, computed as the standard deviation of the error divided by $\sqrt{900}$.

which is similar to what is done experimentally [49]. The case of 2X noise has two times the variance of the added white noise before filtering, and is also con-

sidered. We used the same base-calling logic and filter settings for 1X and 2X noise, though the filter settings were optimized for robust level detection at the 1X noise condition. The tradeoff in noise filtering and level-detection fidelity is central to sequencing error performance; therefore, settings would be optimized for the given S/N and time resolution constraints imposed by the instrument and enzyme rate in actual experiments. Example current traces show the difference between no noise, 1X and 2X noise (see Figure 2.3A). The error rates for multi-read consensus sequences are generated as a function of the number of reads, and compared to the analytic curve for the 50 nucleotide sequence (see Figure 2.3B).

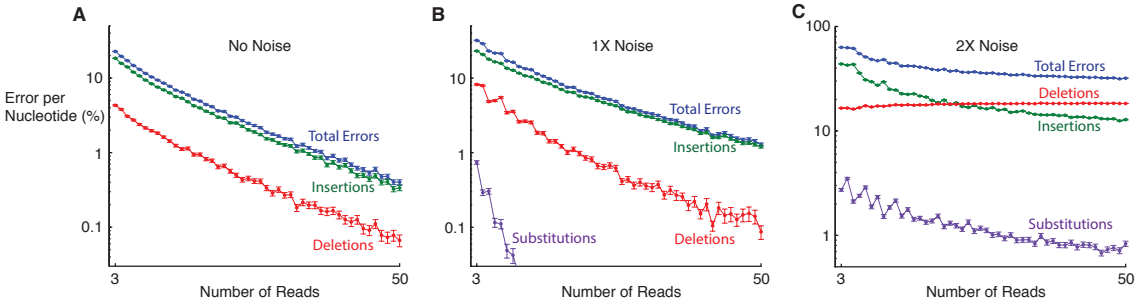

Figure 2.4: Breakdown of mean error rates into insertions, deletions, and substitutions. (A) Simulated nanopore signal with no additive noise. Insertions account for the majority of the total mean error rate and substitutions do not contribute at all. (B) Simulated nanopore signal with 1X noise. Like the no noise case, insertions account for the majority of the total mean error rate. Substitutions play a small role when the number of reads is few, but quickly decrease to zero. (C) Simulated nanopore signal with 2X noise. The additional noise results in a nearly equal contribution from insertions and deletions to the total mean error rate. Substitutions also play a larger role.

In the absence of noise, the error-rate performance of our computed consensus sequences matches the analytic trend exactly, validating our simulation and base-calling algorithm. The errors for both analytic and simulated (noiseless) trends are broken down as 82% insertions and 18% deletions on average, with no substitutions (Figure 2.4). The largest source of error is insertions because 68% of the nucleotides in this specific sequence are non-repeating, and only insertion

errors are possible in the noiseless case since every level is detectable. As the number of reads is increased, insertions become the dominant remaining error source. When 1X noise is added, the filtering required to reduce noise causes faster levels to go undetected, creating an increase in the fraction of deletions for each estimated sequence. Specifically, the fastest dwell that our level-detection method can robustly detect is 170 μs (Figure 2.5), and for a mean enzyme duration of 1 ms, $1 - 1/\exp(0.17/1) = 0.16$ or 16% of dwells are too fast for robust detection. The presence of noise can also cause fast levels to transiently appear at the wrong amplitude, resulting in substitutions. An increasing fraction of deletions and substitutions are observed in the error breakdown for consensus sequences at 1X noise, particularly for a low number of reads (for 3 read consensus, 72% insertions, 26% deletions, 2% substitutions, Figure 2.4). As the number of reads is increased, insertions become the dominant remaining error source, consistent with the noiseless case (Figure 2.4). At 1X noise, we varied the base-to-amplitude mapping to test if our original mapping choice was biasing the error rate performance with noise. The results show no significant difference in the error rate trends (Figure 2.6). When noise is further increased to 2X, a substantial new source of error is that spurious level-changes induced by the noise are detected, causing substitutions, deletions and incorrect calculation of durations. Insertions remain a large source of error (40-70%), and deletions become the greatest source of error for consensus sequences using more than 25 reads (Figure 2.4). While the base-calling performance at 2X noise is unacceptably bad, it should again be qualified that the filtering and base-calling logic was not re-optimized for the 2X noise case but kept the same as for the 1X noise case. Practically, both filtering and logic will be optimized according to the noise and level-detection performance of a given nanopore platform.

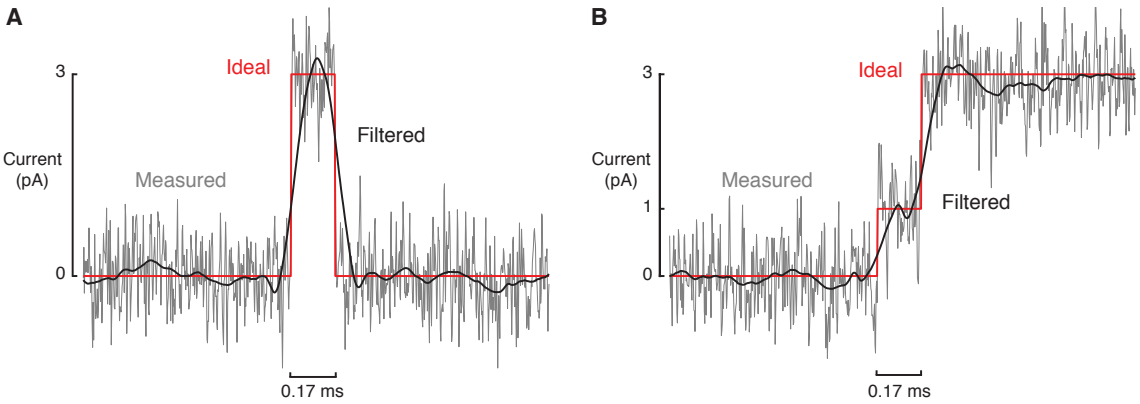

Figure 2.5: Worst case scenarios that affect the minimum dwell time for detecting ionic current levels. Simulation of a measured ionic current signal from nanopore experiments (grey), additionally filtered signal for step detection (black), and the noiseless ionic current levels (red). (A) A short ionic current level taking the form of a pulse in the measured signal is difficult to detect if its gradient is too steep, its peak too narrow, or its maximum amplitude occurs outside of the threshold. (B) A short intermediate ionic current level between two longer levels is difficult to detect if its gradient does not sufficiently flatted out.

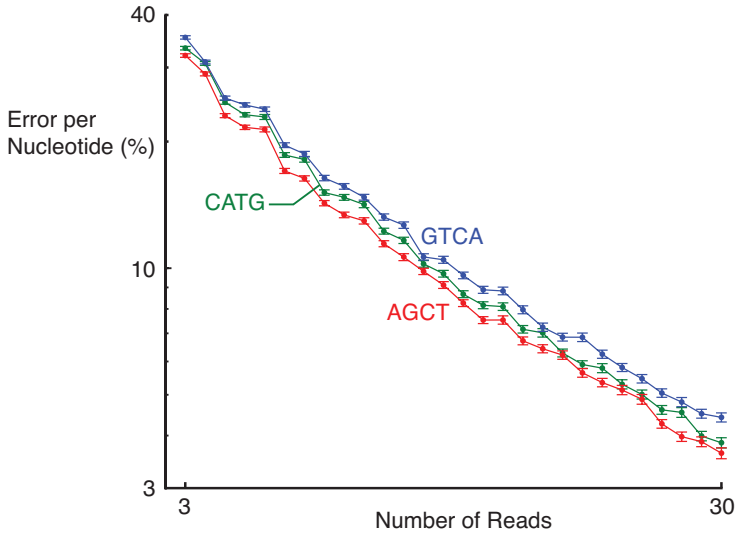

Figure 2.6: Effect of changing nucleobase amplitude mappings on mean error rate. Simulated nanopore signals with 1X noise. Amplitudes (in pA) assigned to bases decrease from left to right, i.e. for the curve CATG, base-amplitude mappings are C→3, A→2, T→1, and G→0. The curve AGCT reflects the base-amplitude mapping used in this work. Rearranging the amplitude mappings has virtually no effect on the mean error rate.

With an enzyme that follows a single exponential distribution, the mean dwell time τ will not be precisely known in practice. Using an estimate $\hat{\tau}$ of the true mean τ in the base-calling logic will incur errors. If $\hat{\tau} < \tau$, the length of each sequence will be overestimated causing insertion errors. Likewise, $\hat{\tau} > \tau$ will cause underestimation of sequence lengths and result in deletion errors. Additionally, the larger or smaller $\hat{\tau}/\tau$ becomes, the greater the errors. As an example, if $x = 1.3$ is computed with known τ and for a level corresponding to the single-nucleotide C, the estimated and correct length is $k_{\text{est}}(n) = 1$ from equation (2.1). However, if $\hat{\tau} = 0.85\tau$ is used to compute x, then it becomes $x = 1.53$ and equation (2.1) produces an insertion error with estimated length $k_{\text{est}}(n) = 2$. To assess the effect of incorrectly estimating τ on error rate performance, we considered two extreme cases at 1X noise: overestimating the mean dwell by double ($\hat{\tau} = 2\tau$), and underestimating the mean dwell by half ($\hat{\tau} = 0.5\tau$). The incorrect estimates for the mean were used in the calculation of $x = (\sum_{l=1}^{n} s_j^l)/(n\hat{\tau})$, which is used to compute the length estimate (k_{est}) of each sequence at each detected level in equation (2.1). As expected, overestimating the mean dwell creates deletion errors, with an error rate of 16% that persists even up to 30 reads (Figure 2.7). Error rate performance is considerably worse when underestimating the mean dwell time by half, with a persistent error over 100% that is comprised almost exclusively of insertion errors (Figure 2.7).

Assuming the mean enzyme dwell time τ is known, we considered also the effect of different τ values on the error rate performance, again using the 1X noise condition. The filtering required for robust amplitude-level detection at 1X noise results in an increasing fraction of levels that go undetected as τ is decreased, and error rates are considerably worse as τ decreases below 1 ms. On the other hand, our base-calling method applied to 1X noisy signals is observed to perform as well

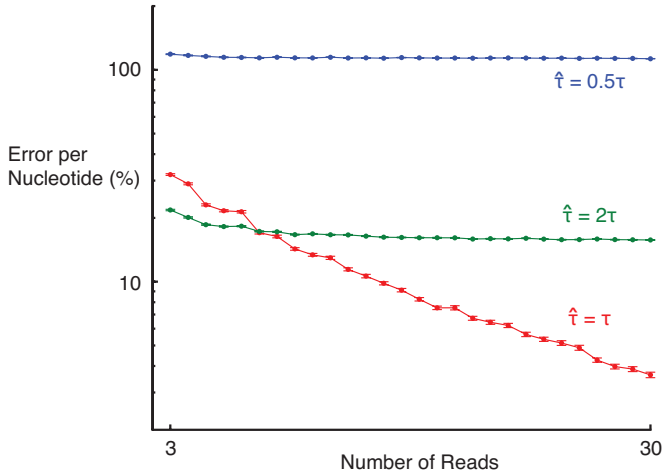

Figure 2.7: Effect of estimating the mean dwell time on the mean error rate. Simulated nanopore signals with 1X noise, a true mean dwell time of 1 ms, and varying estimates ($\hat{\tau}$) of the mean dwell time (τ) used for base-calling. Underestimating the mean dwell time results in more nucleotides being assigned to each ionic current level, which increases the number of insertions along with the mean error rate. Overestimating the mean dwell time results in fewer nucleotides being assigned to each ionic current level, which increases the number of deletions. This does not increase the mean error rate for a small number of reads because while the number of deletions is increased, the number of insertions is also decreased. Since insertions are the main drivers of the mean error rate, this actually improves the mean error rate for a small number of reads. In both cases, increasing the number of reads does little to improve the mean error rate.

as is theoretically possible (i.e., matching the analytic trends derived for noiseless signals) when $\tau > 10$ ms (Figure 2.8). The phi29 enzyme as a replication-driven ratchet has mean dwell $\tau = 36$ ms, computed as the reported 25 ms median dwell [12] divided by $\ln(2)$, but this does not suggest that the theoretically optimal error rate is achievable. Specifically, our idealization ignores backtracking that is experimentally observed with phi29, and the noise and channel sensing characteristics do not match our idealization. Nonetheless, $\tau = 36$ ms means that 99.5% of levels are resolvable if the setup can robustly detect 170 μs (Figure 2.5). Although a slower enzyme will reduce the number of deletion errors caused by filtering out fast events, it also means lower throughput. The viability of a commercial nanopore

sequencer will be determined by both the throughput and the error rates that are achievable, and these are a function of the scale of the multi-channel array that can be incorporated (fluidics, circuitry) into a platform of a given size [48].

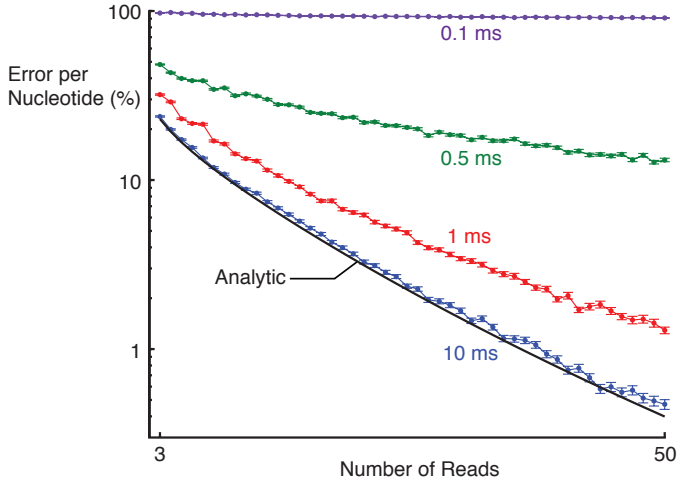

Figure 2.8: Effect of mean dwell time on the mean error rate. Simulated nanopore signals with 1X noise and varying mean dwell times. The mean error rates decrease with an increase in the mean dwell time and eventually converge to the analytic error rate.

2.6 Discussion

Our error analysis shows the need to reread the same molecule at each pore, or to read identical copies of the molecule serially or in parallel pores, when ionic current nanopore sequencing is used in conjunction with enzymes to control DNA motion. Systemic errors caused by the channel's inability to sense and differentiate specific sequences may or may not be present for a given pore. If present, such errors define an error rate threshold below which the platform cannot go, regardless of the number of reads. Random indel errors, on the other hand, can be reduced by increasing the number of reads, and we provided the first analytic expression that defines the best possible rate of error reduction as a function of

the number of reads.

Reduction in instrument complexity is an advantage for prospective nanopore devices that may trump any disadvantages associated with higher systematic error threshold or indel error frequency, though this will only become clear when such devices become available to users. Specifically, the prospective device presumably eliminates the need to build or amplify sequencing libraries, reduces the complexity of fluidics required during the sequencing operation (unlabeled nucleotides), and could make resequencing permissible with no fluid exchange. Even with the same raw error rate (5-15%) and read lengths (250 bp - 10 kbp) as Pacific Biosciences RS platform [91], a considerably less complex device can be much cheaper and portable. There is presently no "cheap, quick and dirty" sequencing technology; however, a hand-held nanopore sequencer may be such a technology. Even with modestly higher error rates, long read-length and portable sequencing platforms would undoubtedly find applications, e.g., for fast re-sequencing or targeted sequencing of pathogen strains [45], provided the user interface is as simple as other devices used routinely in clinical settings.

We conclude this work with a brief discussion on assigning error probabilities to sequences, as this is a forward-looking issue that will benefit from basic research as nanopore sequencing technologies come to market. Assigning a statistical measure of confidence to sequencing data is important for determining the suitability of sequencing results for a given application, and also for providing a quantitative basis for comparing data generated from different technologies [8]. The *de facto* metric for comparing the probability of error for a sequence across platforms is the position-specific quality score (Q-score). Quality scores originated with the base calling program *phred*, which uses an algorithm and a four-parameter set associated with the error characteristics of the Sanger method to compute the

score [19]. The accuracy of the quality scores has been key to the utility of Sanger sequencing data [30]. The quality scores currently reported for next-generation high-throughput sequencing techniques are on the same numerical scale as phred quality scores, but are not as accurate [46, 8, 30]. Quality scores are less accurate in part because the parameters derived for the Sanger sequencing method do not isomorphically (in a mathematical sense) capture error characteristics of the other sequencing methods. To identify an accurate metric of base quality for a nanopore sequencing method, appropriate parameters built on the base-call error characteristics of nanopore signals needs to be identified. More broadly, until a universal standard is developed for defining accuracy for next-generation sequencing, the value of combining sequence data from different technologies will not reach its full potential.

2.7 Acknowledgments

Hongyun Wang and William Dunbar contributed the mathematical framework for analytic validation of the simulation results and text detailing its derivation. I designed and ran the simulation experiments, developed and ran the code to analyze the results, and wrote the majority of the text. This work was supported by National Institutes of Health grant 1R21HG006561-01.

Chapter 3

Parameter Estimation

The following chapter describes excerpts from works published in *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* ([54]) and *Proceedings of the 51st IEEE Conference on Decision and Control* ([56]) for which I was first author.

3.1 Introduction

Early nanopore work used constant voltages to examine DNA and enzyme-bound DNA complexes [4], but more recently, the use of time-varying voltages has expanded the capabilities of the nanopore. For example, nanopore-DNA interactions [3] and polymerase-DNA interactions on the nanopore [90, 57] have been measured at the single molecule level using active control with step-changing voltages. Voltage ramps have been used for nanopore dynamic force spectroscopy, with the aim of modeling the molecular bond energy landscape [17]. With the assistance of custom hardware and filtering, sinusoidal voltage patterns have made it possible to monitor the presence of DNA in the pore at zero DC voltage [18, 40]. This application has the aim of producing a zero-force DNA sensor, in which (DC

bias) force on the DNA can be removed while still sensing the presence of the DNA in the pore.

A challenge with time-varying voltages is that the capacitive elements in the system contaminate the measured ionic current for a step voltage change. In prior work combining active control with nanopores, it was shown that this step-induced transient response limits the time-resolution for detecting DNA or DNA-protein dynamics [90], making it impossible to measure the population of molecular responses that are faster than the transient settling time (up to 30% in [90]). Sinusoidal voltages persistently excite the capacitive elements in the system, and thus continually mask the true value of the conductance. For generic time-varying voltages, we require the use of an estimator to recover the channel conductance as the parameter used to characterize the state of molecules captured in the nanopore. The following sections summarize previous work that developed a least-squares parameter estimation (LSPE) algorithm and a Kalman filter for estimating the nanopore channel conductance under voltage-varying conditions, including step and sinusoidal voltages, with the objective of inferring the channel conductance parameter as continuously as possible.

3.2 Least Squares Parameter Estimation (LSPE)

The classical method of least-squares is based around the concept that the unknown parameters of a dynamical system can be accurately estimated given a sufficiently accurate system model and a large enough set of observed response data. The method dictates that the best estimates of the unknown parameters are the most probable values that minimize the sum of the squares of the difference between the observed response in the data set and the predicted response of the system, hence the name 'least-squares'. When a system model is linear and

the noise on the observed data is uncorrelated with a constant variance and a mean of zero (i.e. white noise), the Gauss-Markov theorem states that the least-squares estimator is the best linear unbiased estimator of the unknown parameter values. The modeled approximation of the biological nanopore system fulfills these requirements making the method of least-squares a good candidate for estimating the conductance of the nanopore channel. The LSPE algorithm summarized in this section is shown through simulations to provide efficient online estimation of the channel conductance during step-changing voltages, and continuous estimation during sinusoidal voltage inputs, with realistic noise superimposed on the data.

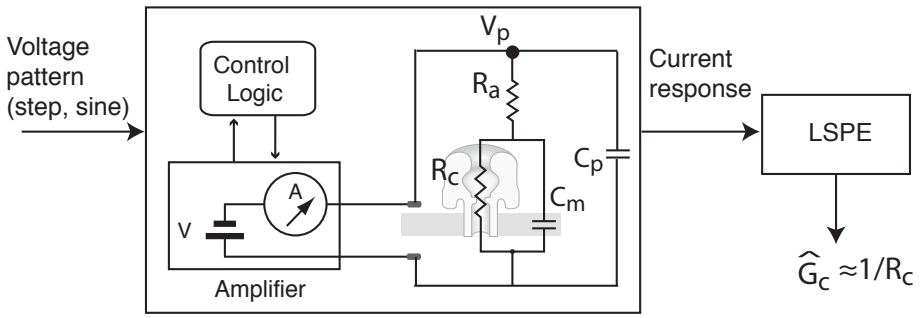

Figure 3.1: An amplifier applies voltage and measures the ionic current through the nanopore channel. Control logic is used to monitor the current and control the input voltage pattern. The known input signal and the measured current response are used by the LSPE algorithm to estimate $\widehat{G}_c \approx G_c = 1/R_c$, the conductance of the nanopore channel. In the circuit model of the system, R_c is the resistance of the channel, C_m and C_p are the membrane and parasitic capacitances, respectively, V_p is the voltage at the output of the amplifier, and R_a is the electrolytic access resistance.

3.2.1 Nanopore System Model

The four-state model of the biological nanopore system in the Laplace domain has the transfer function $H(s)$ from the input voltage $V(s)$ to the output current

$I(s)$ given by

$$H(s) = \frac{C_\Sigma s + G_c}{a_1 s^4 + a_2 s^3 + a_3 s^2 + a_4 s + 1} \tag{3.1}$$

where $C_\Sigma = C_p + C_m$ (pF) is the combined (membrane and parasitic) capacitance of the system, G_c (nS) is the channel conductance of the nanopore and a_1, a_2, a_3 and a_4 are characteristic of the Bessel filter. For consistency of units, time is in milliseconds and frequency is in kHz. We can ignore R_a in the model since it is negligible ($\sim 10^{-4}$ GΩ) compared to R_c (3 GΩ). In another work, we have used system identification tools to validate this model with experimental data [22]. The Bessel filter variables are defined in terms of the 4^{th}-order reverse Bessel polynomial coefficients and the -3 dB cutoff frequency f_c as

$$(a_1, a_2, a_3, a_4) = \frac{(1, 10f, 45f^2, 105f^3)}{105f^4} \tag{3.2}$$

$$\text{with} \quad f = \frac{2\pi f_c}{2.113917675},$$

where the denominator constant was identified to move the -3 dB cutoff frequency of the filter to f_c. The continuous-time state space representation of equation (3.1) in control canonical form is

$$\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t); \;\; t \geq 0 \tag{3.3}$$

with column vector $x = [x_1; x_2; x_3; x_4]$ and matrices

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{1}{a_1} & -\frac{a_4}{a_1} & -\frac{a_3}{a_1} & -\frac{a_2}{a_1} \end{bmatrix},$$

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} \frac{G_c}{a_1} & \frac{C_\Sigma}{a_1} & 0 & 0 \end{bmatrix}.$$

In the simulations in section 3.2.3, white noise is added to u and y (with different variances). The system model equation (3.3) and LSPE algorithm can be extended to incorporate explicit models of noise (white or colored), with such noise models being experimentally identified. This extension is not done here for brevity.

3.2.2 Least-Squares Parameter Estimation Algorithm

To construct the LSPE algorithm and simulate the response of the nanopore system the continuous-time model represented by equation (3.3) is discretized and converted into discrete-time using the delta operator form. Discretization beings with the solution to (3.3)

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau$$
$$y(t) = Ce^{At}x(0) + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau.$$

The sample period Δ defines sample times $t_k = k * \Delta$. The input signal is assumed to be piece-wise constant between the sample times: $u(t) = u(t_k)$ for

all $t \in [t_k, t_{k+1})$. Using this, the continuous solution is converted to discrete time form as

$$x(t_{k+1}) = A_{\mathrm{d}} x(t_k) + B_{\mathrm{d}} u(t_k), \quad y(t_k) = C_{\mathrm{d}} x(t_k), \tag{3.4}$$

with $A_{\mathrm{d}} = e^{A\Delta}$, $B_{\mathrm{d}} = \left(\int_0^{\Delta} e^{A(\Delta-\tau)} d\tau\right) B$, and $C_{\mathrm{d}} = C$. The matrix A is invertible, so the matrix B_{d} can be rewritten as $B_{\mathrm{d}} = A^{-1}\left(e^{A\Delta} - I\right) B$.

Equation (3.4) is the traditional discrete time shift operator form, which models the absolute displacement of the state vector from sample to sample, whereas equation (3.3) models the infinitesimal increment of the state vector defined by the time derivative. This underlying characteristic of the continuous time state-space equations is more accurately modeled in discrete time using the delta operator form [23]. Also known as the divided difference operator form, the delta operator form models the change in the absolute displacement of the state vector from sample to sample over a given sample period. Using the delta operator, the discrete time state-space model takes the form

$$\left.\begin{aligned} x_\delta(t_k) &= A_\delta x(t_k) + B_\delta u(t_k) \\ x(t_{k+1}) &= x(t_k) + x_\delta(t_k)\Delta \\ y(t_k) &= C_\delta x(t_k), \end{aligned}\right\} \tag{3.5}$$

with $A_\delta = (A_{\mathrm{d}} - I)/\Delta$, $B_\delta = B_{\mathrm{d}}/\Delta$, and $C_\delta = C_{\mathrm{d}} = C$.

Algebraically, the sampled output can be written in terms of the system parameters, the state vector and the initial condition by recursively evaluating equation

(3.5). Beginning with t_1, the solution of the sampled output at t_n takes the form

$$y(t_n) = \frac{G_c}{a_1}\left[x_1(t_0) + \sum_{i=0}^{n-1} x_{\delta,1}(t_i)\Delta\right] + \frac{C_\Sigma}{a_1}\left[x_2(t_0) + \sum_{i=0}^{n-1} x_{\delta,2}(t_i)\Delta\right] \tag{3.6}$$

The matrix expression of interest that relates the output to the system parameters G_c and C_Σ can now be defined as

$$\begin{bmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_n) \end{bmatrix} = \left[\begin{array}{c|c} Q_1 & Q_2 \end{array}\right] \begin{bmatrix} G_c/a_1 \\ \\ C_\Sigma/a_1 \end{bmatrix}$$

with

$$Q_1 = \begin{bmatrix} x_1(t_0) + x_{\delta,1}(t_0)\Delta \\ x_1(t_0) + x_{\delta,1}(t_0)\Delta + x_{\delta,1}(t_1)\Delta \\ \vdots \\ x_1(t_0) + \sum_{i=0}^{n-1} x_{\delta,1}(t_i)\Delta \end{bmatrix}$$

and

$$Q_2 = \begin{bmatrix} x_2(t_0) + x_{\delta,2}(t_0)\Delta \\ x_2(t_0) + x_{\delta,2}(t_0)\Delta + x_{\delta,2}(t_1)\Delta \\ \vdots \\ x_2(t_0) + \sum_{i=0}^{n-1} x_{\delta,2}(t_i)\Delta \end{bmatrix}$$

which is written in vector notation as

$$y^{1,n} = Qz$$

where the matrix $Q = [Q_1\ Q_2] \in \mathbb{R}^{n\times 2}$ and the column vector $z = [G_c/a_1;\ C_\Sigma/a_1] \in \mathbb{R}^2$.

The least-squares approximation problem is based upon finding the best estimate $\hat{z}$ of the vector z that minimizes

$$\|Qz - y^{1,n}\|^2$$

where $\|\cdot\|$ represents the Euclidean norm. Since the matrix Q has more rows than columns and has full column rank, the least-squares approximation problem has a unique solution [6] in the form

$$\hat{z} = (Q^T Q)^{-1} Q^T y^{1,n}.$$

Once the least-squares solution $\hat{z}$ is computed, the estimates of the channel conductance and the system capacitance are $[\widehat{G}_c;\ \widehat{C}_\Sigma] = \hat{z} * a_1$.

The channel conductance of the nanopore changes when DNA is captured and translocates through the nanopore. These capture events occur on a micro-to-millisecond time scale [4]. Thus, the LSPE algorithm must be able to estimate changes in G_c on these time scales. This is accomplished through sequential implementation of the algorithm on overlapping windows of length n that span the input and output data sets of length N, where $N \gg n$.

3.2.3 Simulations

The performance of the LSPE algorithm was tested in simulations with step-changing and sinusoidal voltages. To emulate realistic experimental conditions, white noise was added to the input (0.2 mV RMS) and filtered output (1.5 pA RMS) with variances close to those observed experimentally [90] (noise is white up to 10 kHz bandwidth). Also, the value of G_c was set to 1/3 nS for positive voltages and 2/9 nS for negative voltages, consistent with values for experiments performed in 0.3 M KCl buffered solution [90]. The performance of the LSPE algorithm is compared here to the performance of a simple 'I/V method', defined as estimating the conductance by $I_p(t_k)/V_p(t_k)$ at each sample time t_k. When voltage is constant, the current is constant unless changes in G_c occur, for example, if DNA is captured in the nanopore or polymerase bound to DNA dissociates from the DNA [90, 57]. Thus, when V_p is constant for a sustained period, the I/V method produces an accurate estimate for G_c. To be of value in estimating G_c, the LSPE should perform comparably to the I/V method when V_p is constant, and outperform the I/V method when V_p is time-varying.

3.2.3.1 Step-Changing Input

For a step-changing input, the output current stays constant except when the input transitions from one level to another. The switching of the input voltage produces a transient response in the output current the duration of which is dependent on the amplitude of the input voltage, the amount of capacitance in the system C_Σ, and the Bessel filter cutoff frequency f_c. For this work, a cutoff frequency of 1kHz was used to provide a sufficiently long settling time to test the LSPE algorithm without contaminating the signal with too much noise.

At 1 kHz bandwidth, 250 kHz sample rate and without noise, the step-response

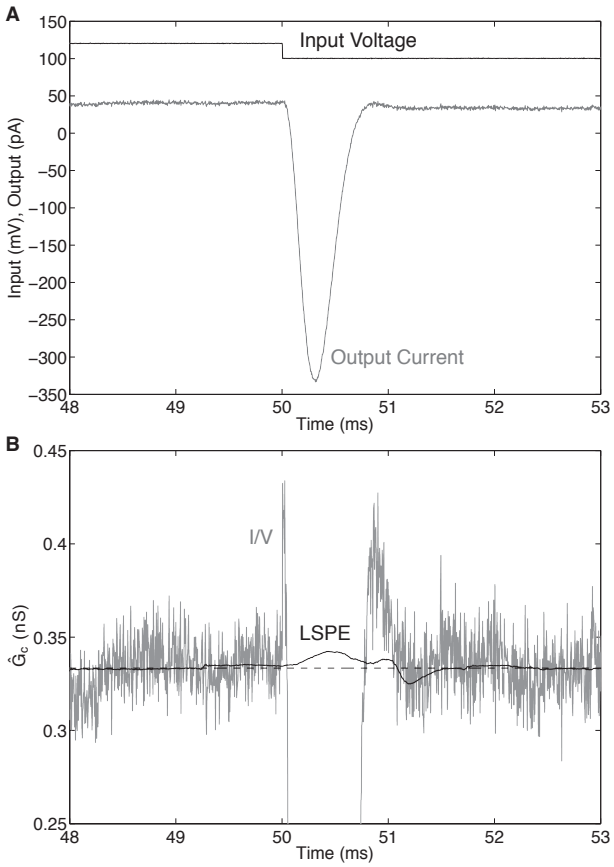

Figure 3.2: (A) Voltage step response (120 to 100 mV) of the nanopore system model. (B) A comparison of the LPSE and I/V methods for generating $\widehat{G}_c$. The I/V method has a larger steady-state standard deviation (1.36×10^{-2} nS) and a much larger overshoot (3.669 nS) in response to a step change than the LSPE algorithm (7.927×10^{-4} nS and 9.708×10^{-3} nS).

settling time of the LSPE estimate of $\widehat{G}_c$ is 0.996 ms, compared to 1.412 ms for the I/V method. That is, the LSPE estimate converges faster (70%) than the output current does. Practically, capacitance compensation on the recording amplifier can speed the current settling time (and thus the I/V method's estimate). However, the I/V method with a compensated current will, in general, not work in both step and sinusoidal conditions without heuristic tuning of the compensation settings for each set of conditions (voltage pattern, bandwidth), while the LSPE algorithm

works universally.

The performance of the LSPE algorithm for step voltages is shown in Figures 3.2 and 3.3. The LSPE algorithm outperforms the I/V method by producing an estimate of $\widehat{G}_c$ with a smaller standard deviation. One might argue that the LSPE algorithm is simply acting as a filter, and the performance of the I/V method could be improved if the current were first filtered. In fact, the LSPE algorithm is not a filter but an estimator, recursively computing the value of $\widehat{G}_c$ that minimizes the error between the measured current and current modeled by the discrete-time form of equation (3.3). Although additional low-pass filtering of the current would reduce the standard deviation of the I/V estimate, the filter would further increase the settling time of the estimate.

3.2.3.2 Sinusoidal Input

For a sinusoidal voltage input, the output current is constantly in a transient state, with the capacitive elements in the system being persistently excited. This has a positive effect on the LSPE algorithm in that once $\widehat{G}_c$ converges, it does not diverge again even though both input and output signals are non-constant.

The I/V method does not produce accurate values of $\widehat{G}_c$ for sinusoidal voltages, as expected, but we report the results here for comparison. The performance of the LSPE algorithm for sinusoidal input voltages is shown in Figures 3.4 and 3.5. In Figure 3.4, $G_c = 1/3$ nS since the input stays positive. The I/V estimate has a large standard deviation and follows a 10 Hz sinusoidal pattern of the measurements, never converging to G_c. The I/V estimate briefly reaches the true value of G_c only at the peaks of the sinusoidal input voltage since these are the locations where the current and voltage are momentarily constant. This also holds for a sinusoidal input that changes polarity, shown in Figure 3.5. The

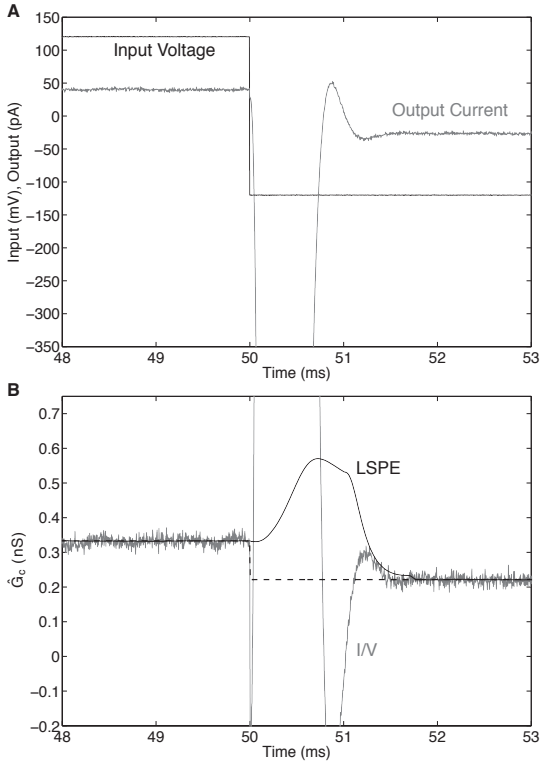

Figure 3.3: (A) Voltage step response (120 to −120 mV) of the nanopore system model. (B) A comparison of the LPSE and I/V methods for generating $\widehat{G}_c$. The voltage sign change at 50 ms causes a step change in G_c from 1/3 to 2/9 nS. The two methods have comparable settling times, with the LSPE algorithm having a smaller steady-state standard deviation (8.898×10^{-4} nS) and overshoot (0.349 nS) than the I/V method (1.34×10^{-2} nS and 36.57 nS).

change in polarity results in a step change in G_c, which the LSPE algorithm tracks well (Fig. 3.5). The LSPE estimate is noisier than when the voltage maintains a constant polarity (Fig. 3.4), but $\widehat{G}_c$ remains centered around the true values of G_c (1/3 nS and 2/9 nS), whereas the I/V estimate ranges between 3.6×10^3 nS and -2.1×10^{-4} nS.

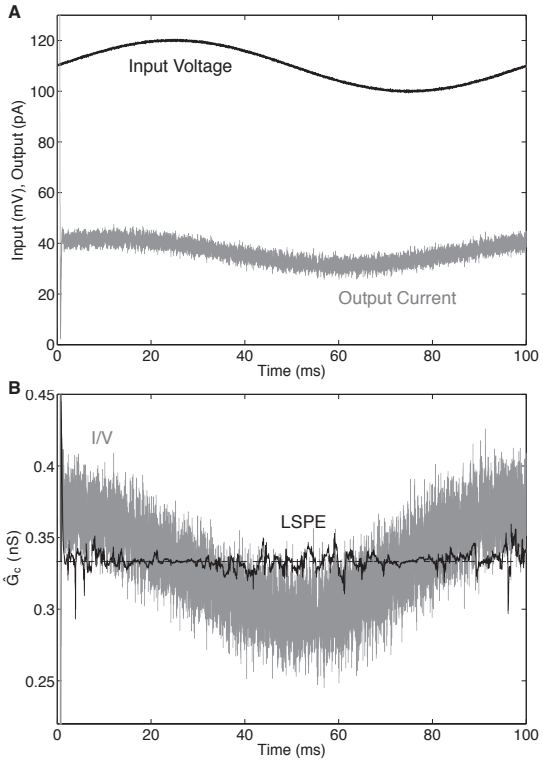

Figure 3.4: (A) Sinusoidal voltage response (10 mV peak-to-peak, 10 Hz, 110 mV DC offset) of the nanopore system model. (B) A comparison of the LPSE and I/V methods for generating $\widehat{G}_c$. The I/V method's estimate has a larger standard deviation (2.8×10^{-2} nS) than the LSPE algorithm (5.4×10^{-3} nS) and does not generate accurate estimates.

3.2.4 Discussion

The LSPE algorithm presented in this section provides a reasonably accurate means for estimating the channel conductance of a nanopore under voltage-varying conditions. The algorithm consistently achieves better performance (in terms of convergence time and standard deviation of the estimate) than the simple I/V method for both step-changing and sinusoidal input voltages. Since variance is improved, DNA or DNA-protein events that can be detected by the measured current (i.e., there is sufficient single-to-noise ratio) are easier to detect through

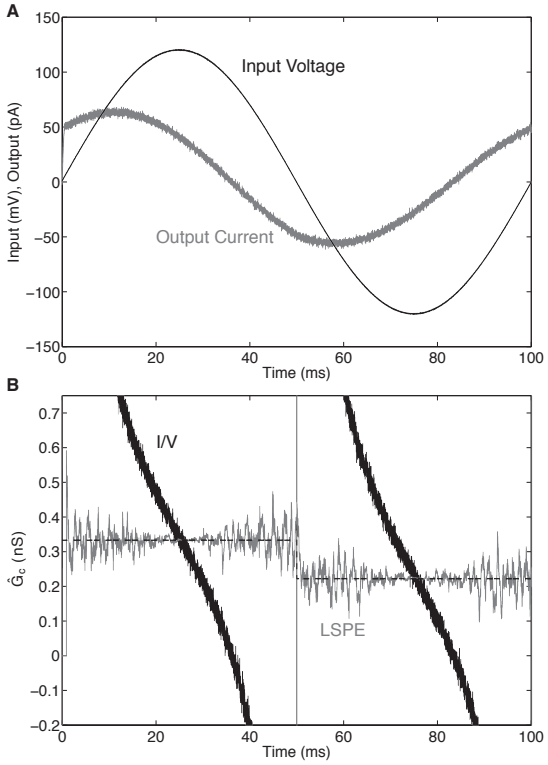

Figure 3.5: (A) Sinusoidal voltage response (120 mV peak-to-peak, 10 Hz, 0 mV DC offset) of the nanopore system model. (B) A comparison of the LPSE and I/V methods for generating $\widehat{\widehat{G}}_c$. The voltage sign change at 50 ms causes a step change in G_c from 1/3 to 2/9 nS. The I/V method does not generate accurate estimates, whereas the LSPE algorithm does track the change in G_c.

the use of the LSPE algorithm.

For this initial effort, we focused on an online implementation that uses fixed-length windows of past data to generate the estimated conductance value. Future work will explore improving the algorithm's performance by varying the window length based on detected rates of change of the data [35], and by incorporating forgetting-factors in the sequential implementation [44]. Also, an offline implementation that makes use of future windows to compute the estimate can be developed to further improve the detection resolution of rapid DNA-protein dissociation events that follow voltage changes in active control experiments [90],

[57].

The cited advantage of AC-signal detection (absent DC bias) is that nanopore/analyte interactions can be measured while reducing the effects of electroosmosis, electrophoresis, and protein deformation that accompany large DC biases [18]. In [18], custom hardware (lock-in amplifier) and software permit high frequency (10–20 mV, 1–2 kHz f_w) sinusoidal voltage inputs. The LSPE derived here cannot track G_c at sinusoidal frequencies above 50 Hz (data not shown). Future work will explore if and how well the LSPE estimate may track the presence of DNA in the pore at sinusoidal voltages around 0 mV (no DC bias), at 5–50 Hz frequencies, as an alternative to the high frequency method in [18].

3.3 Kalman Filtering

In the previous section we derived a simple least-squares parameter estimator (LSPE) to recover the conductance of the nanopore channel. While the LSPE algorithm performed better than the simple method of dividing current by voltage, its performance can be improved upon, especially for sinusoidal voltages. In this section we summarize the development of a Kalman filter estimator that more accurately recovers the open nanopore channel conductance during time-varying voltages. The filter is tested first in simulations with realistic process and measurement noise, and then on nanopore experiment data using sinusoidal voltage inputs. The filter is shown to recover the step changes in open channel conductance that occur when voltage changes polarity with a high degree of accuracy.

3.3.1 Nanopore System Model

The same transfer function for the four-state model of the biological nanopore system described in section 3.2.1 is used for developing the Kalman filter. This model was chosen because it is the simplest model that correctly describes the nanopore system and eliminates G_c and C_Σ from the dynamics matrix. Unlike in the LSPE algorithm, the transfer function is transformed from the frequency domain into state space using the observer canonical form instead of the control canonical form. This form makes it possible for the system model to remain linear when the parameters G_c and C_Σ are added to the state vector.

3.3.1.1 Extension of the system model

The method of Kalman filtering can be used to estimate the states of a linear stochastic system [29]. In order to use this method to estimate the system parameters G_c and C_Σ, the state vector x must be extended to include the parameters as states. This results in an extended system model of the form

$$\dot{x}_\mathrm{e}(t) = A_\mathrm{e}x_\mathrm{e}(t) + B_\mathrm{e}u(t) + w, \quad y(t) = C_\mathrm{e}x_\mathrm{e}(t) + v \tag{3.7}$$

with column vector $x_\mathrm{e} = [x_1\ x_2\ x_3\ x_4\ G_c\ C_\Sigma]^T$ and matrices

$$A_\mathrm{e} = \begin{bmatrix} A & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_\mathrm{e} = \begin{bmatrix} B \\ 0 \\ 0 \end{bmatrix}$$

$$\text{and} \quad C_\mathrm{e} = \begin{bmatrix} C & 0 & 0 \end{bmatrix}.$$

Once G_c and C_Σ are included in the extended state vector, equation (3.7) becomes nonlinear since G_c and C_Σ are also parameters in B_e. This is corrected by rearranging the system equations to get

$$\dot{x}_\mathrm{e}(t) = \Phi(t)x_\mathrm{e}(t) + w, \quad y(t) = C_\mathrm{e}x_\mathrm{e}(t) + v \tag{3.8}$$

where

$$\Phi(t) = \begin{bmatrix} 0 & 0 & 0 & -\dfrac{1}{a_1} & \dfrac{u(t)}{a_1} & 0 \\ 1 & 0 & 0 & -\dfrac{a_4}{a_1} & 0 & \dfrac{u(t)}{a_1} \\ 0 & 1 & 0 & -\dfrac{a_3}{a_1} & 0 & 0 \\ 0 & 0 & 1 & -\dfrac{a_2}{a_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

As with the LSPE algorithm, the continuous-time model of the system represented by equation (3.8) is discretized and converted into discrete-time. However, this time the typical shift operator form is used for simplicity.

3.3.2 Kalman Filter

The Kalman filter is the statistically optimal estimator for estimating the state of a linear nondeterministic system [24]. It solves the problem of optimal linear filtering by recursively calculating an estimate of a system's state using the previous state estimate along with knowledge about the system dynamics, noise statistics and measured input and output data. This recursive process can be broken down into two steps: prediction and update. In the prediction step, the a

priori state estimate $\hat{x}^-$ is extrapolated using the state transition matrix Φ and the *a posteriori* state estimate $\hat{x}^+$ from the previous sample time such that

$$\hat{x}^-(t_k) = \Phi(t_{k-1})\hat{x}^+(t_{k-1}). \tag{3.9}$$

The *a priori* error covariance matrix P^- is also extrapolated using Φ along with the *a posteriori* error covariance matrix P^+ and the process noise covariance matrix Q from the previous sample time such that

$$P^-(t_k) = \Phi(t_{k-1})P^+(t_{k-1})\Phi(t_{k-1})^T + Q(t_{k-1}). \tag{3.10}$$

The initial values for the *a priori* state estimate are chosen to represent the expected values for the nanopore channel conductance and system capacitance for a given set of experimental conditions. The initial values for the *a priori* error covariance matrix are chosen to put a greater emphasis on driving the estimates of G_c and C_Σ to their true values more quickly than the other states.

In the update step, the Kalman gain K as well as $\hat{x}^+$ and P^+ are obtained from the equations

$$\begin{aligned} K(t_k) &= P^-(t_k)C^T\left[CP^-(t_k)C^T + R\right]^{-1} \\ \hat{x}^+(t_k) &= \hat{x}^-(t_k) + K(t_k)\left[y(t_k) - C\hat{x}^-(t_k)\right] \\ P^+(t_k) &= \left[I - K(t_k)C\right]P^-(t_k), \end{aligned} \tag{3.11}$$

which utilize the predicted *a priori* values as well as the measurement sensitivity vector and measured output data. Also during this step, the process noise coupling and covariance matrices are updated to reflect the newly calculated *a posterior* state estimates.

3.3.2.1 Change detection

In voltage-varying experiments, the open nanopore channel conductance is observed to undergo a step-like change when the input voltage switches polarity. When DNA is in the pore, step-like changes are also observed, signaling the dissociation of a protein from DNA [90, 4], or the motion of DNA through the pore driven by the catalytic action of an enzyme [12, 57]. In order to track and estimate the changing conductance with the Kalman filter, it is necessary to reset the *a priori* error covariance matrix each time a change is detected. This is because P^- is the only time-varying parameter used to calculate the Kalman gain, which is what drives the state estimate to the correct value.

To perform change detection, we uses a two-sided cumulative sum (CUSUM) hypothesis test. The CUSUM algorithm uses a distance measure combined with a stopping rule to determine when a change in parameters has taken place [25]. The distance measure is used to quantify the error between the actual system output and the estimated system output at each sample time. The stopping rule gives an alarm when the distance measure becomes too large. This alarm is the signal to the Kalman filter that a change has been detected, so P^- is reset to its initial condition.

3.3.3 Simulations

The discrete-time four-state nanopore system model and Kalman filter were implemented in MATLAB to test the performance of the Kalman filter in simulations with sinusoidal voltages. The Kalman filter was tested over a range of input frequencies and amplitudes in an effort to characterize the effect of the input parameters on the performance of the estimator. Random zero-mean Gaussian noise with a covariance of 0.2 mV^2 and 1.5 mV^2 was added to the input and output re-

spectively to emulate realistic experimental conditions [90]. The true values of G_c used (1 nS for positive voltages and 0.876 nS for negative voltages) were chosen to be consistent with experimentally observed values in 1 M KCl buffered solution, derived by fitting a slope to I-V curve data recorded over a set of constant voltages (data not shown). The input voltage and output current were sampled at 250 kHz and the Bessel filter cutoff frequency was set to 5 kHz, values equal to those used in experiments. The performance of the Kalman filter was assessed on the basis of the root-mean-squared (RMS) error of the G_c estimates taken just before voltage changes polarity after the Kalman filter has converged.

3.3.3.1 Results

An example sinusoidal input, with a frequency of 1 Hz and a peak-to-peak amplitude of 200 mV, and simulated output are shown in Figure 3.6A. At this frequency, the capacitance is barely excited, and the current and voltage are nearly in phase. The continuous estimate of G_c produced from this data is shown in Figure 3.6B.

The Kalman filter performs best for inputs with higher amplitudes and lower frequencies. As shown in Figure 3.7, the RMS estimation error for a frequency of 100 Hz dropped from 7.049×10^{-2} nS to 8.109×10^{-4} nS as the amplitude was increased from 1 mV to 100 mV. This shows a nearly linear dependence between the RMS estimation error and amplitude. Higher amplitudes necessarily create a larger signal-to-noise ratio (SNR) making it easier for the Kalman filter to detect and estimate changes in the system's state as the amplitude increases. The dependence between frequency and the RMS estimation error, also shown in Figure 3.7, is more quadratic with the RMS estimation error for an amplitude of 1 mV only increasing from 2.550×10^{-3} nS to 7.049×10^{-2} nS as the frequency

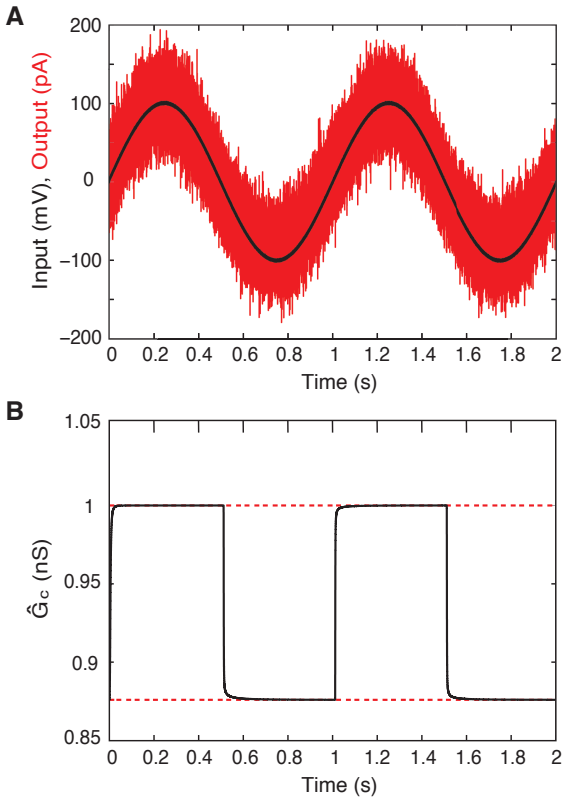

Figure 3.6: (A) Sinusoidal voltage response (200 mV peak-to-peak, 1 Hz) of the simulated nanopore system. (B) Kalman filter estimation of G_c, the nanopore channel conductance. The Kalman filter is able to produce accurate estimates of the nanopore channel conductance with an RMS estimation error as small as 5.788×10^{-7} nS.

was increased from 1 Hz to 100 Hz.

At lower frequencies, the value of G_c changes more slowly providing the Kalman filter with more data points to use for estimation and change detection between values. Therefore, the estimator performs better at lower frequencies than at higher frequencies. When the input was set to the lowest frequency (1 Hz) and the highest amplitude (100 mV) to achieve maximum performance (Figure 3.6A), the Kalman filter was able to produce estimates of G_c (Figure 3.6B) with an RMS estimation error of only 5.788×10^{-7} nS.

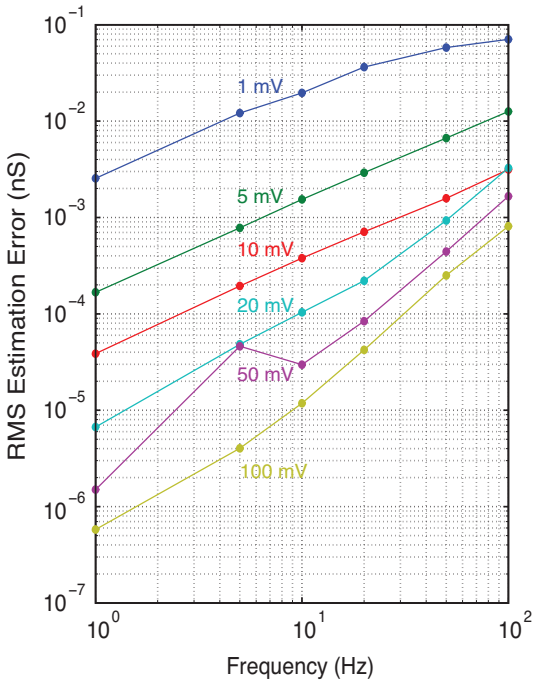

Figure 3.7: The RMS estimation error as a function of the input frequency for a range of different amplitudes using simulated data. The RMS estimation error shows an inverse relationship and nearly linear dependence on amplitude and a direct relationship and a nearly quadratic dependence on the frequency.

3.3.4 Experiments

To validate the results obtained using simulated data, the Kalman filter was applied to real experimental data gathered from the nanopore device. The experiments were conducted in 1 M KCl buffered solution, in the absence of DNA molecules, and with all other conditions identical to the simulations. The performance of the Kalman filter was also measured in terms of the RMS estimation error of the G_c estimates for both amplitude and frequency in the same way as in the simulations using known values of G_c.

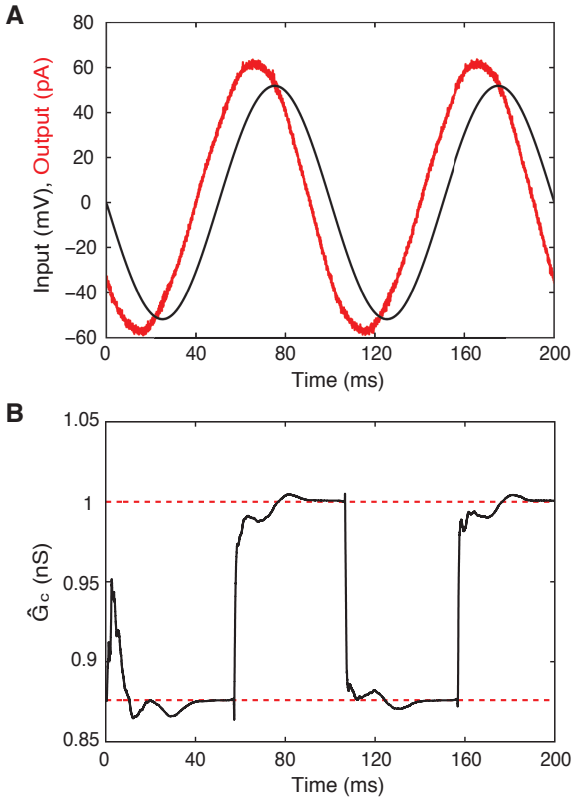

Figure 3.8: (A) Sinusoidal voltage response (100 mV peak-to-peak, 10 Hz) of the actual nanopore system. (B) Kalman filter estimation of G_c, the nanopore channel conductance. With the current noise model, the Kalman filter is able to produce accurate estimates of the nanopore channel conductance from real experimental data with an RMS estimation error as small as 7.2×10^{-4} nS.

3.3.4.1 Results

An example sinusoidal input, with a frequency 10 Hz and a peak-to-peak amplitude of 100 mV, and recorded output are shown in Figure 3.8A. At this frequency, the capacitance is excited enough to see the phase lead in the current when compared to the voltage input.

In general, the Kalman filter still performed best at lower frequencies and higher amplitudes with the lowest RMS estimation error of 7.117×10^{-4} nS occurring for the data shown in Figure 3.8. The error trends for our preliminary

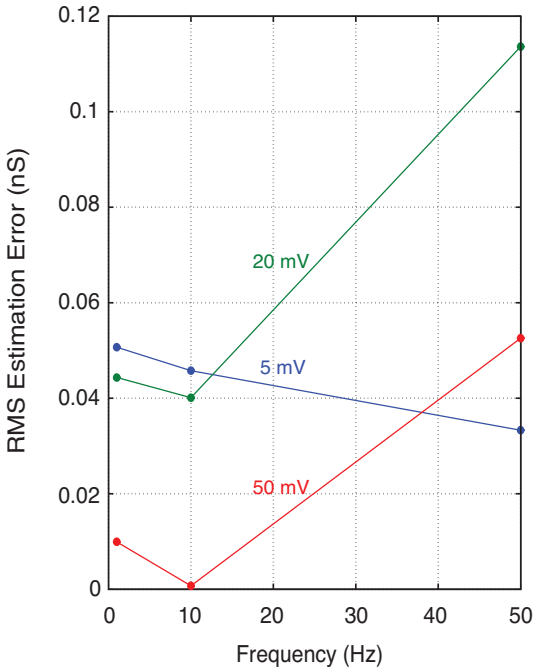

Figure 3.9: The RMS estimation error as a function of the amplitude and frequency of the input voltage using experimental data. More experimental data is needed to validate the relationships and dependences seen in the simulated data between the RMS estimation error and the amplitude and frequency.

testing of the Kalman filter with experimental data are shown in Figure 3.9. The linear and quadratic dependences between the RMS estimation error and the amplitude and frequency shown in simulations were not as evident. We suspect that a primary reason for the larger errors is that the noise model used by the Kalman filter is not adequate. In equation (3.3), the process and measurement noise are assumed to be white, but this assumption may be over simplistic. While our work in system identification of the nanopore system showed that reasonable model fitting results are obtained with this assumption, it also points to the possibility of colored noise in the system [22]. The performance of the Kalman filter is a direct reflection of the accuracy of the system model [24], and a more realistic noise model should improve performance. Future work will involve reapplying the

filter experimentally with revised noise model structures.

3.3.5 Discussion

We have presented our initial work on developing a Kalman filter for continuously estimating the channel conductance of a nanopore in time-varying voltage experiments. The simulation results with sinusoidal voltages show that the Kalman filter produces accurate estimates for a range of amplitudes and frequencies, and provides a basis for choosing input parameters to maximize the performance of the estimator. In general, the time varying voltages are chosen to meet a control objective that involves positioning of the DNA in the nanopore [90]. Though we have not yet used sinusoidal voltages for active control of DNA, these voltages can be used to maintain observability of the channel conductance as continuously as possible, and are already in use in other nanopore experiments [40]. The preliminary experimental results show that the Kalman filter can produce accurate conductance estimates for real nanopore data, and we anticipate that the performance will be improved with the addition of a more accurate noise model. By providing a means to accurately estimate the state of a molecule in a nanopore under voltage-varying conditions, the Kalman filter can lend itself to further expanding the capabilities of the nanopore instrument for science and technological applications. Future work will focus on improving the model used by the Kalman filter, and online implementation of the estimator to facilitate recovery of sequence information during voltage varying experiments [12].

3.4 Acknowledgments

Raj Maitra provided the experimental data. William Dunbar and Don Wiberg assisted in developing the system models used for the LSPE and Kalman filter. I designed and ran the simulation experiments, developed and ran the code to analyze the results, and wrote the text.

Chapter 4

Conclusion

The projects presented in this thesis represent fundamental groundwork that has been laid towards the goal of actualizing a novel method of single molecule detection and interrogation; the two-pore architecture. In Chapter 2 we investigated the need for rereading a single DNA molecule to gain statistical confidence about information contained in its sequence of nucleobases. While this study was focused on the specific application of sequencing, its lessons translate to any nanopore-based single molecule sensing application and thus provide ample motivation for the development of a nanopore device capable of active control. In Chapter 3 we presented two different parameter estimation methods and demonstrated their necessity and utility for achieving active control over a single molecule that has been captured in a nanopore.

This work merely represents the start of an ongoing project with a substantial amount of work still left to be done to realize the two-pore architecture. Apart from the complexities of fabricating such a device, the other main challenge that remains is developing the active control logic. In order to use modern control theory to actively control the motion of molecules through the two-pore architecture, a system model must first be created. This process can build on the work that has

already been done for modeling the biological nanopore system [22]. The model can be honed through the comparison of simulated output to actual data collected from the two-pore architecture until the mathematical model accurately reflects the performance of the physical system. Once the model has been perfected, it can be employed for use with active control methods.

Actively controlling molecules translocating through the two-pore system will require inferring the state of the two nanopores as continuously as possible. The Kalman filter developed in this work (Chapter 3), can be updated with the new system model to achieve this. Besides updating the system model, a new noise model for the Kalman filter must also be developed to better encompass the range of colored and white noise exhibited by the system. Once the updated Kalman filter has been thoroughly tested and verified to provide accurate continuous estimates of the nanopore channel conductances, it can be incorporated into an automated control program.

The automated control program can be used to coordinate and switch the transmembrane voltages for the two nanopores in concert. The first phase of the control program will be the capture phase where both voltages can be used to electrophoretically drive a molecule into both nanopores. When the Kalman filter senses that a molecule has been captured in both pores, the control program can switch to the holding phase. This phase involves using the two voltages to apply equal and opposite force on the captured molecule to keep it from leaving either pore. With the molecule held in both pores, the control program can switch to the sensing phase. In this phase, the voltages continue to apply opposing forces, but the amplitude of one of the voltages is increased slightly so that the captured molecule starts to slowly move in the direction of the increased voltage. This controlled motion of the molecule can allow features of the molecule (such

as the nucleobases of ssDNA or RecA filaments on dsDNA) to be sensed with high fidelity. Once the region of interest on the molecule has passed through the nanopores, the voltage amplitudes are reversed. This will cause the molecule to slowly traverse back through the nanopores in the opposite direction. Repeated application of this sensing phase of the control program can enable the structure of a single molecule to be reread as many times as necessary to gain statistical confidence about the sensed features. Once the desired number of rereads has been met, the control program can switch back to the capture phase, which will allow the captured molecule to leave the nanopores and clear the way for the next molecule.

The development of the active control logic and fabrication of the two-pore device is already well underway. Hopefully the work presented in this thesis proves to be helpful in achieving the great promise that the two-pore architecture is poised to bring to the scientific community.

Appendix A

Supporting Information for Error Analysis of Idealized Nanopore Sequencing

A.1 Identifiability of DNA sequences from ionic current amplitude

If the ssDNA passing through the pore is controlled by an enzyme on the pore, the ssDNA moves in 1 nt steps, with the dwell time of each ssDNA position being exponentially distributed, and step-transitions that are instantaneous compared to the measurement bandwidth [12, 49]. An appropriate idealization for the signal is a *pulse-train*, defined by a set of M amplitudes and a sequence of measured dwell times. From the single-channel recording and analysis literature [67, 83, 64], there are a set of techniques that can be applied to estimate the pulse-train idealization from the noisy recorded data. For sequencing, the pulse train would be compared to a library of amplitudes identified through control experiments with known

sequences. In this section, we consider challenges associated with having a limited number of distinct amplitude levels in the idealization.

If n nucleotides affect the ionic current, then $M = 4^n$ amplitude levels are sufficient to unambiguously identify the sequence. Of course, all 4^n amplitude levels may not be necessary to unambiguously identify the sequence; however, it is straightforward to construct cases for which less than 4^n amplitude levels results in systematic errors. For the purpose of synthesizing the idealization from noisy data, each amplitude level must have a signal-to-noise ratio (SNR) of at least 2 for idealization by half-amplitude methods, or at least 1.5 by Markov-based methods [83]. For $n \geq 3$, as in the case of the MspA nanopore [49], achieving $M = 4^n$ amplitude levels with sufficient SNR may not be possible. As stated, we consider specific examples in which having fewer that 4^n amplitude levels makes it impossible to unambiguously identify the sequence. We construct examples for $n = 2, 3, 4$, assuming sequences are identified right-to-left as they pass through the pore. Thus, $AGCTTAG$ with $n = 4$ would be identified $TTAG$, then $CTTA$, etc. We refer to M "distinct" amplitude levels when each level has sufficient SNR for detection. The case for $n = 1$ nucleotide affecting the current amplitude is considered first, and is the simplest.

Proposition 1. If $n = 1$, then $M = 4^n = 4$ distinct amplitude levels are necessary to unambiguously identify the sequence.

Proof. This follows trivially. Suppose $n = 1$ and $M = 3$. Then two of the four nucleotides generate the same current amplitude. There is no way to reconcile which of these nucleotides is present in the sensing region, using solely current amplitude. If $M = 2$ or 1, then one or none of the nucleotides are identifiable, respectively. □

Next, for the case $n = 2$, we construct a case in which having fewer than

$4^n = 16$ distinct amplitude levels makes it impossible to resolve all sequences.

Proposition 2. Suppose $n = 2$ and there are $M < 4^n = 16$ distinct amplitude levels. Let $X, Y \in \{A, T, G, C\}$ and $X \neq Y$. If the three pairs in the set $\{XX, XY, YX\}$ generate the same amplitude I_1, then there are an infinite number of sequences that cannot be identified from the pulse-train. In particular, no subsequence $Z_1 \cdots Z_m$ can be identified within the sequence $XZ_1 \cdots Z_m X$, provided $Z_i \in \{X, Y\}$ for $i = 1, ..., m$ $(m \geq 1)$ and each Y is separated by one or more Xs.

Proof. Without loss of generality, let $X = C$ and $Y = A$, and assume the pairs in the set $\{CC, CA, AC\}$ generate the same amplitude I_1. Then the nucleotide $Z_1 \in \{A, C\}$ within the triple CZ_1C cannot be identified. We can show this by considering an example sub-sequence $S_1 = TCACG$ to be identified. The amplitude that registers CG (assumed to be identifiable) can be used to choose yCG upon detecting I_1, with $y = A$ or C. After the second I_1 is detected (assuming a tracking counter is enabled) we have $xyCG$ with $xy \in \{CC, CA, AC\}$. Next, TC is detected (assuming it is identifiable), which constrains the value of $x = C$. The value for y, however, cannot be resolved. Additionally, any subsequence constructed from A and C and nested within $C \cdots C$ cannot be identified, provided each A is nested within Cs. An example is the subsequence CCC within $C\underline{CCC}C$, which is indistinguishable from the underlined subsequences within $C\underline{CAC}C$, $C\underline{ACA}C$, $C\underline{ACC}C$, and $C\underline{CCA}C$. Also, the longer the nested subsequence, the larger the set of subsequences that are indistinguishable. □

If one adds YY to the set in Proposition 2, the result is a larger number of unidentifiable subsequences.

Proposition 3. Suppose $n = 2$ and there are $M < 4^n = 16$ distinct amplitude levels. Let $X, Y \in \{A, T, G, C\}$ and $X \neq Y$. If the four pairs in the set

$\{XX, XY, YX, YY\}$ generate the same amplitude I_1, then there are an infinite number of sequences that cannot be identified from the pulse-train. In particular, no subsequence $Z_1 \cdots Z_m$ within the sequence $Z_0 Z_1 \cdots Z_m Z_{m+1}$ can be identified, with $Z_i \in \{X, Y\}$ for $i = 0, ..., m+1$ $(m \geq 1)$.

Proof. The proof follows the same logic as in the proof of Proposition 2, with the unidentifiable subsequence being nested within $X \cdots X$, $Y \cdots X$, $X \cdots Y$ or $Y \cdots Y$. □

To see the increase in the number of sequences that cannot be resolved, let $X = A$ and $Y = C$ and assume $\{CC, CA, AC, AA\}$ generate the same amplitude. Then the nucleotide $Z_1 \in \{A, C\}$ cannot be identified within any of the triples: CZ_1C, AZ_1C, CZ_1A, or AZ_1A. There is also a greater number of subsequence permutations that cannot be resolved for a given subsequence length $m > 1$. As an example, again with $X = A$ and $Y = C$ and assuming $\{CC, CA, AC, AA\}$ generate the same amplitude, the subsequence $CCCC$ with $m = 4$ is not identifiable from within $C\underline{CCCCC}$, $A\underline{CCCCC}$, $C\underline{CCCCC}A$ or $A\underline{CCC}A$. Moreover, all $2^m = 16$ four-letter combinations of A and C are indistinguishable from $CCCC$.

We consider next $n = 3$, which approaches the sensitivity of the biological pore MspA [49] and matches the claimed sensitivity of the nanopores developed by Oxford Nanopore Technologies. It is unlikely that all $M = 4^3 = 64$ distinct amplitudes are available for idealization.

Proposition 4. Suppose $n = 3$ and there are $M < 4^n = 64$ distinct amplitude levels. Let $X, Y \in \{A, T, G, C\}$ and $X \neq Y$. If the four triples in the set $\{XXX, XXY, XYX, YXX\}$ generate the same amplitude I_1, then there are an infinite number of sequences that cannot be identified from the pulse-train. In particular, no subsequence $Z_1 \cdots Z_m$ can be identified within the sequence

$XXZ_1 \cdots Z_m XX$, provided $Z_i \in \{X, Y\}$ for $i = 1, ..., m$ $(m \geq 1)$ and each Y is separated by two or more Xs.

Proof. Without loss of generality, let $X = C$ and $Y = A$, and assume the elements in the set $\{CCC, CCA, CAC, ACC\}$ generate the same amplitude I_1. Then A and C are indistinguishable within the sequences $CCACC$ and $CCCCC$, respectively. We can show this by considering an example sub-sequence $S_2 = TCCACCG$ to be identified. The amplitude that registers CCG (assumed to be identifiable) can be used to choose $yCCG$ upon detecting I_1, with $y = A$ or C. After the second I_1 is detected (assuming a tracking counter is enabled) we have $yzCCG$ with $yz \in \{AC, CA, CC\}$. After the third I_1 is detected, we have $xyzCCG$ with $xyz \in \{CCC, CCA, CAC, ACC\}$. Next, TCC is detected (assuming it is identifiable), which constrains the value of $xy = CC$. The value for z cannot be resolved. Following the generalization for this example, it is straightforward to show that CC, AC and CA are indistinguishable within $CCCCCC$, $CCACCC$ and $CCCACC$, respectively. The longer the nested subsequence, the larger the set of subsequences that are indistinguishable. □

Proposition 5. Suppose $n = 4$ and there are $M < 4^n = 256$ distinct amplitude levels. Let $X, Y \in \{A, T, G, C\}$ and $X \neq Y$. If the five elements in the set

$$\{XXXX, XXXY, XXYX, XYXX, YXXXX\}$$

generate the same amplitude I_1, then there are an infinite number of sequences that cannot be identified from the pulse-train. In particular, no subsequence $Z_1 \cdots Z_m$ can be identified within the sequence $XXXZ_1 \cdots Z_m XXX$, provided $Z_i \in \{X, Y\}$ for $i = 1, ..., m$ $(m \geq 1)$ and each Y is separated by three or more Xs.

Proof. The proof follows the same logic as the proofs for Propositions 2-4. Without loss of generality, let $X = C$ and $Y = A$, and assume the elements in the set

$$\{CCCC, CCCA, CCAC, CACC, ACCC\}$$

generate the same amplitude I_1. Then A and C (where underlined) are indistinguishable within the sequences $CCC\underline{A}CCC$ and $CCC\underline{C}CCC$, respectively. We can show this by considering an example sub-sequence $S_2 = TCCCACCCG$ to be identified. The amplitude that registers $CCCG$ (assumed to be identifiable) can be used to choose $zCCCG$ upon detecting I_1, with $z = A$ or C. After the second I_1 is detected (assuming a tracking counter is enabled) we have $yzCCCG$ with $yz \in \{AC, CA, CC\}$. After the third I_1 is detected, we have $xyzCCCG$ with $xyz \in \{CCC, CCA, CAC, ACC\}$. After the fourth I_1 is detected, we have $wxyzCCCG$ with

$$wxyz \in \{CCCC, CCCA, CCAC, CACC, CACC, ACCC\}$$

Next, $TCCC$ is detected (assuming it is identifiable), which constrains the value of $wxy = CCC$. The value for z cannot be resolved. Following the generalization for this example, it is straightforward to show that CC, AC and CA are indistinguishable within $CCC\underline{CC}CCC$, $CCC\underline{AC}CCC$ and $CCC\underline{CA}CCC$, respectively. The longer the nested subsequence, the larger the set of subsequences that are indistinguishable. □

The results in Propositions 2-5 show that there may be sequences which cannot be identified by amplitude level classification. Moreover, the examples do not cover all possible cases where identifiability is lost; they show only the existence of cases where identifiability is lost. All cases should be enumerated as part of

efforts to sequence based on ionic current. Additionally, the cases shown are not unreasonable, in the sense that such sequences might be expected to have a common amplitude, particularly for $n = 3, 4$. Until control experiments reveal which sequences cannot be robustly separated by distinct amplitudes, and for what n value(s), it is not clear if the distinct amplitude levels that register in the ionic current will be sufficient to identify intact ssDNA sequences.

Bibliography

[1] P.Yu Apel, Yu.E Korchev, Z Siwy, R Spohr, and M Yoshida. Diode-like single-ion track membrane prepared by electro-stopping. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 184(3):337 – 346, 2001.

[2] Thomas Aref, Mikas Remeika, and Alexey Bezryadin. High-resolution nanofabrication using a highly focused electron beam. *Journal of Applied Physics*, 104(2):–, 2008.

[3] M Bates, M Burns, and A Meller. Dynamics of DNA molecules in a membrane channel probed by active control techniques. *Biophysical Journal*, 84:2366–2372, 2003.

[4] Seico Benner, Roger J. A. Chen, Noah A. Wilson, Robin Abu-Shumays, Nicholas Hurt, Kate R. Lieberman, David W. Deamer, William B. Dunbar, and Mark Akeson. Sequence-specific detection of individual DNA polymerase complexes in real time using a nanopore. *Nature Nanotechnology*, 2:718–724, 2007.

[5] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–62, Sep 1997.

[6] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. D. 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.*, 26(10):1–8, Sep 2008.

[8] W. Brockman, P. Alvarez, S. Young, M. Garber, G. Giannoukos, W.L. Lee, C. Russ, E.S. Lander, C. Nusbaum, and D.B. Jaffe. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.*, 18(5):763–770, 2008.

[9] T. Z. Butler, M. Pavlenok, I. M. Derrington, M. Niederweis, and J. H. Gundlach. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc. Natl. Acad. Sci.*, 105(52):20647–52, 2008.

[10] S. Chang, S. Huang, H. Liu, P. Zhang, F. Liang, R. Akahori, S. Li, B. Gyarfas, J. Shumway, B. Ashcroft, J. He, and S. Lindsay. Chemical recognition and binding kinetics in a functionalized tunnel junction. *Nanotechnology*, 23(23):235101, 2012.

[11] P. Chen, J. Gu, E. Brandin, Y. Kim, Q. Wang, and D. Branton. Probing single DNA molecule transport using fabricated nanopores. *Nano. Lett.*, 4(11):2293–98, 2004.

[12] Gerald M. Cherf, Kate R. Lieberman, Hytham Rashid, Christopher E. Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology*, doi:10.1038/nbt.2147, 2012.

[13] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotech.*, 4(4):265–270, 2009.

[14] S.L. Cockroft, J. Chu, M. Amorin, and M. R. Ghadiri. A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution. *J. Amer. Chem. Soc.*, 130(3):818–820, 2008.

[15] C. Dekker. Solid-state nanopores. *Nature Nanotech.*, 2:209–215, 2007.

[16] I. M. Derrington, T. Z. Butler, M. D. Collins, E. Manrao, M. Pavlenok, M. Niederweis, and J. H. Gundlach. Nanopore DNA sequencing with MspA. *Proc. Natl. Acad. Sci. U.S.A.*, 107(37):16060–5, Sep 2010.

[17] O. K. Dudko, J. Mathé, and A. Meller. Nanopore force spectroscopy tools for analyzing single biomolecular complexes. *Meth. Enzymol.*, 475:565–89, Jan 2010.

[18] E N Ervin, R Kawano, R White, and H White. Simultaneous alternating and direct current readout of protein ion channel blocking events using glass nanopore membranes. *Anal. Chem*, 80(6):2069–2076, 2008.

[19] Brent Ewing and Phil Green. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.*, 8:186–194, 1998.

[20] Michael D. Fischbein and Marija Drndić. Electron beam nanosculpting of suspended graphene sheets. *Applied Physics Letters*, 93(11):113107, 2008.

[21] S. Garaj, W. Hubbard, A. Reina, J. Kong, D. Branton, and J. A. Golovchenko. Graphene as a subnanometre trans-electrode membrane. *Nature*, 467:190–194, Sep 2010.

[22] Daniel R. Garalde, Christopher R. O'Donnell, Raj D. Maitra, Donald M. Wiberg, Gang Wang, and William B. Dunbar. Modeling the biological nanopore instrument for biomolecular state estimation. *accepted, under re-review, IEEE Transactions on Control Systems Technology*, 2012.

[23] Graham C. Goodwin, Richard H. Middleton, and H. Vincent Poor. High-speed digital signal processing and control. *Proceedings of the IEEE*, 80(2):240–259, February 1992.

[24] Mohinder S. Grewal and Angus P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. John Wiley and Sons, Inc., second edition, 2001.

[25] Fredrik Gustafsson. *Adaptive Filtering and Change Detection*. John Wiley and Sons Ltd, first edition, 2000.

[26] B. Gyarfas, F. Olasagasti, S. Benner, D. Garalde, K. R. Lieberman, and M. Akeson. Mapping the position of DNA polymerase-bound DNA templates in a nanopore at 5 Å resolution. *ACS Nano*, 3(6):1457–1466, 2009.

[27] A. R. Hall, A. Scott, D. Rotem, K. K. Mehta, H. Bayley, and C. Dekker. Hybrid pore formation by directed insertion of α-haemolysin into solid-state nanopores. *Nature Nanotech.*, 5(12):874–7, Dec 2010.

[28] A. R. Hall, S. van Dorp, S. Lemay, and C. Dekker. Electrophoretic force on a protein-coated DNA molecule in a solid-state nanopore. *Nano Lett.*, 9(12):4441–4445, 2009.

[29] Simon Haykin. *Kalman Filtering and Neural Networks*. John Wiley and Sons, Inc., 2001.

[30] Robert A. Holt and Steven J. M. Jones. The new paradigm of flow cell sequencing. *Genome Res.*, 18(6):839–846, June 2008.

[31] S. Howorka, S. Cheley, and H. Bayley. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nat. Biotechnol.*, 19:636–9, 2001.

[32] 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. *Nature Nanotech.*, 5(12):868–73, Dec 2010.

[33] S. M. Iqbal, D. Akin, and R. Bashir. Solid-state nanopore channels with DNA selectivity. *Nature Nanotech.*, 2:243–248, 2007.

[34] M. Jain, H. Olsen, B. Paten, and M. Akeson. The oxford nanopore minion: delivery of nanopore sequencing to genomic community. *Genome Biol.*, 17(239), 2016.

[35] Jin Jiang and Youmin Zhang. A novel variable-length sliding window blockwise least-squares algorithm for on-line estimation of time-varying parameters. *International Journal of Adaptive Control and Signal Processing*, 18(6):505–521, June 2004.

[36] M. J. Kim, M. Wanunu, D.C. Bell, and A. Meller. Rapid fabrication of uniformly sized nanopores and nanopore arrays for parallel DNA analysis. *Adv. Mater.*, 18:3149–53, 2005.

[37] Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W. Kinzler, and Bert Vogelstein. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 108(23):9530–5, Jun 2011.

[38] N. Kobayashi, K. Tamura, and T. Aotsuka. Pcr error and molecular population genetics. *Biochemical genetics*, 37(9-10):317–321, 1999.

[39] Ernest T. Lam, Alex Hastie, Chin Lin, Dean Ehrlich, Somes K. Das, Michael D. Austin, Paru Deshpande, Han Cao, Niranjan Nagarajan, Ming Xiao, and Pui-Yan Kwok. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.*, 30(8):771–776, 08 2012.

[40] Daniel K. Lathrop, Eric N. Erivin, Geoffrey A. Barrall, Michael G. Keehan, Ryuji Kawano, Michael A. Krupka, Henry S. White, and Andrew H. Hibbs. Monitoring the Escape of DNA from a Nanopore Using an Alternating Signal. *Journal of the American Chemical Society*, 132(6):1878–1885, 2010.

[41] J. Li, D. Stein, C. McMullan, D. Branton, M.J. Aziz, and J.A. Golovchenko. Ion-beam sculpting at nanometre length scales. *Nature*, 412:166–9, 2001.

[42] F. Liang, S. Li, S. Lindsay, and P. Zhang. Synthesis, Physicochemical Properties, and Hydrogen Bonding of 4(5)-Substituted 1-H-Imidazole-2-carboxamide, a Potential Universal Reader for DNA Sequencing by Recognition Tunneling. *Chem. Eur. J.*, 18(19):5998–6007, 2012.

[43] K. R. Lieberman, G. M. Cherf, M. J. Doody, F. Olasagasti, Y. Kolodji, and M. Akeson. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *J. Am. Chem. Soc.*, 132(50):17961–72, 2010.

[44] Lennart Ljung and Svante Gunnarsson. Adaptation and tracking in system identification—a survey. *Automatica*, 26(1):7–21, 1990.

[45] Nicholas J. Loman, Chrystala Constantinidou, Jacqueline Z.M. Chan, Mihail Halachev, Martin Sergeant, Charles W. Penn, Esther R. Robinson, and Mark J. Pallen. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.*, 10(9):599–606, Sep 2012.

[46] Nicholas J. Loman, Raju V. Misra, Timothy J. Dallman, Chrystala Constantinidou, Saheer E. Gharbia, John Wain, and Mark J. Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.*, 30(5):434–9, May 2012.

[47] B. Luan, H. Peng, S. Polonsky, S. Rossnagel, G. Stolovitzky, and G. Martyna. Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. *Phys. Rev. Lett.*, 104:2381031–4, 2010.

[48] Raj D Maitra, Jungsuk Kim, and William B Dunbar. Recent advances in nanopore sequencing. *Electrophoresis*, 33(23):3418–28, December 2012.

[49] E. A. Manrao, I. M. Derrington, A. H. Laszlo, K. W. Langford, M. K. Hopper, N. Gillgren, M. Pavlenok, M. Niederweis, and J. H. Gundlach. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.*, 30:349–353, 2012.

[50] Christopher A. Merchant, Ken Healy, Meni Wanunu, Vishva Ray, Neil Peterman, John Bartel, Michael D Fischbein, Kimberly Venta, Zhengtang Luo, A.T. Charlie Johnson, and Marija Drndic. Dna translocation through graphene nanopores. *Nano Lett.*, 10:2915–2921, 2010.

[51] David Mosén-Ansorena, Ana María Aransay, and Naiara Rodríguez-Ezpeleta. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC bioinformatics*, 13:192, January 2012.

[52] Liviu Movileanu, Jason P Schmittschmitt, J Martin Scholtz, and Hagan Bayley. Interactions of peptides with a protein pore. *Biophysical journal*, 89(2):1030–45, August 2005.

[53] K S Novoselov, a K Geim, S V Morozov, D Jiang, Y Zhang, S V Dubonos, I V Grigorieva, and a a Firsov. Electric field effect in atomically thin carbon films. *Science (New York, N.Y.)*, 306(5696):666–9, October 2004.

[54] C. O'Donnell and W. Dunbar. Least-squares estimation of nanopore channel conductance in volage-varying experiments. In *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, 2012.

[55] C. O'Donnell, H. Wang, and W. Dunbar. Error analysis of idealized nanopore sequencing. *Electrophoresis*, 34(15):2137–2144, 2013.

[56] C. O'Donnell, D. Wiberg, and W. Dunbar. A kalman filter for estimating nanopore channel conductance in voltage-varying experiments. In *51st IEEE Conference on Decision and Control*, pages 2304–2309, 2012.

[57] Felix Olasagasti, Kate R Lieberman, Seico Benner, Gerald M Cherf, Joseph M Dahl, David W Deamer, and Mark Akeson. Replication of individual DNA molecules under electronic control using a protein nanopore. *Nature Nanotechnology*, 5(11):798–806, 2010.

[58] T. Perkins, R. Dalal, P. Mitsis, and S. Block. Sequence-dependent pausing of single lambda exonuclease molecules. *Science*, 301:1914–1918, 2003.

[59] Andrew Pollack. Company unveils DNA sequencing device meant to be portable, disposable and cheap. *The New York Times*, Feb 18, 2012, p. B2.

[60] Stas Polonsky, Steve Rossnagel, and Gustavo Stolovitzky. Nanopore in metal-dielectric sandwich for DNA position control. *Applied Physics Letters*, 91(15):153103, 2007.

[61] H. W. C. Postma. Rapid Sequencing of Individual DNA Molecules in Graphene Nanogaps. *Nano Lett.*, 10(2):420–5, 2010.

[62] R. F. Purnell, K. K. Mehta, and J. J. Schmidt. Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore. *Nano Lett.*, 8(9):3029–3034, 2008.

[63] R. F. Purnell and J. J. Schmidt. Discrimination of single base substitutions in a DNA strand immobilized in a biological nanopore. *ACS Nano*, 3(9):2533–2538, 2009.

[64] F. Qin, A. Auerbach, and F. Sachs. Hidden Markov modeling for single channel kinetics with filtering and correlated noise. *Biophys. J.*, 79(10):1928–1944, 2000.

[65] B.E. Ragle and J.B. Wardenburg. Anti-alpha-hemolysin monoclonal antibodies mediate protection against staphylococcus aureus pneumonia. *Infect. Immun.*, 2009.

[66] A. I. Roca and M. M. Cox. RecA protein: structure, function, and role in recombinational DNA repair. *Progress in nucleic acid research and molecular biology*, 56:129–223, 1997.

[67] B. Sakmann and E. Neher, editors. *Single-Channel Recording*. Plenum Press, 2nd edition, 1995.

[68] M. Sanchez-Cespedes, P. Parrella, S. Nomoto, D. Cohen, Y. Xiao, M. Esteller, C. Jeronimo, R. C. K. Jordan, T. Nicol, W. M. Koch, M. Schoenberg, P. Mazzarelli, V. M. Fazio, and D. Sidransky. Identification of a Mononucleotide Repeat as a Major Target for Mitochondrial DNA Alterations in Human Tumors. *Cancer Res.*, 61(19):7015–7019, 2001.

[69] G. F. Schneider, S. W. Kowalczyk, V. E. Calado, G. Pandraud, H. W. Zandbergen, L. M. K. Vandersypen, and C. Dekker. DNA translocation through graphene nanopores. *Nano. Lett.*, 10(8):3163–7, Aug 2010.

[70] R. Smeets, U. Keyser, M. Wu, N. Dekker, and C. Dekker. Nanobubbles in Solid-State Nanopores. *Physical Review Letters*, 97(8):088101, August 2006.

[71] R. M. M. Smeets, U. F. Keyser, N. H. Dekker, and C. Dekker. Noise in solid-state nanopores. *Proc. Natl. Acad. Sci.*, 105(2):417–421, 2008.

[72] R M M Smeets, Stefan W Kowalczyk, a R Hall, N H Dekker, and Cees Dekker. Translocation of RecA-coated double-stranded DNA through solid-state nanopores. *Nano letters*, 9(9):3089–96, September 2009.

[73] Ralph M M Smeets, Ulrich F Keyser, Diego Krapf, Meng-Yue Wu, Nynke H Dekker, and Cees Dekker. Salt dependence of ion transport and DNA translocation through solid-state nanopores. *Nano letters*, 6(1):89–95, January 2006.

[74] L. Song, M. Hobaugh, C. Shustak, S. Cheley, H. Bayley, and J. E. Gouaux. Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science*, 274(5294):1859–1866, 1996.

[75] Claudia Stahl, Susanne Kubetzko, Iris Kaps, Silke Seeber, Harald Engelhardt, and Michael Niederweis. Mspa provides the main hydrophilic pathway through the cell wall of mycobacterium smegmatis. *Molecular Microbiology*, 40(2):451–464, 2001.

[76] D. Stoddart, A. J. Heron, J. Klingelhoefer, E. Mikhailova, G. Maglia, and H. Bayley. Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore. *Nano Lett.*, 10(9):3633–7, Sep 2010.

[77] A. J. Storm, J. H. Chen, X. S. Ling, H. W. Zandbergen, and C. Dekker. Fabrication of solid-state nanopores with single-nanometre precision. *Nat. Mater.*, 2003.

[78] J. Thompson and P. Milos. The properties and applications of single-molecule dna sequencing. *Genome Biol.*, 12(2):217, 2011.

[79] M. Tsutsui, Y. He, M. Furuhashi, S. Rahong, M. Taniguchi, and T. Kawai. Transverse electric field dragging of DNA in a nanochannel. *Sci. Rep.*, 2(394), Jan 2012.

[80] M. Tsutsui, M. Taniguchi, K. Yokota, and T. Kawai. Identifying single nucleotides by tunnelling current. *Nature Nanotech.*, 5:286–290, Jan 2010.

[81] Makusu Tsutsui, Sakon Rahong, Yoko Iizumi, Toshiya Okazaki, Masateru Taniguchi, and Tomoji Kawai. Single-molecule sensing electrode embedded in-plane nanopore. *Scientific reports*, 1:46, January 2011.

[82] S. van Dorp, U. Keyser, N. Dekker, C. Dekker, and S. G. Lemay. Origin of the electrophoretic force on DNA in solid-state nanopores. *Nature Phys.*, 5:347–351, Jan 2009.

[83] L Venkataramanan and F J Sigworth. Applying hidden Markov models to the analysis of single ion channel activity. *Biophysical journal*, 82(4):1930–42, April 2002.

[84] B. M. Venkatesan and R. Bashir. Nanopore sensors for nucleic acid analysis. *Nature Nanotech.*, 6:615–624, 2011.

[85] B. M. Venkatesan, D. Estrada, S. Banerjee, X. Jin, V. E. Dorgan, M. Bae, N. R. Aluru, E. Pop, and R. Bashir. Stacked Graphene-Al_2O_3 Nanopore Sensors for Sensitive Detection of DNA and DNA-Protein Complexes. *ACS Nano.*, 6(1):441–450, 2012.

[86] B. M. Venkatesan, A. B. Shah, J. Zuo, and R. Bashir. DNA Sensing Using Nanocrystalline Surface-Enhanced Al2O3 Nanopore Sensors. *Adv. Mater.*, 20(8):1266–1275, 2010.

[87] Bala Murali Venkatesan, Brian Dorvel, Sukru Yemenicioglu, Nicholas Watkins, Ivan Petrov, and Rashid Bashir. Highly sensitive, mechanically stable nanopore sensors for dna analysis. *Adv. Mater.*, 21:2771–2776, 2009.

[88] Deqiang Wang, Stefan Harrer, Binquan Luan, Gustavo Stolovitzky, Hongbo Peng, and Ali Afzali-Ardakani. Regulating the Transport of DNA through Biofriendly Nanochannels in a Thin Solid Membrane. *Scientific reports*, 4:3985, January 2014.

[89] Ruoshan Wei, Daniel Pedone, Andreas Zürner, Markus Döblinger, and Ulrich Rant. Fabrication of metallized nanopores in silicon nitride membranes for single-molecule sensing. *Small*, 6(13):1406–1414, 2010.

[90] Noah A. Wilson, Robin Abu-Shumays, Brett Gyarfas, Hongyun Wang, Kate R. Lieberman, Mark Akeson, and William B. Dunbar. Electronic control of DNA polymerase binding and unbinding to single DNA molecules. *ACS Nano*, 3:995–1003, 2009.

[91] Jiekun Xuan, Ying Yu, Tao Qing, Lei Guo, and Leming Shi. Next-generation sequencing in the clinic: Promises and challenges. *Cancer Letters*, 340(2):284 – 295, 2013. Next Generation Sequencing Applications in Cancer Research.

[92] Jingmin Zhang, Liping You, Hengqiang Ye, and Dapeng Yu. Fabrication of ultrafine nanostructures with single-nanometre precision in a high-resolution transmission electron microscope. *Nanotechnology*, 18(15):155303, 2007.