UNIVERSITY OF CALIFORNIA SAN DIEGO

Towards Single-Molecule Nanopore DNA and Protein Sequencing

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

in

Materials Science and Engineering

by

Wenxu Zhang

Committee in charge:

Professor Xiaohua Huang, Chair
Professor Vineet Bafna
Professor Gert Cauwenberghs
Professor Zhaowei Liu
Professor Donald Sirbuly

2021

The dissertation of Wenxu Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to Professor Xiaohua Huang, my lifetime mentor and friend, for his great guidance through my graduate studies and research. I am so lucky to be guided by him in the last five years. He provides me a lot of freedom to explore various projects, so I could have time to think and enjoy science, though some research projects could be quite challenging.

I would like to thank the rest of my committee members, Professor Vineet Bafna, Professor Gert Cauwenberghs, Professor Zhaowei Liu, and Professor Donald Sirbuly for their time and advice.

It is an honor to work with a group of talents who are knowledgeable, capable, and responsible. It is a cherished treasure in my whole life. My gratitude is here to all current and former members of the lab. Especially, I want to thank Sylvia Liang for contributing significantly to the nanopore single-molecule protein identification project. This work would not have been possible without her. I want to thank Norman Luong for providing his knowledge and experience in biological experiments and mechanical design. He gave me a lot of help in research and life. I want to thank Dr. Eric Chu for the valuable discussion in research. He also provides a lot of advice and helps in my life. Thanks Dr. Huang, Sylvia, and Yan for editing this thesis.

Additionally, I am grateful to my family for giving me their support and encouragement. I would like to thank Yan for her love, support, and accompany. I am grateful to my parents for their selfless sacrifices and support.

Chapter 2, in part, is currently being prepared for submission for publication of the material by Wenxu Zhang, and Xiaohua Huang. The dissertation author was the primary investigator and author of this material.

Chapter 3, in part, is currently being prepared for submission for publication of the material by Sylvia Liang, Wenxu Zhang, and Xiaohua Huang. Sylvia Liang and the dissertation author were the primary investigators and authors of this material.

Chapter 4, in part, is currently being prepared for submission for publication of the material by Wenxu Zhang and Xiaohua Huang. The dissertation author was the primary investigator and author of this material.

Chapter 5, in part, is currently being prepared for submission for publication of the material by Wenxu Zhang and Xiaohua Huang. The dissertation author was the primary investigator and author of this material.

VITA

2015	B.S. University of Science and Technology of China
2017	M.S. University of California San Diego
2021	Ph.D. University of California San Diego

ABSTRACT OF THE DISSERTATION

Towards Single-Molecule Nanopore DNA and Protein Sequencing

by

Wenxu Zhang

Doctor of Philosophy in Materials Science and Engineering

University of California San Diego, 2021

Professor Xiaohua Huang, Chair

Precise determination of DNA and protein sequences is essential to understanding biological systems. After more than 30 years of development, nanopore technology has been demonstrated to be capable of real-time long-read sequencing of single DNA molecules. However, current nanopore DNA sequencing is still limited by low consensus accuracy due to systematic error in reading homopolymers and other sequences with similar levels of ionic

blockage. While nanopore DNA sequencing technology is revolutionizing genomics research, whether an analogous nanopore technology can be used to identify and sequence protein molecules remains largely unexplored. In the first part of this dissertation, I investigated the feasibility of ultraaccurate nanopore DNA sequencing based on experimental sequencing data and computational modeling. In the second part of the dissertation, I investigated the theoretical feasibility of protein identification and sequencing using nanopore technologies. I also presented experimental work on the development of an integrated opto-electrical system for nanopore fabrication process and its potential application in single-molecule protein identification using high-speed optical detection.

Chapter 1 Introduction

1.1 Overview of sequencing technologies

Deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and protein play essential roles in all known forms of life. They all are composed of fundamental building blocks. DNA is a molecule consisting of two chains. Both chains consist of four nucleotides linked together by covalent bonds. The sequence of the nucleotides in the DNA molecules encodes the instructions for forming all other cellular components. Unlike double-stranded DNA, RNA is a single-stranded molecule. Several types of RNA molecules exist in cells, and each type carries out specific functionality, such as protein synthetic and gene regulation. Proteins, which comprise one or more long chains of amino acid residues, carry out most functions in living systems. The function of a protein depends on its amino acid sequencing. Quantitative understanding of biological operating systems and human diseases requires precise determination of DNA, RNA, and protein sequence.

1.1.1 Overview of DNA sequencing technologies

DNA sequencing developed by Sanger(1-3) uses the random early termination of DNA synthesis which is caused by dideoxynucleoside triphosphates (ddNTPs). In the Sanger sequencing, standard nucleotides, chain-terminating nucleotides (ddNTPs) and DNA polymerases are mixed. DNA synthesis is interrupted by ddNTPs and then DNA fragments with different lengths are generated. The lengths of those fragments are measured by gel electrophoresis, which can achieve separation with single nucleotide resolution. The DNA sequence is determined after four rounds of synthesis and electrophoresis, with one of four ddNTPs (ddATP, ddTTP, ddGTP, and ddCTP) and all normal dNTPs being added each time. The relative positions of the various bands on the gel encode the DNA sequence.

Illumina sequencer, a wildly used next-generation sequencing (NGS) platform, significantly increases sequencing throughputs(1). In Illumina sequencer, the genome sequences are first fragmented and immobilized on a solid surface. The immobilized single molecule template is then amplified into a dense cluster by bridge amplification on a silicon chip. The sequencing steps comprise cycles of biochemistry reaction and fluorescence imaging. The images are then analyzed to extract the DNA sequence. NGS is much cheaper and faster compared to Sanger sequencing. However, NGS is limited to short read length (50 – 500 bp), and the whole sequencing takes from hours to days to complete. The sequencing machine is also heavy and expensive.

Third-generation sequencing technologies offer the capability of sequencing a single molecule in real-time. Compared to NGS technologies, third-generation technologies can achieve a read length of more than 100kb. Two widely utilized sequencers, Pacific Bioscience (PacBio) and Oxford Nanopore Technologies (ONT), use different strategies to accomplish single-molecule long-read DNA sequencing. PacBio uses a zero-mode waveguide, optically observing polymerase-mediated synthesis in real time(4, 5). ONT uses ionic flow, electronically sensing nucleotides when a DNA molecule is translocated through a nanoscale pore.

1.1.2 Overview of RNA and protein sequencing technology

The current standard workflow on RNA sequencing is involved with library preparation and complementary DNA (cDNA) sequencing(6, 7). However, cDNA amplification will bring bias, distortion of relative abundances, and dropout of some species(8).

Edman degradation(9) and mass spectrometry are currently available methods for protein sequencing. In Edman degradation, the amino-terminal residue is cleaved from the peptide and then identified. This process is repeated until all amino acids in the peptide are determined. Mass

spectrometry offers an alternative method for protein sequencing. The protein molecule is first ionized and then moved under the electric and magnetic fields. The path of molecule movement is uniquely determined by its mass-to-charge ratio, and can be used to distinguish amino acids with very high accuracy. However, these methods require substantial effort, large quantities of protein and expensive instruments.

1.2 Development of nanopore technology

In 1989, David Deamer and George Church individually proposed the idea of nanopore sequencing(10). In an electrolytic solution environment, a thin membrane separates two liquid chambers. When a nanoscale pore is present on the thin membrane, ions could transport through the pore freely. Due to ion transportation, a small current could be observed with a bias voltage applied on this thin membrane. When a single-stranded DNA is translocated through the pore, different nucleotides will block the pore to a different extent and induce current fluctuation since each nucleotide has its unique size and chemical properties. The current readout will be decoded to its DNA sequence by a properly designed algorithm, such as Viterbi algorithm or Long Short-Term Memory (LSTM) neural network.

A well-behaved nanopore is essential to achieving high sequencing accuracy. First, the nanopore should have a similar diameter compared to the size of nucleotides. The thickness of the nanopore should also be compatible with the length of nucleotides. Otherwise, a more than the ideal number of nucleotides will be sensed simultaneously, and the current signal will be highly convoluted. It poses challenges to signal analysis. Second, nanopore with atomic-level positioning is essential to sequencing. Nanopores produced from different batches need to be homogeneous. A high variation in the pore property will cause inconsistent current fluctuation. While reproducibly fabricating solid-state nanopores(11, 12) with similar geometry could be

challenging, a biological nanopore(10, 13-17) designed with a bottom-up strategy could easily achieve single-atomics level positioning.

Besides precisely controlled nanopore fabrication technology, the temporal resolution for a single nucleotide during nanopore sensing needs to improve in order to increase sequencing accuracy. On the one hand, polymerases(14, 18), helicases(19, 20) are used to slow down DNA translocation to enable a better signal reading per nucleotide. On the other hand, current amplifier(21) with MHz level data acquisition bandwidth has been developed to acquire more data per nucleotide. However, higher bandwidth usually introduces a higher noise level. It also brings difficulty in signal processing.

Nanopore sequencing offers a wide range of unique benefits. It shows potential in genome assembly(22, 23), pathogen evolution(24-26), and other applications. As a direct and real-time sequencing platform, it eliminates unwanted amplification bias. It does not require any surrogate markers. Additionally, nanopore sequencing could directly detect base modification(27), while detecting methylation with bisulfite sequencing in the NGS requires additional chemical reactions.

In addition to DNA sequencing, nanopore technology has been used to develop a parallel, directly, amplification-free RNA sequencing method(8). Compared to NGS, nanopore sequencing does not need the reversed transcriptase step from RNA to cDNA. The direct RNA sequencing method also enables the analysis of RNA modifications(28).

Single-molecule protein fingerprinting and sequencing with nanopore technology is still in the nascent stage(29, 30). Enzymatic control of protein unfolding and translocation through the α-hemolysin has been demonstrated(31). However, it lacks the resolution to distinguish single amino acids.

1.3 Scope of the dissertation

The objective of this dissertation work is to develop new theories and technologies for nanopore sequencing of DNA and protein.

In this work, we first laid the foundations for ultra-accurate DNA nanopore sequencing. We then investigated the feasibility of single-molecule protein identification and sequencing using a nanopore. An opto-electrical system was developed to monitor nanopore controlled dielectric breakdown process. It could be additionally utilized for future protein fingerprinting experiments with synchronized optical and electrical measurement.

Chapter 2 Feasibility of Ultra-accurate Nanopore DNA Sequencing

2.1 Abstract

Nanopore DNA sequencing has great promise to enable accurate haplotype-resolved sequencing and assembly of individual human genomes. However, the current technology offered by Oxford Nanopore Technologies (ONT) is limited by low single-pass and consensus accuracy due to systematic errors in reading homopolymers and other sequences with similar levels of ionic blockage. We used computational modeling to investigate the nucleotide-to-nucleotide transition events through the CsgG nanopore employed in the ONT platform and found that the current profiles follow a sinusoid, which is observable in experimental sequencing data. Our modeling suggested that the CsgG nanopore has the dimensions required for accurate DNA sequencing. We performed computational sequencing of the human genome at 20 kHz detection bandwidth and showed that ultra accurate nanopore DNA sequencing is feasible (zero error in 140 million bp sequenced at 20x). Our modeling of how nanopore geometries influence current signal detection also provides insights for future nanopore engineering.

2.2 Introduction

Nanopore technology has enabled direct and rapid electronic sequencing of single DNA molecules with long read lengths with a portable device(10, 32, 33). It is now possible to sequence and assemble human genomes using the ONT platforms(33). However, the full potential of the technology has not been realized due to the relatively low consensus accuracy that can be achieved (99.97% at 100x)(33-35). The inability to increase consensus accuracy has been primarily ascribed to systematic errors, mostly deletions and insertions, in reading homopolymers and certain sequence combinations having similar levels of ionic current blockage(33-36). The lengths of the homopolymers are usually estimated by dwell time(35, 37).

Even though a helicase is employed to control DNA translocation, the velocity of nucleotide-to-nucleotide translocation still varies stochastically and has a relatively broad distribution(35, 37), which contributes to systematic errors in homopolymer sequencing. Systematic errors could also arise for certain sequence combinations that have similar or statistically indistinguishable current blockage levels(38). In principle, these systemic errors can be completely eliminated if the nucleotide-to-nucleotide transition events can be reliably detected and sequences with similar current levels can be resolved.

In this work, we used computational modeling to investigate the nucleotide-to-nucleotide transition events and whether the events are observable in experimental sequencing data acquired with MinION R9.4 chemistry. We also studied the physical origins of the current signal variations to better understand the characteristics of the measured current profiles. Finally, we investigated the theoretical single-pass and consensus sequencing accuracies that can be achieved using the current generation of CsgG nanopore device, and the influence of nanopore geometries on the detectability of the nucleotide-to-nucleotide transition events.

2.3 DNA modeling

We used finite element analysis (FEA) and the Poisson Nernst-Planck (PNP) equation system to compute ionic currents through nanopores. The geometry of the mutant CsgG nanopore was constructed based on the crystal structure of the wild-type protein(39), and further refined using experimental open-pore currents (Supplementary Figure 2.1a)(8, 33). Six consecutive DNA bases contribute to current blockade level. Constructing the large number of 3D models of 4096 DNA hexamers with fine structural features for FEA would be very challenging and computationally unmanageable. Therefore, we modeled DNA as a linear structure composed of connected cylinders with axial symmetry, which enabled us to use 2D

modeling to acquire data for the 3D models. The deoxyribonucleosides (dA, dC, dG and dT) are modeled as cylinders with the same height but of different diameters, and the phosphodiester linker is also modeled as a cylinder (Figure 2.1a).

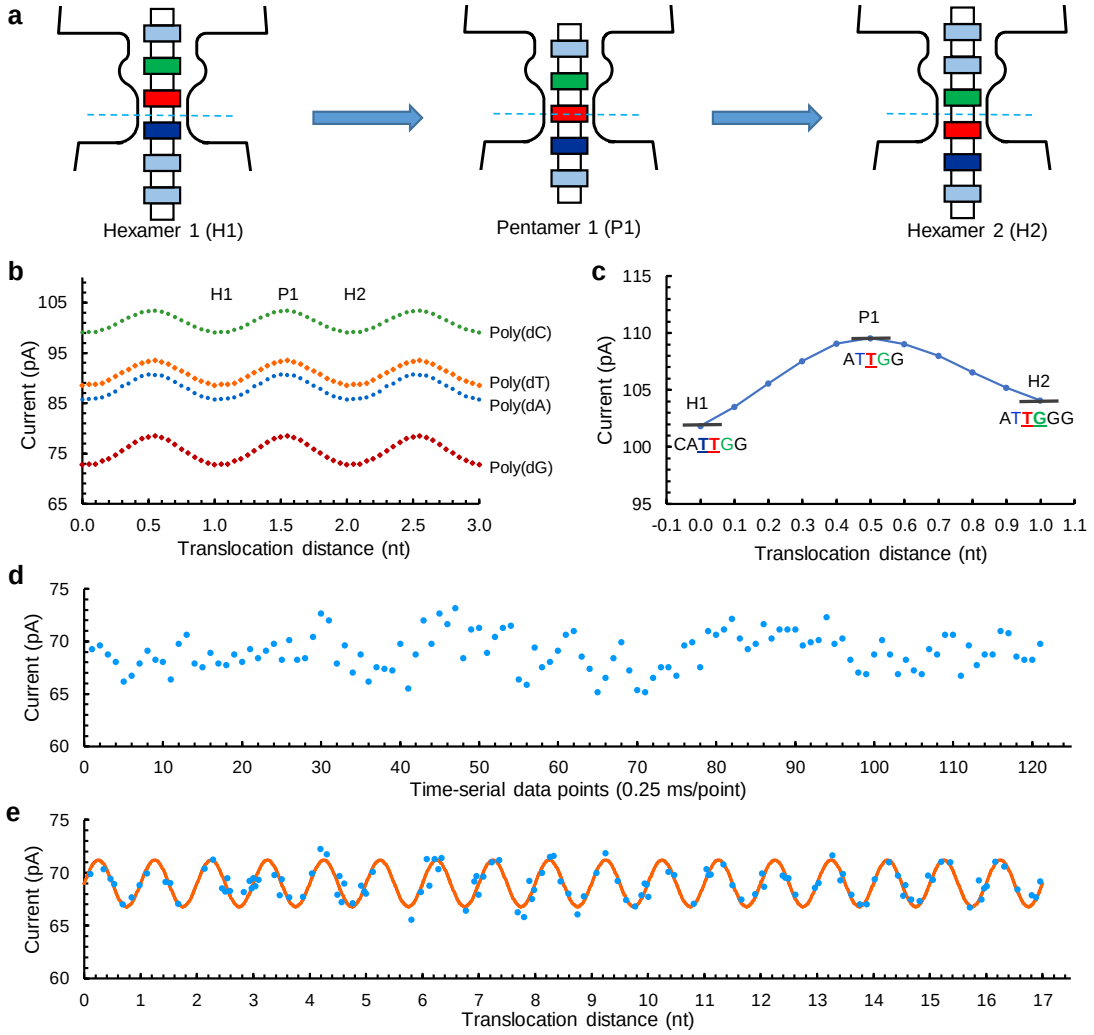


Figure 2.1 Sinusoid profiles of nucleotide-to-nucleotide DNA translocation through a nanopore.

(a) DNA translocation through a mutant CsgG nanopore. The translocation from one hexamer (H1) to the next hexamer (H2) transits through a pentamer (P1). The dashed line indicates the estimated narrowest constriction site. Illustrated are the 2D cross sections through the nanopore vertical axis. (b) Current profiles of homopolymers. The deoxyribonucleosides of the homopolymers are modeled as oval cylinders and the translocation was simulated by computing the current levels of the models at twenty equally spaced positions along the vertical axis of the nanopore. (c) Current profile of a non-homopolymer transition. The translocation from CATTGG to ATTGGG transits through a pentamer ATTGG. The pentamer and hexamers were modeled as circular cylinders, and translocation was simulated by computing the current levels of the models at eleven equally spaced positions along the vertical axis of the nanopore. (d) Raw experimental sequencing data of a homopolymer poly(dT)$_{17}$. The data was extracted from a region covering 23-base poly(dT)$_{23}$. For clarity, the signal for the last three dT bases at each boundary was removed. (e) Mapping of experimental data to a sine wave function. The evenly spaced time-serial data points are mapped well to a sine wave with 17 periods in the spatial domain using a Viterbi algorithm ($R^2 = 0.75$).

We developed a genetic algorithm to determine the diameters and heights that best represent the deoxyribonucleosides and phosphodiester linker based on experimental mean current levels of DNA hexamers(40) (Detail in Methods). Our initial attempt to use one diameter to model each nucleoside did not result in a good correlation between the model and experimental data ($R^2 \approx 0.5$). Therefore, we optimized the diameters by modeling each deoxyribonucleoside with 16 slightly different diameters depending on its nearest neighbors (Supplementary Tables 2.1 and 2.2). The optimized set of parameters were used to construct the models of 4096 DNA hexamers. Each hexamer was positioned inside the mutant CsgG nanopore with the center phosphodiester linker at the approximated narrowest constriction site (Fig. 2.1a), and FEA and PNP equation system were used to compute the ionic current level through the nanopore device (Supplementary Fig. 2.1b). Experimentally, several discrete current levels are observed for each hexamer(40), perhaps due to the stepwise behavior of the helicase(38, 41). Therefore, we compared the current levels of our models to the experimental mean values of the corresponding hexamers. The overall correlation is surprisingly good ($R^2 = 0.91$, Supplementary Fig. 2.3).

In addition to current level variations, significant signal variations within each current level are also observed in the experimental data(40). Since both vertical translocation and lateral movements contribute simultaneously to the current signals, it is not possible to pinpoint precisely the physical origins of the variations. However, by modeling each process separately, we can determine the origins and magnitudes of the variations. We computed the standard deviation of the signal variations associated with the lateral movements by positioning the DNA hexamer at various accessible locations from the nanopore center. With the simplistic circular

cylinder model of DNA, we found that the signal variation is much greater than the experimental value (2.91 pA vs. 0.80 pA for $(dC)_6$). To provide a more realistic model to mimic the overall shape of the deoxyribonucleosides, we modeled each nucleoside as an oval cylinder. The dimensions of the oval cross sections were optimized based on several constraints (Supplementary Figs. 2.4-2.6, and Supplementary Table 2.3). Using the oval cylinder models, we constructed the 3D models of four homopolymers, and computed the standard deviations of current levels due to lateral movements, including translations and rotations. The standard deviations for $(dA)_6$, $(dC)_6$, $(dG)_6$, and $(dT)_6$ were determined to be 0.77, 0.59, 0.97 and 0.74 pA, respectively, correlating very well with the experimental values (Supplementary Table 2.4).

2.3 Homopolymers translocation

Next, we investigated whether the nucleotide-to-nucleotide translocation of homopolymers through the mutant CsgG nanopore produces an observable periodic pattern. The current profiles were simulated by positioning the homopolymers at twenty consecutive locations along the vertical axis of the nanopore. Interestingly, we found that the current blockage profile of each nucleotide-to-nucleotide translocation event follows one period of a sine wave and the root mean square (RMS) amplitudes of the sinusoids range from 1.53 to 1.99 pA for the four different homopolymers (Fig. 2.1b, Supplementary Fig. 2.7 and Table 2.5). Non-homopolymers also follow a similar behavior with an average RMS amplitude of 1.72 pA (Fig. 2.1c, Supplementary Fig. 2.8a, b). Given that the RMS amplitudes are substantially greater than the signal variations due to lateral movements and measurement electronics noise (~ 1 pA), there is a possibility that the sinusoid patterns are observable in the experimental sequencing data. Since the sinusoid patterns of homopolymers are visually more apparent (Fig. 2.1b), we investigated

the possibility by examining the raw experimental current signal profiles of 25622 homopolymer regions (20-40 bases long) in chromosome 20 of the human genome(33).

The experimental current profiles are acquired at a fixed frequency (4 kHz), but the DNA translocation velocity is not uniform. Therefore, the sinusoid pattern is warped in the spatial domain and may not be apparent in the temporal profiles. We developed a Viterbi algorithm to unwarp and map the current profiles into a sine wave function as predicted by our modeling. Not surprisingly, the current profiles of many homopolymers are mapped well to sine wave functions (Fig. 2.1e and Supplementary Fig. 2.9a-h). We found that ~70% of the nucleotide-to-nucleotide transition events are mapped well to a sine wave period. The remaining 30% of the events are mapped poorly because they either have two or fewer measured data points or the data points cluster in the spatial domain, perhaps due to the stepwise behavior of the helicase(38) (Supplementary Fig. 2.10). The inability to reliably detect the 30% or so transition events very likely leads to a high systematic error in decoding the lengths of many homopolymers. Based on the mapping results, we also determined data point and translocation velocity distributions (Supplementary Fig. 2.11). About 15% of nucleotide-to-nucleotide translocation events have three or fewer measured data points. Due to the sinusoid behavior of the current profiles and uneven spatial distribution of the limited number of data points acquired, the assignment of a single value to the current level of a hexamer is very problematic, potentially resulting in an uncertainty greater than the signal variations of lateral motions or electronic measurement noise. The large uncertainty could also make the current levels of many hexamers indistinguishable, leading to higher systematic errors.

Our modeling results and analysis of the experimental data described above suggested that the current version of the mutant CsgG nanopore has the geometry required for highly

accurate DNA sequencing, and that the performance of the current ONT platform is limited by the detection bandwidth (4 kHz). We investigated whether ultra accurate DNA sequencing can be achieved if the bandwidth is increased to 20 kHz, which is achievable with a standalone (e.g. Axopatch 200B) or custom-designed CMOS amplifier(42). First, we modeled each hexamer to hexamer transition event by translocating the DNA in eleven spatially equal steps through the mutant CsgG nanopore using 2D models (Fig. 2.1a, c). The current levels of all possible hexamer to hexamer transition events were determined using 4096 x 11 models (Supplementary Fig. 2.8a, b). Signal variations due to lateral motions and electronic noise were simulated using a Gaussian distribution. The stochastic variations of translocation velocity were simulated using an inverse Gaussian distribution function (Supplementary Figs. 2.11, 2.12). We simulated the current profiles of 140,000 fragments of 1,000-base long sequences randomly selected from the reference human genome(33). Twenty current profiles (10 for the template strand and 10 for the reverse complement) were generated for each sequence. The simulated current profiles closely resemble those of the raw experimental sequencing data (Fig. 2.2a, b). The profiles were processed with a bandpass filter and segmented into three-level current profiles. The signals were then decoded using a Viterbi algorithm and a 3-state hidden Markov model (HMM) (Fig. 2.2c. Supplementary Fig. 2.13). Surprisingly, we achieved a high single-pass accuracy of 99.4% and extremely high consensus accuracy (zero error in 140 million bp at 20x sequencing depth). At a higher measurement bandwidth, the nucleotide-to-nucleotide transition events can be reliably detected, allowing for accurate homopolymer sequencing and the complete elimination of insertion and deletion errors. We also found that potential systematic errors due to decoding certain sequences with essentially indistinguishable current profiles can be completely eliminated by sequencing the complementary strands (Supplementary Fig. 2.14).

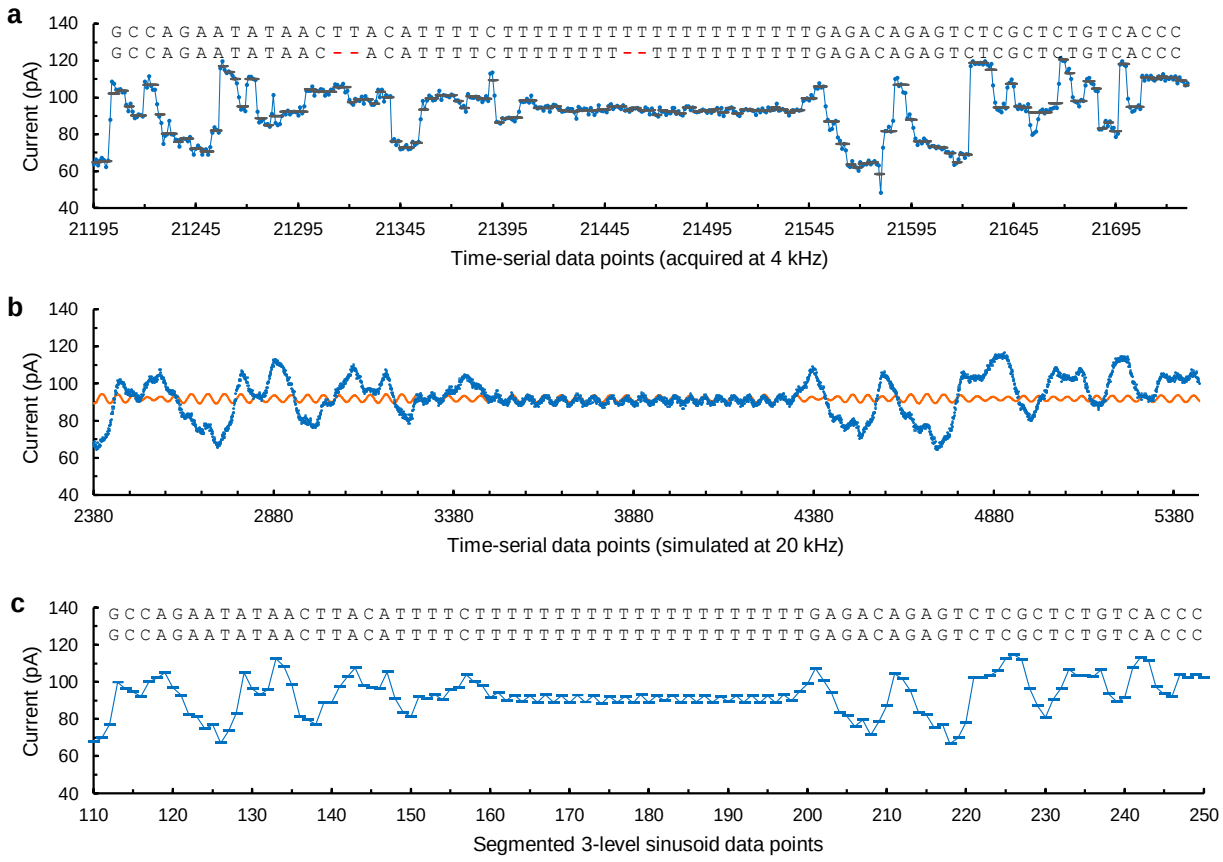


Figure 2.2 Sequence decoding using the sinusoid profiles of DNA translocation and a 3-state HMM.
(a) Raw experimental current profiles and decoded sequence. The profile was extracted from published data(33). The current profile (dotted line) was segmented into current levels (horizontal bars) using Metrichor and decoded using Guppy. The decoded sequence, which has two deletion errors, is shown below the reference sequence. (b) Simulated sequencing current profile. The same reference sequence from (a) was used for the simulation. The simulated profile (blue dots) is shown along the filtered data (orange) after the application of a 400-500 Hz bandpass filter. The DC component of the signal has been added to the filtered data for visualization purpose. (c) Sequence decoding using a 3-state HMM. The simulated data were segmented into 3-level current profiles based on the sinusoid pattern (Fig. 2.1b, c) and decoded using a Viterbi algorithm and a 3-state HMM. In this case, the sequence is decoded with 100% accuracy, identical to the reference.

2.4 Optimal nanopore geometry

Furthermore, we investigated how the nanopore geometries influence the amplitude of the sinusoid pattern, and current level variations (Fig. 2.3 and Supplementary Table 2.6). Unexpectedly, we found that the amplitude oscillates when the height of a nanopore with 1.0 nm

diameter decreases from 1.2 nm to 0.6 nm, and that a nanopore with 1.0 nm in both height and diameter produces an amplitude less than the signal variation due to lateral movements (Fig. 2.3a, c). We also found that the sinusoid amplitude decreases quickly below the lateral signal variation as the diameter of a nanopore with 0.8 nm height increases from 1.0 nm to 1.4 nm (Fig. 2.3b, d). A nanopore with a diameter of 1.0 nm and a height of 0.8 nm gives a very pronounced root mean square amplitude (3.9 pA), much greater than that of the mutant CsgG nanopore (1.53 pA) and lateral signal variation (~1 pA). Engineering such a nanopore to further improve the detectability of nucleotide-to-nucleotide transition events and sequencing accuracy is potentially feasible considering that protein nanopores with similar dimensions already exist in nature and are being employed for DNA sequencing (e.g. the height of MspA nanopore is ~0.6 nm(43, 44) and the diameter of wild-type CsgG nanopore is only 0.95 nm(39)).

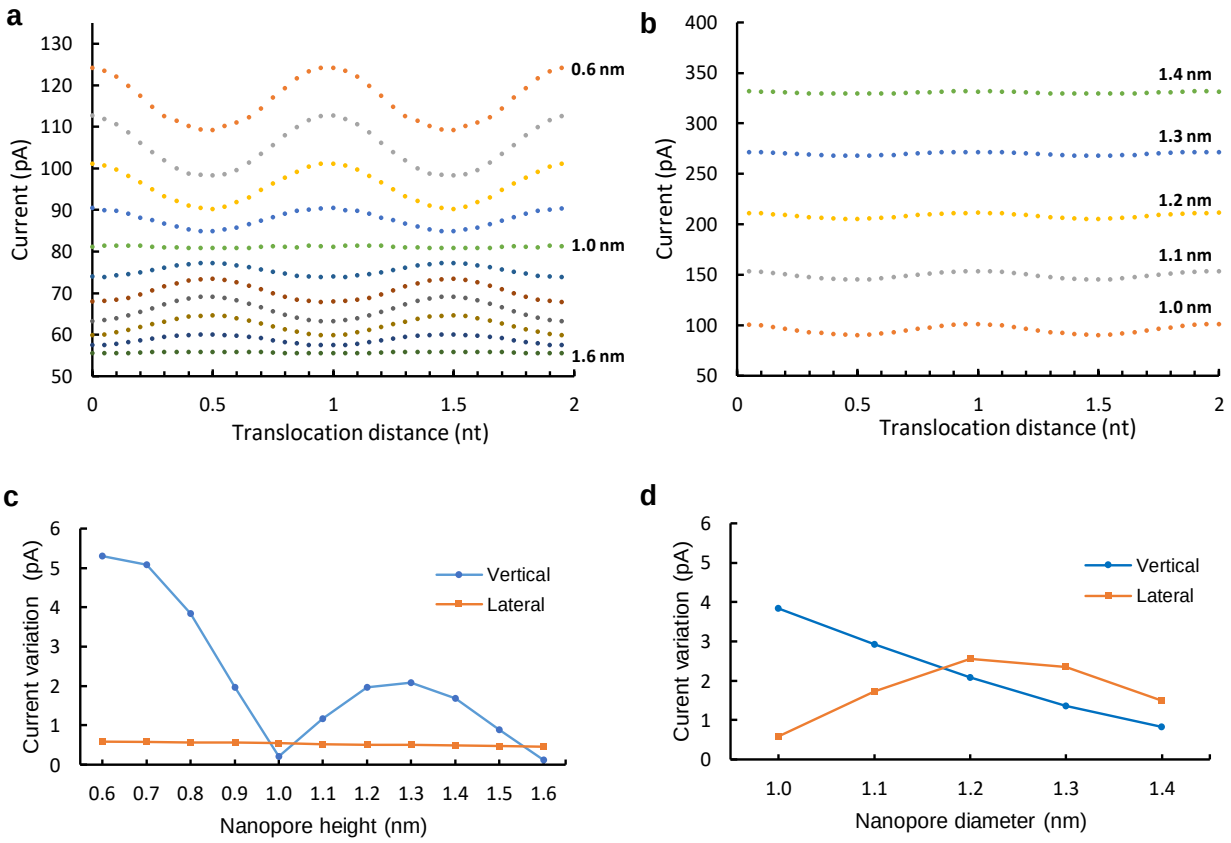


Figure 2.3 Influence of nanopore geometries on detecting nucleotide-to-nucleotide translocation.

The nanopores are modeled as circular cylinders. The homopolymer $(dC)_9$ is used as the DNA model and dC is modeled as an oval cylinder. (a) Influence of nanopore height on current profile of homopolymer translocation. The diameter of the nanopore is kept constant at 1.0 nm while the height is varied from 0.6 nm to 1.6 nm (top to bottom in 0.1 nm step). (b) Influence of nanopore diameter on current profile. The height of the nanopore is kept constant at 0.8 nm while the diameter is varied from 1.0 nm to 1.4 nm. (c) Current signal variation vs. nanopore height. Root mean squares of the sine wave signals in (a) are shown along the standard deviations of signal variations due to lateral movements. (d) Current signal variation vs. nanopore diameter. Root mean squares of the sine wave signals in (b) are shown along the standard deviations due to lateral movements.

2.5 Methods

2.5.1 Materials and general methods

Finite element analysis (FEA) was performed using a Multiphysics software package (COMSOL, Inc.). The nanopore sequencing data acquired using MinION R9.4 chemistry were obtained from Jain et al(33). We used the current levels for DNA hexamers and current signal standard deviations, and software (Tombo, Scrappie, Metrichor and Guppy V2.1.3) provided by ONT for our analysis(45). Custom programs were written using MATLAB (MathWorks, Inc.) or Python.

2.5.2 Physical model of mutant CsgG nanopore

The geometry for the engineered mutant CsgG nanopore was extracted based on the X-ray crystal structure of the wild-type protein and the open-pore current of the mutant protein. First, a surface contour map was taken from the crystal structure of *Escherichia coli* CsgG (Supplementary Fig. 2.1a)(39). The fine contour lines were then simplified by representing the curved surfaces of the two constriction areas with smooth arcs and the surfaces of the other less curvy areas with straight lines. Finally, the contour of the nanopore was approximated by slightly changing the curvatures of the two constriction areas such that the open-pore current of our model is in agreement with the experimental value of the mutant CsgG protein pore, which was approximated from raw sequencing data (about 200 pA under typical operation conditions of 300 to 500 mM KCl and 180 to 250 mV applied voltage across the nanopore)(8, 33). The open-pore current of our CsgG model was computed using FEA and the PNP equation system described below.

2.5.3 Physical model of nanopore device and calculations of ionic currents

The overall device is constructed as a cylinder of 200 nm in both diameter and height, which is divided into two equal chambers (*cis* and *trans*), with a nanopore protein positioned at the center separating the chambers (Supplementary Fig. 2.1b). FEA was used to calculate the electromagnetic field and ion fluxes through the pore. The potential and ions concentrations were determined using the Poisson equation:

$$\nabla^2\phi = -F(C_{K^+} - C_{Cl^-})/\varepsilon_0\varepsilon_r,$$

where ϕ is the potential inside the nanopore device, F is the Faraday constant, C_{K^+} is the concentration of K^+, C_{Cl^-} is the concentration of Cl^-, ε_0 is the dielectric permittivity of vacuum, and ε_r is the relative dielectric permittivity of the medium. Nanopore surface charge may influence the flux of ions. In our model, however, the nanopore surface is assumed neutral. Both the *cis* and *trans* chambers contain 400 mM KCl and a 180 mV potential is applied across the chambers. Water in confined space has been reported to have a very low dielectric permittivity ($\varepsilon_r = 2$)(46). DNA in free solution also has a low dielectric permittivity ($\varepsilon_r = 6$)(47). Its permittivity in confined space has not been determined, but is likely very low as well. Therefore, we used a low relative permittivity ($\varepsilon_r = 2$) for both water and DNA inside the nanopore region.

The ion fluxes were computed using the Nernst-Planck Equation:

$$\vec{J} = -D_{diff}\left(\nabla C + \frac{eZC}{k_BT}\nabla\phi\right),$$

where $\vec{J}$ is the ion flux, D_{diff} is the effective diffusion coefficient, C is the concentration of ions, e is the elementary charge, Z is the ionic charge, k_B is the Boltzmann constant, and T is the absolute temperature. The diffusion coefficients of K^+ and Cl^- were set to 1.95×10^{-9} m^2/s and 2.03×10^{-9} m^2/s, respectively. T was set to 293.15 K. The hydrodynamic radii of K^+ and Cl^- were assumed equal with a value of 3.3 Å as reported(48). When the distance of the ion is one hydrodynamic radius or greater from the nanopore surface, the diffusion coefficient is expected to be constant as in bulk solution (D^0_{diff}). As the ion approaches the surface of the nanopore, its diffusion coefficient is expected to decrease due to increasing friction. The effective diffusion coefficient (D_{diff}) was modeled as a function of distance (d) from the nanopore surface: $D_{diff} = \alpha D^0_{diff}$, where $\alpha = d/3.3$ for d < 3.3 Å and α = 1 for $d \geq$ 3.3 Å.

To minimize computations, the models were discretized using mesh sizes with finer meshes for the locations closer to the surfaces of the DNA and nanopore, and with larger mesh sizes for other areas. Since our nanopores and cylinder models of DNA have cylindrical symmetry around the vertical axis through the center, only 2D models are required for FEA. The ion fluxes for the 3D models were calculated by integrating across the vertical cross-sectional area of the nanopore. We calculated the ionic fluxes using five cross-sectional areas perpendicular to the vertical axis of the nanopore between the deoxyribonucleosides and phosphodiester linker in the center of the pore. The median value of the five calculated fluxes was used to represent the ionic current level. The calculated current was then normalized by multiplying its value with a correction factor of 0.89, which is the ratio of the open pore current of experimental measurement (200 pA) to that calculated from our model (224 pA). The use of 2D models for finite element analysis dramatically reduced the amount of computations required,

allowing us to model tens of thousands of configurations using a desktop computer (Intel Core i9 4 GHz processor with 16 cores/32 threads running and 128 GB of DDR4 RAM). The extensive modeling would not be feasible using all-atom molecular dynamic simulations.

2.5.4 Modeling of DNA as a linear structure of connected circular cylinders

Briefly, we used FEA and PNP system to compute the ionic current levels of ten thousand models of DNA hexamers with various combinations of putative diameters and heights through the mutant CsgG nanopore. We developed an algorithm for accurate prediction of the current levels of any hexamer (Supplementary Fig. 2.2). An optimal set of diameters and heights of the nucleosides and phosphodiester linker was then obtained by evolving the dimensions of the DNA hexamers such that their current levels best fit the corresponding experimental data (Supplementary Fig. 2.3).

We first investigated whether DNA can be approximated as a simple circular cylinder model with axial symmetry, and whether the dimensions of the nucleosides and phosphodiester linker can be extracted from the current levels of 4096 hexamers that are determined from experimental nanopore DNA sequencing data. We modeled DNA as a rigid linear structure of connected cylinders. The deoxyribonucleosides (dA, dC, dG and dT) were modeled as cylinders with the same height, but of different diameters. The phosphodiester linker was also modeled as a cylinder. Our aim was to determine the optimal set of diameters and heights based on experimental data such that the computed ionic current levels of all 4096 hexamers through the CsgG nanopore have a good correlation to those measured experimentally. The task turned out to be quite challenging due to the extensive modeling required to explore the large parameter space (millions of combinations). Therefore, we developed a genetic algorithm to overcome the challenge.

To reduce the amount of modeling required, we first established a method for predicting current level of any DNA hexamer. Ten thousand models of hexamers with various combinations of putative diameters and heights were constructed and their ionic current levels through the CsgG nanopore were determined using FEA and PNP equation system. The diameters and heights were randomly sampled within a reasonable ranges estimated from the crystal and chemical structures of DNA. The model dataset (*M*) contains the height and diameter parameters, and the current levels of the hexamers:

$$M = \{(D_1, I_1), (D_2, I_2), ..., (D_i, I_i), ..., (D_{10000}, I_{10000})\},$$

where D_i is an instance of the 10,000 models and I_i is the corresponding current level. D_i is represented as:

$$D_i = \{h_{i-Base}, h_{i-Linker}, d_{i1}, d_{i2}, d_{i3}, d_{i4}, d_{i5}, d_{i6}, d_{i-Linker}\},$$

where $h_{i\text{-}Linker}$ is the linker height, h_{Base} is the height of the deoxyribonucleosides, d_{i1} to d_{i6} are the diameters at the 1st to 6th positions along the hexamer, and $d_{i\text{-}Linker}$ is the diameter of the phosphodiester linker. The model dataset was used to predict the current level of any hexamer with any dimensions based on the similarity of the dimensions of its constituents to those in the model dataset. Since the DNA base closer to the narrowest constriction site of the nanopore contributes more to current blockade than the ones further away, the contributions of the hexamer $x = \{h_{Base}, h_{Linker}, d_1, d_2, d_3, d_4, d_5, d_6, d_{Linker}\}$ at different positions were weighted by a factor $f = \{\frac{\partial I}{\partial h_{Base}}, \frac{\partial I}{\partial h_{Linker}}, \frac{\partial I}{\partial d_1}, \frac{\partial I}{\partial d_2}, \frac{\partial I}{\partial d_3}, \frac{\partial I}{\partial d_4}, \frac{\partial I}{\partial d_5}, \frac{\partial I}{\partial d_6}, \frac{\partial I}{\partial d_{Linker}}\}$. The partial derivative is taken at the center of the parameter space. The algorithm was implemented using MATLAB as described below.

Algorithm 2.1 Algorithm for predicting the current level of any DNA hexamer

Input: Dimensions of nucleosides and linkers of a given hexamer (x), weight factor f, and the model dataset (M):

$$x=\{h_{Base}, h_{Linker}, d_1, d_2, d_3, d_4, d_5, d_6, d_{Linker}\},$$

$$M=\{(D_1, I_1), (D_2, I_2), ..., (D_i, I_i), ..., (D_{10000}, I_{100000})\},$$

where $D_i=\{h_{i-Base}, h_{i-Linker}, d_{i1}, d_{i2}, d_{i3}, d_{i4}, d_{i5}, d_{i6}, d_{i-Linker}\}$

Output: Predicted current level (I_x) of the input hexamer x.

Outline of algorithm:

For all D_i in M

Calculate the dimensional similarity distance between x and D_i using similarity distance (SD):

$SD_i=\sqrt{\sum_{j=1}^{9} f_j(x_j-D_{i,j})^2}$, where the subscript j represents each of the 9 corresponding sub elements in f, x and D_i.

End

Identify the k hexamers that are most similar to x based on similarity (with the smallest SD values, k = 7);

Compute the current level (I_x) for x by averaging the current levels of the most similar hexamers.

The accuracy of our prediction algorithm was assessed by leave-one-out cross validation. The correlation between the predicted and the model current levels is excellent with a root mean square error of 1.90 pA and R^2 of 0.996 (Supplementary Fig. 2.2).

Next, we developed a genetic algorithm to evolve a set of parameters for the diameters and heights of the deoxyribonucleosides and phosphodiester linker using experimentally determined current levels of 4096 DNA hexamers. The experimental data set consists of:

$$E=\{(E_1, I_{E1}), (E_2, I_{E2}), ..., (E_i, I_{Ei}), ..., (E_{4096}, I_{E4096})\},$$

where E_i is one of the 4,096 DNA hexamer and I_{Ei} is its corresponding experimental current level. The experimental current data were provided online by ONT(40). Our initial attempt to use only one diameter to model each nucleobase gave reasonable correlation between the model and

experimental current levels ($R^2 \approx 0.5$). However, it is well known that each nucleobase interacts with its two nearest neighbors, thus the effective hydrodynamic radius of each nucleobase is expected to vary slightly depending on its two nearest neighbors. Therefore, we optimized the diameters by modeling each base with 16 slightly different diameters depending on its two nearest neighbors.

The parameter space consists of the diameters and heights for four deoxyribonucleosides and a phosphodiester linker. Considering nearest neighbor effect, the entire set of parameters can be represented by $\alpha = \{H_{base}, H_{Linker}, D_{Linker}, (D_A)_{1-16}, (D_C)_{1-16}, (D_G)_{1-16}, (D_T)_{1-16}\}$, where H_{Base} is the height of the nucleosides, H_{Linker} is the height of the phosphodiester linker, D_{Linker} is the diameter of the phosphodiester linker, and $(D_A)_{1\text{-}16,}$ $(D_C)_{1\text{-}16,}$ $(D_G)_{1\text{-}16}$ and $(D_T)_{1\text{-}16}$ are the diameters of the dA, dC, dG and dT, respectively, each with 16 different nearest neighbor pairs. Each individual in the population P consisting of 4096 hexamers is represented as $P_i = \{X_1, X_2, X_3, X_4, X_5, X_6, \}$, where X_1 *to* X_6 is any one of the nucleosides (dA, dC, dG or dT). First, the parameter set α was initialized with a random set of values roughly within the ranges estimated from the chemical structures of DNA. The current level for each of the hexamer P_i was calculated using the 10,000-model dataset and the current prediction algorithm (Algorithm 2.1) as described above. The fitness of the parameters was evaluated by the sum of the root mean square errors between the predicted and the experimental current levels for all hexamers. The set of parameters giving the smallest root mean square error has the best fitness. The population was then updated with the parameters that gave the best fitness and the evolution process continued until convergence or termination. The algorithm was implemented using MATLAB as follows.

Algorithm 2.2 Genetic algorithm for optimizing the diameters and heights of deoxyribonucleosides and phosphodiester linker based on experimental data

Input: A reference experimental dataset (E) and 10k model dataset (M):

$$E=\left\{(E_1,I_{E1}),(E_2,I_{E2}),...,(E_i,I_{Ei}),...,(E_{4096},I_{E4096})\right\}$$

$$M=\left\{(D_1,I_1),(D_2,I_2),...,(D_i,I_i),...,(D_{10000},I_{10000})\right\},$$

Output: The diameters and heights of the cylinders used to represent the four deoxyribonucleosides and the phosphodiester linker. To account for nearest neighbor effect, a set of 16 slightly different diameters is determined for each nucleoside.

Outline of the algorithm:

Initialize the first-generation population using a set of parameters;
While the termination condition of the algorithm is not satisfied **do**
  For each individual P_i in the population **do**
    Generate the dimensional parameters of all hexamers;
    Predict the current levels of all hexamers using the prediction algorithm (Algorithm 1) and model data set M;
    Compute sum of the root mean square errors between predicted and experimental current levels as a fitness function using the predicted current levels and experimental data set E;
  End
  Update the next generation of population based on the fitness function;
End
Output the optimized set of parameters.

After the extraction of a set of optimal parameters for the cylinder models of the nucleosides and phosphodiester linker (Supplementary Tables 2.1 and 2.2), the current levels for all DNA hexamers through the MinION CsgG nanopore were determined using FEA and PNP equation system as described earlier. The correlation between the modeling results and experimental data was calculated (Supplementary Fig. 2.3).

2.5.5 Modeling of DNA as connected oval cylinders

Even though the four deoxyribonucleosides share the same deoxyribose moiety, each nucleoside has a different size and shape. To provide a more realistic model for the DNA bases, we modeled each nucleoside with an oval cylinder with one axis of reflection symmetry. The

shape of each nucleoside was constructed using three circular arcs based on several constraints (Supplementary Figs. 2.4-2.6) as follows: 1) the cross sectional area be equal to that of the cylinder models; 2) the ionic current level be equal to that of the cylinder models; 3) the diameter of one of the arcs (r_s) be equal to the diameter of the cylinder representing the phosphodiester linker, 4) the distance between the centers of the two arcs on the long axis (d) be chosen such that the standard deviation of the signal variations due to the movements of the center of mass of the model along the long axis from the center of the nanopore is close to the same-level standard deviation of the experimental data.

First, r_s was set to be equal to the radius of the phosphodiester linker determined earlier using the cylindrical models. Various values were assigned to d to calculate r_b and R based on the constraint that the total cross sectional area of the oval be equal to that of the circular cylinder model, and the current level be equal to that of the cylinder model. A 3D model was constructed for each homopolymer hexamer. The current level of the hexamers with center of mass positioned at several different distances from the horizontal center of the CsgG nanopore were computed using FEA and PNP equation system. Since the current levels for equal displacement of center of mass of the oval along both the long and short axes from the nanopore center are found to be almost identical (Supplementary Fig. 2.5b), the standard deviations of the current signals due to lateral movements from the nanopore center can be calculated. If we assumed each position in the horizontal cross section of the nanopore is accessible to the DNA hexamer with equal probability, the standard deviation of current variations due to lateral movements can be determined by:

$$\sigma = \sqrt{\int_0^{l_{\max}} \frac{2lI_l^2}{l_{\max}^2} dl - \left(\int_0^{l_{\max}} \frac{2lI_l}{l_{\max}^2} dl \right)^2},$$

where l is the lateral displacement from the nanopore center (at which $l = 0$), l_{max} is the maximum accessible displacement, I_l is the current level when the DNA hexamer is laterally displaced l distance from the pore center. For each homopolymer hexamer with the various putative d values, the standard deviations of current variations were computed. The optimal set of parameters (r_s, r_b, d and R) was then determined by minimizing the difference between the standard deviations of the model and experimental data (Supplementary Fig. 2.6, Supplementary Table 2.3).

2.5.6 Modeling of nucleotide-to-nucleotide translocation of DNA homopolymers through nanopores

The nucleotide-to-nucleotide translocation of four different DNA homopolymers through the mutant CsgG nanopore were simulated by computing the current levels of homopolymers positioned at 20 consecutive locations along the vertical axis of the nanopore (Fig. 2.1a). The nucleosides were modeled as oval cylinders with their cross sectional center of mass positioned along the nanopore vertical axis. In addition to electric-field-driven translocation along the vertical axis of the nanopore, the DNA molecule may also undergo rapid rotation and lateral translational movements inside the nanopore due to thermal kinetic energy. These thermal motions also introduce variations in the current levels. Therefore, we also computed the current level variations of the homopolymers as described above (Supplementary Table 2.4). For comparison, we also computed the current profiles for homopolymers modeled as circular cylinders. The profiles are also identical to the profiles obtained with the oval cylinder models (Supplementary Fig. 2.7).

2.5.7 Mapping of homopolymer sequencing data to sine wave functions

We extracted the raw current profiles of 25622 sequences containing between 20-40-base homopolymers on chromosome 20 of the human genome that has been sequenced using MinION R9.4 chemistry for our analysis(33). The raw current signals were aligned to the genome sequences using Tombo and decoded using Metrichor and Guppy. The decoded sequences were aligned to the reference genome sequences based on the time stamps in the processed data. Prior to mapping, the baselines for the current profiles and the signal boundaries of the homopolymers were determined. The baselines were calculated by averaging using a window of 11 data points. The boundaries of the current signals for the homopolymers were first estimated using Tombo and Metrichor segmentation data, and then further narrowed down based on the variation of several nearby data points. Two bases at both ends of the signal were not included in the analysis to exclude the influence of the nearby non-homopolymer sequences.

The experimental data were acquired at a fixed frequency of about 4 kHz and a read rate of about 450 nt/s (~9 data points per base), but the translocation process is stochastic due to thermal motions and stepwise behavior of the helicase. Therefore, the data points acquired during the translocation from one nucleotide to another are spatially warped, and therefore are not expected to fit literally to a sine wave. To investigate whether the data follow a sinusoid pattern if the time-serial data points are mapped or unwarped into spatial data with equal spacing in distance, we developed a method based on the Viterbi algorithm for mapping the experimental data to sine wave functions.

The experimental time-serial data for each homopolymer is represented by: $I = \{I_1, I_2, \ldots, I_i, \ldots, I_L\}$, L is the total number of data points. The associated baseline data is represented by: $B = \{B_1, B_2, \ldots, B_i, \ldots, B_L\}$. The sinusoid model data of an N base long homopolymer represented by

N sine wave periods were digitized into data points with equal spatial spacing: $M = \{S_1, S_2, \ldots, S_j, \ldots, S_z\}$, where $z = N \times k$ and k is the number of data points per sine wave period. To account for potentially large warping in the experimental data, we used a large number of data points per sine wave period (k = 100). We assumed that DNA translocation velocity for the temporal data points collected at fixed 4 kHz frequency follows an inverse Gaussian distribution function. The sine wave was assumed to have an amplitude of 2.2 pA, which is approximately equal to the lowest value determined from the experimental and modeling data (Supplementary Table 2.5). The algorithm was implemented using Python as described below.

Algorithm 2.3 Viterbi algorithm for mapping experimental data to sine wave functions

Input: 1) Experimental current data *I* and baseline data *B*, both with *L* data points;

2) Model data *M* of *N* sine wave periods, each of which has *k* data points;

3) Standard deviation (σ) of the current signal variation due to thermal lateral movements;

4) Specified mean (μ) and shape parameter (λ) for the inverse Gaussian function ($\varphi_{g,m}$);

5) Frequency of data sampling rates (*f*).

Output: Map the experimental temporal data points onto the periods of a sine wave function with equal spacing in distance.

Outline of algorithm:

Initialize: map the first signal I_1 onto the first sine wave period.

for $i = 1$ to k, calculate

$$p_{I_1}^{i} = \exp\left(-\frac{(I_1 - B_1 - S_i)^2}{2\sigma^2}\right)$$

For $i = 2$ to L **do**

For $m = i$ to $k \times i$, ($m = k \times N$ if $i > N$), determine

$$p_{I_i}^{S_m} = \max_{g}\left(p_{I_{i-1}}^{S_g} \varphi_{S_g}^{S_m} \phi_{I_i,B_i}^{S_m}\right) \text{ for } g \text{ from } (m-k) \text{ to } (m-1), \ (g = 1 \text{ if } (m-k) < 1),$$

where

$$\varphi_{S_g}^{S_m} = \sqrt{\frac{k^3\lambda}{2\pi f^3 (m-g)^3}} \exp\left(-\frac{\lambda(fm - fg - k\mu)^2}{2\mu^2 (m-g) kf}\right),$$

$$\phi_{I_i,B_i}^{S_m} = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(I_i - B_i - S_m)^2}{2\sigma^2}\right);$$

$$q_{I_i}^{S_m} = \arg\max_{g}\left(p_{I_{i-1}}^{S_g} \varphi_{S_g}^{S_m} \phi_{I_i,B_i}^{S_m}\right) \text{ for } g \text{ from } (m-k) \text{ to } (m-1), \ (g = 1 \text{ if } (m-k) < 1);$$

End

End

$$Q_L = \arg\max_{g}\left(p_{I_L}^{S_g}\right) \text{ for } g \text{ from 1 to } k \times N$$

For $i = L - 1$ to 1, calculate

$$Q_i = q_{I_{i+1}}^{S_{Q_{i+1}}}$$

End

Assign each experimental data point ($I_1, I_2, \ldots, I_i, \ldots, I_L$) onto a position of a sine wave function $(Q_1, Q_2, \ldots, Q_i, \ldots, Q_L)$.

Up on close examination of the raw current signals, we found that the velocity distribution, noise level and baseline signal could vary substantially for individual homopolymer region.

Therefore, we accounted for the local variations by refining the parameters based on the current profile of the individual homopolymer region. The mapping was performed iteratively using the algorithm. In the first iteration, a set of average parameters that were determined from all the data (μ = 450 nt/s, λ = 800 nt/s and σ = 0.9 pA) was used for the mapping. The parameters (μ, λ and σ) were then refined based on the mapped positions of the experimental data points relative to the sine wave periods and another round of mapping was performed. The process was repeated until the mean square error between experiment current levels and sine wave function could not be minimized any further. The coefficient of determination ($R^2 = 1 - \sum(I_{Measured} - I_{Model})^2 / \sum(I_{Measured} - I_{Baseline})^2$) was calculated and used as one measure of mapping quality score. In addition, the quality of the mapping was also quantified by the distribution of the data points in 6 equally spaced sub-segments of each sine wave period: good ($\geq$ 1 data point in at least 5 sub-segments), fair ($\geq$ 1 data point in 3 or 4 sub-segments), and poor ($\leq$ 2 sub-segments have data points) (Supplementary Fig. 2.10)

To further validate our approach, we also mapped simulated evenly spaced sine wave data with a similar amplitude (2.2 pA) and added signal variation (standard deviation σ = 0.9 pA), and simulated data with random noise (normal distribution with standard variation σ = 2.0 pA) (Supplementary Fig. 2.9g, h)

2.5.8 Translocation velocity distribution

We determined the mean and variation of translocation velocity using both homopolymer mapping data and 1 million base randomly selected from the experimental sequencing data (Supplementary Fig. 2.11a-d). For the homopolymers, we extracted the number of data points in each sine wave period (and thus nucleotide) to obtain the distribution (data points/nt), and

calculated the translocation velocity distribution. For the other sequencing data, per-base data points were determined based on the segmented current levels (using the software Tombo). The distribution was then transformed into velocity distribution. The means and standard deviations of the distributions were then determined.

2.5.9 Simulations of nanopore sequencing of human genome

We expect that nucleotide-to-nucleotide transition events of all DNA hexamers follow a sinusoid. If observable and considered, the sinusoid patterns not only allow for the accurate counting of the numbers of bases in homopolymer regions, but also enable higher single-pass basecalling accuracy and consensus accuracy. To simulate the current profiles associated with the nucleotide-to-nucleotide transition events requires the construction and analysis of a large number of 3D models for all 4096 DNA hexamers (20 x 4096), which is computationally very demanding or impractical.

To make our modeling computationally manageable, we used several strategies to simplify our modeling and computations. First, we used circular cylinder models to represent the DNA hexamers and pentamers. Because of the axial symmetry of the circular cylinder models, the current levels of the hexamers and pentamers through CsgG nanopore can be modeled and computed using 2D modeling. We found that there is a good correlation between the model and experimental data ($R^2 = 0.91$, Supplementary Fig. 2.3) and the current profiles of the cylinder models and the oval models of the four DNA homopolymers are almost identical (Fig. 2.1b and Supplementary Fig. 2.8). Second, the current signal variation due to lateral movements and measurement electronics noise is approximated using the mean of the standard deviations (0.9 pA) of the experimentally determined values of homopolymers. The value is slightly higher than the mean value computed (0.77 pA) based on the oval cylinder models of four DNA

homopolymers (Supplementary Table 2.4). Third, to model the nucleotide-to-nucleotide translocation of all 4096 hexamers, we constructed 11 models for each transition from one hexamer to another hexamer by positioning the models at 11 equally spaced steps along the vertical axis of the CsgG nanopore. The current levels of all the models (4096 x 11) were computed using FEA and the PNP equation system. The root mean square amplitude of each sine wave associated with each transition event was determined as square root of the difference between the current level of the pentamer and the mean value of the two hexamers. For the computational sequencing experiments, we simulated the current profiles using the same CsgG nanopore employed in the ONT MinION R9.4 chemistry and the same signal variation due to lateral movements (0.9 pA), but at a higher measurement bandwidth (20 kHz). The stochastic behavior of translocation velocity was modeled using an inverse Gaussian distribution:

$$f(v;\mu,\lambda)=\sqrt{\frac{\lambda}{2\pi v^3}}\exp\left(-\frac{\lambda(v-\mu)^2}{2\mu^2 v}\right),$$

where v is the DNA translocation velocity (nt/s), μ is the mean velocity, and λ is the shape parameter that is determined by the standard deviation of the velocity distribution ($\lambda=\mu^3/\sigma_v^2$). The mean and standard deviation of translocation velocity (μ = 450 nt/s and σ_v =239 nt/s) were chosen to mimic closely the experimental conditions of the ONT MinION platform (Supplementary Fig. 2.11a-e). Signal variations due to lateral movements were simulated using a Gaussian distribution. 140 million base pairs (140,000 sequences of 1 kbp fragments) were randomly selected from the human genome (GRCh38) and sequenced at 20x depth (10 each of the two complementary strands).

Next, we developed an algorithm to decode the simulated current profiles into DNA sequences. First, the simulated data were processed using a bandpass filter with a bandwidth of 400 Hz to 500 Hz (using *bandpass* function in MATLAB). Second, the time stamps of the local maximum and minimum values on the filtered signal profiles were used to estimate the positions of current levels corresponding to the pentamers and hexamers. The current levels of the pentamer and hexamer were determined by averaging three current levels (data points) around the crest and trough of each sine wave period, respectively. Third, the current profiles were decoded into DNA sequences using a Viterbi algorithm by finding the optimal paths through the data represented as a 3-state HMM (Fig. 2.1c, Fig. 2.2c, Supplementary Fig. 2.13). In our implementation, we treated the transitions from one state to another differently. Only one possibility is allowed for a transition from a hexamer to a pentamer, but four possible transitions are allowed for the transition from a pentamer to a hexamer. For example, the pentamer transits to hexamer ACGTAC must be ACGTA, and the pentamer ACGTA is allowed to transit to any one of four possible hexamers (ACGTAA, ACGTAC, ACGTAG and ACGTAT). The very ends of the profiles (20 bases at each end) were excluded from the analysis.

The simulated current profiles were processed and decoded. Single-pass accuracy was determined by aligning the decoded consensus sequences for the template and reverse complementary strands of each sequence separately by the progressive alignment method (using *multialign* function in MATLAB). The consensus sequences were then aligned to the reference sequences using the Smith-Waterman algorithm to identify errors. If a consensus basecall does not agree with the reference, the basecall is considered as an error (either mismatch, insertion or deletion). A basecall is considered correct if the consensus basecalls of both strands or the consensus basecall by one strand with a higher quality score agrees with the reference. The

quality score is calculated as the percentage of the consented calls out of the 10 sequencing experiments for each strand. In some rare cases, the consensus basecalls of the two complementary strands do not agree with one another and the percentages of both consensus basecalls out of the 10x sequencing experiments are the same. Out of the 140 million bp sequenced, we encountered 63 such cases. Very interestingly, we found that one of the complementary strands contains a sequence with a current profile which is essentially indistinguishable from another sequence combination, leading to the ambiguity in calling the profile into two different bases with almost equal probability (Supplementary Fig 2.14). In those cases, we calculated the RMS of the Euclidean distance between the current profiles of the two ambiguous calls. A higher quality score was assigned to the strand with a larger RMS distance since the accuracy of the call is greater for a sequence with a profile that is less similar to other sequence(s) (Supplementary Fig. 2.14). As expected, in all 63 cases, the RMS distances of the profiles of the putative pair of one strand are on the order of or less than the signal variation due to lateral movements and electronic noise (with a standard deviation of 0.9 pA), the profiles of the pair cannot be reliably differentiated, leading to ambiguity in basecalling. However, in all cases, the RMS distances of the profiles of the putative pairs of their corresponding complementary strands are much greater than 0.9 pA, enabling accuracy basecalling.

2.6 Conclusions

In conclusion, the use of simple cylinder models to represent DNA enabled us to perform extensive modeling of DNA translocation through nanopores using FEA, and the classical Poisson Nernst-Planck equation system. The work would not be feasible computationally using molecular dynamic simulations. Despite the simplicity of our models and the use of the continuum physics, the results from our modeling of current levels, signal variations due to

vertical translocation and lateral movements correlate well with experimental data. We showed that the current profiles of the nucleotide-to-nucleotide translocation of DNA through CsgG nanopore follow a sinusoid and that the behavior is observable in the experimental sequencing data. Moreover, our modeling also revealed the feasibility of accurate sequencing using the CsgG nanopore, and the bandwidth limitation of the current MinION platform. Using a 20 kHz measurement bandwidth and a 3-state HMM for sequence decoding, we performed computational sequencing of the human genome and demonstrated the feasibility of ultra accurate nanopore DNA sequencing (with single-pass accuracy of 99.4%, zero error in 140 million bp at 20x sequencing depth). Our modeling also provides insights into the physical origins of the current signal variations and the engineering of nanopores to improve the detection of nucleotide-to-nucleotide transition events and sequencing accuracy.

2.7 Supplementary information

Supplementary Table 2.1 Optimized dimensions of deoxyribonucleosides and phosphodiester linker based on experimental data

The dimensions were determined using only models and experimental data for four homopolymers, $(dA)_6$, $(dC)_6$, $(dG)_6$, $(dG)_6$, and $(dT)_6$.

Linker and Base	Height (Å)	Radius (Å)
Phosphodiester linker	2.99	2.31
dA	2.88	3.90
dC		3.65
dG		4.04
dT		3.81

Supplementary Table 2.2 Optimized dimensions of deoxyribonucleosides and phosphodiester linker with consideration of nearest-neighbor effect

The effective radius of each base is dependent on its two nearest neighbors. Sixteen values are determined for each nucleoside.

Radius (Å)		Neighboring nucleoside above				Neighboring nucleoside below
		dA	dC	dG	dT	
Nucleoside (in the middle)	**dA**	**3.90**	4.12	3.88	4.26	dA
		3.73	3.95	3.77	4.09	dC
		4.07	4.18	4.04	4.35	dG
		3.65	3.90	3.63	4.00	dT
	dC	3.23	3.48	3.26	3.85	dA
		3.46	**3.65**	3.41	3.90	dC
		3.33	3.63	3.41	3.91	dG
		3.31	3.57	3.26	3.74	dT
	dG	4.01	4.11	3.93	4.31	dA
		3.85	4.02	3.79	4.20	dC
		4.17	4.15	**4.04**	4.38	dG
		3.76	3.89	3.63	4.05	dT
	dT	3.07	3.23	3.20	3.74	dA
		3.08	3.32	3.24	3.79	dC
		3.33	3.34	3.33	3.80	dG
		3.28	3.48	3.31	**3.81**	dT

Supplementary Table 2.3 Dimensions of oval cylinder models of deoxyribonucleosides

Parameter	Dimensions (Å)			
	dA	dC	dG	dT
Oval long axis	4.34	4.32	4.36	4.33
Oval short axis	3.56	3.17	3.89	3.48
R	4.76	5.67	4.49	4.88
d	3.21	3.42	3.01	3.25
r_b	3.15	2.90	3.40	3.10
r_s	2.31			
Height	2.88			

Supplementary Table 2.4 Current signal variations due to lateral movements inside mutant CsgG nanopore
ND: not determined.

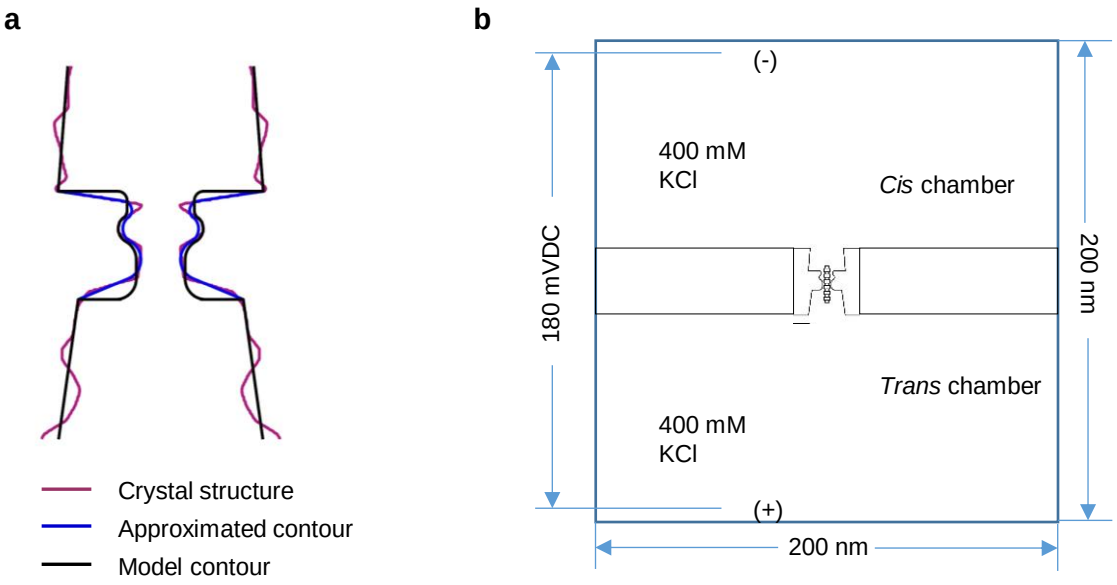
Measured or modeled		Standard deviation of signal variation (pA)			
		$(dA)_6$	$(dC)_6$	$(dG)_6$	$(dT)_6$
Experimental		0.94	0.80	0.95	0.91
Modeling	Oval cylinder	0.77	0.59	0.97	0.74
	Circular cylinder	ND	2.91	ND	ND
Electronic/amplifier noise		~1.0			

Supplementary Table 2.5 Current level variations of nucleotide-to-nucleotide translocation through mutant CsgG nanopore
The homopolymers are modeled as linear connected oval cylinders.

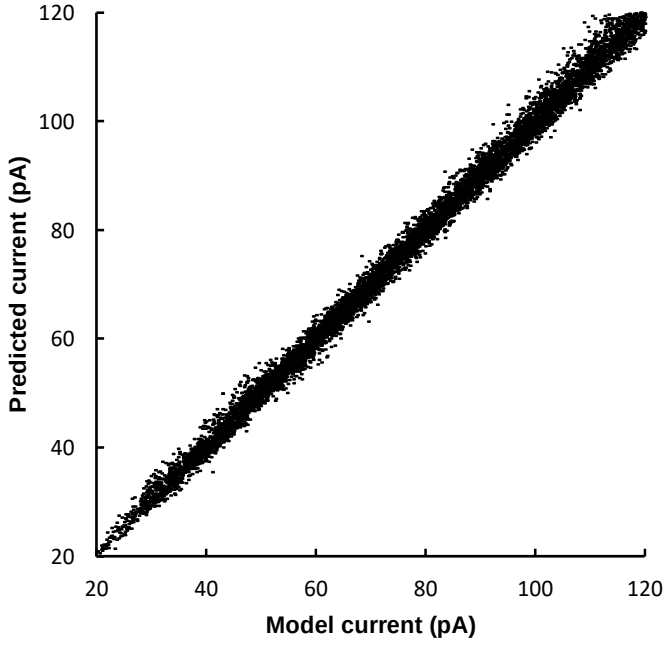
Measured or Modeled		$(dA)_6$	$(dC)_6$	$(dG)_6$	$(dT)_6$
Experimental	Mean current level (pA)	86.49	98.89	73.67	90.68
	Standard deviation of levels (pA)	1.52	1.40	1.51	1.51
	1/2 of peak to peak variation (pA)	2.15	1.98	2.14	2.14
Modeling	Mean current level (pA)	88.10	101.17	75.47	90.92
	Root mean square of sine wave (pA)	1.77	1.53	1.99	1.69
	Amplitude of sine wave (pA)	2.46	2.15	2.86	2.45

Supplementary Table 2.6 | Vertical and horizontal current variations for nanopores with various dimensions. The data were based on modeling of $(dC)_6$.

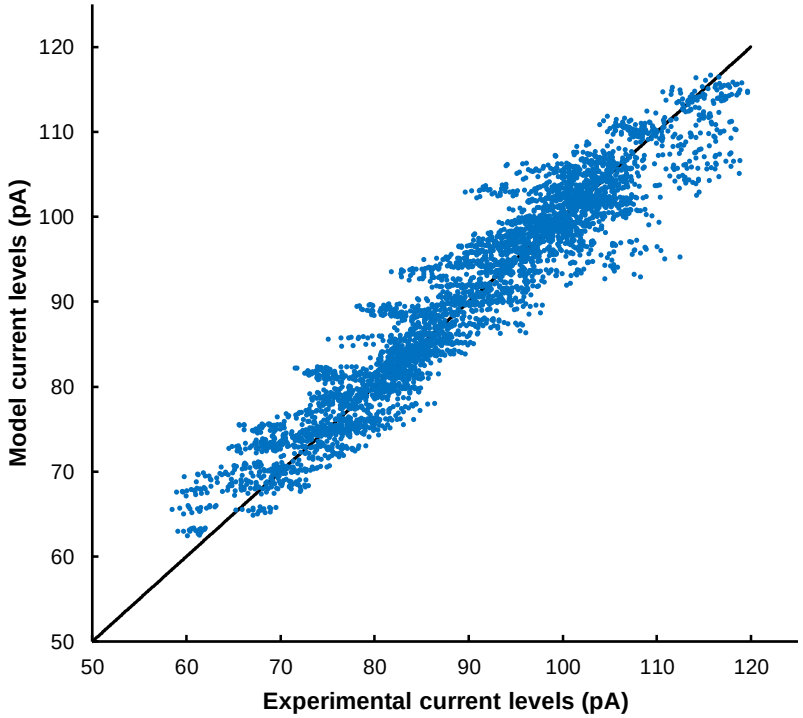
Nanopore Diameter (nm)	Nanopore Height (nm)	Variation due to translocation along vertical axis (pA)	Variation due to lateral horizontal movement (pA)
1.0	0.6	5.30	0.585
	0.7	5.08	0.573
	0.8	3.84	0.567
	0.9	1.97	0.558
	1.0	0.21	0.547
	1.1	1.17	0.524
	1.2	1.96	0.510
	1.3	2.08	0.504
	1.4	1.69	0.489
	1.5	0.89	0.472
	1.6	0.12	0.456
1.1	0.8	2.92	1.733
1.2		2.09	2.556
1.3		1.37	2.353
1.4		0.83	1.488



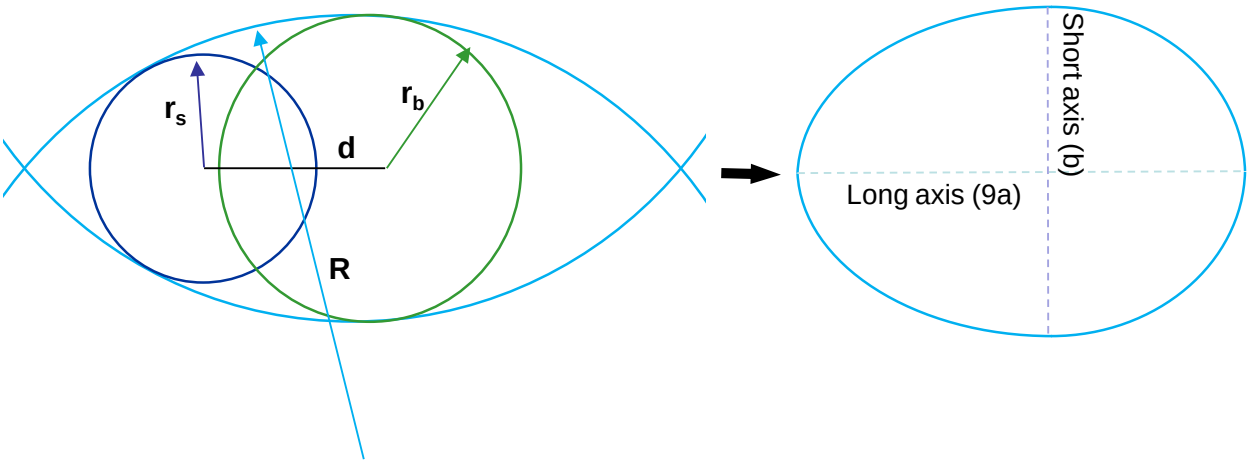
Supplementary Figure 2.1 Mutant CsgG nanopore geometry and nanopore device model. (a) The surface contour map was taken from the crystal structure of the wild-type protein. The 2D contour around the constriction site is approximated by three arcs, which were positioned 1Å from the surface contour of the crystal structure to account for mutations made to the wild-type CsgG nanopore protein. The narrowest constriction point of the nanopore geometry was estimated by modeling of current level using FEA and PNP equation system. (b) A model of the nanopore device used for FEA. Not drawn to scale.



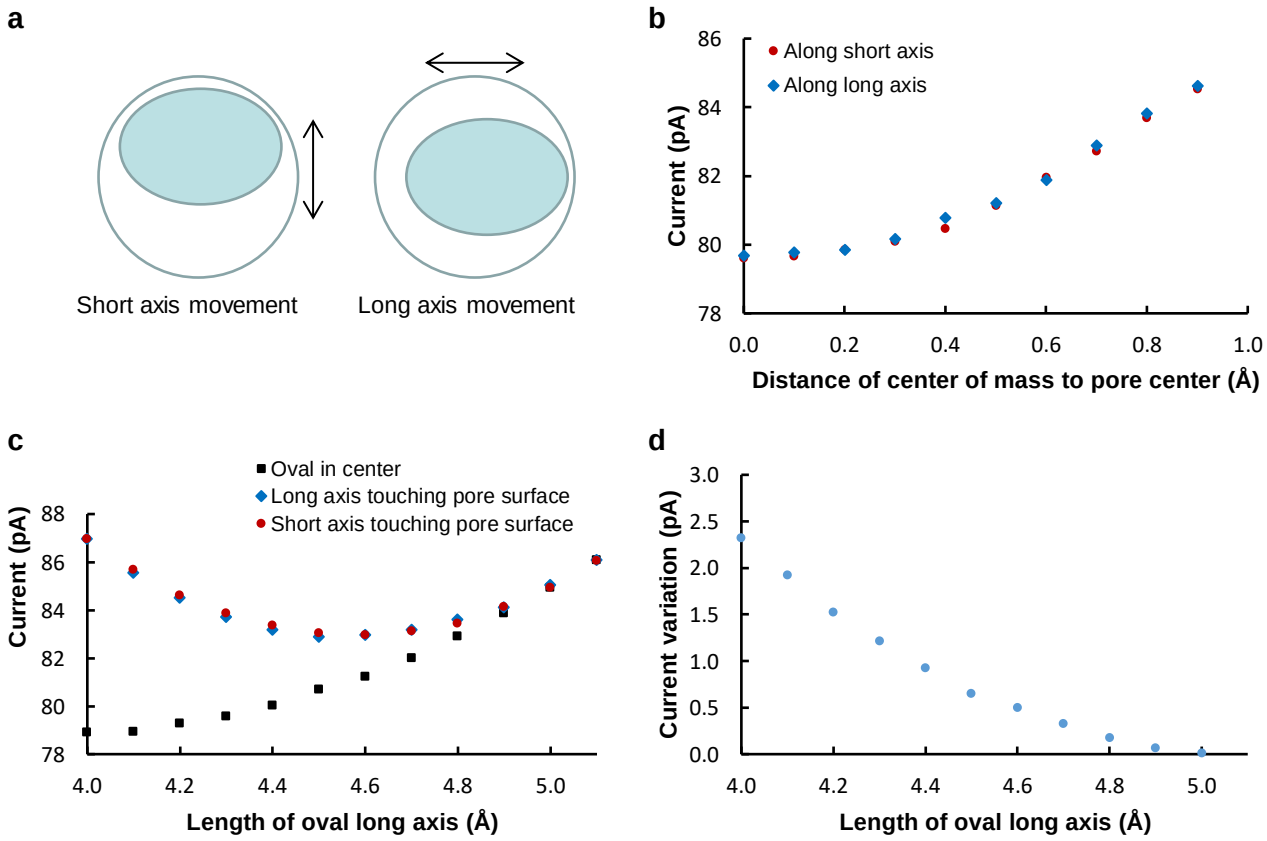
Supplementary Figure 2.2 Accuracy of algorithm for prediction of current levels of DNA hexamers.
The accuracy of our prediction algorithm was assessed by using leave-one-out cross validation. The correlation (R^2) and root mean square error between the predicted current levels and the model current levels were determined to 0.996 and 1.90 pA, respectively.



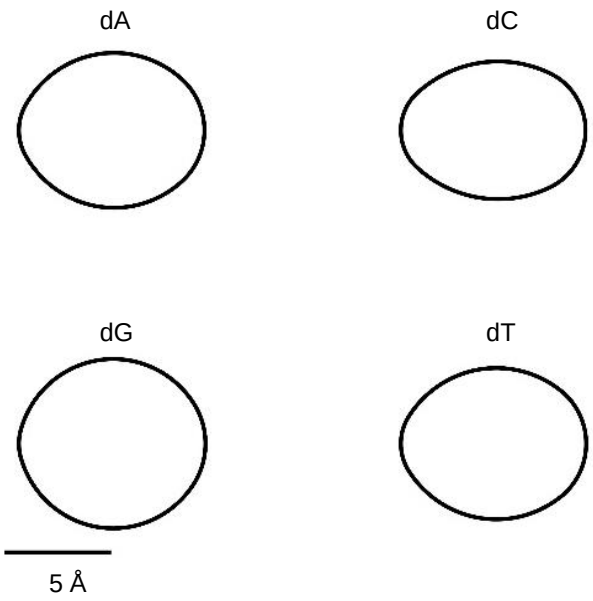
Supplementary Figure 2.3 Correlation of model and experimental current levels of DNA hexamers.
The 4096 hexamers were modeled as circular cylinders. The correlation between the modeling and experimental data is excellent ($R^2 = 0.91$).



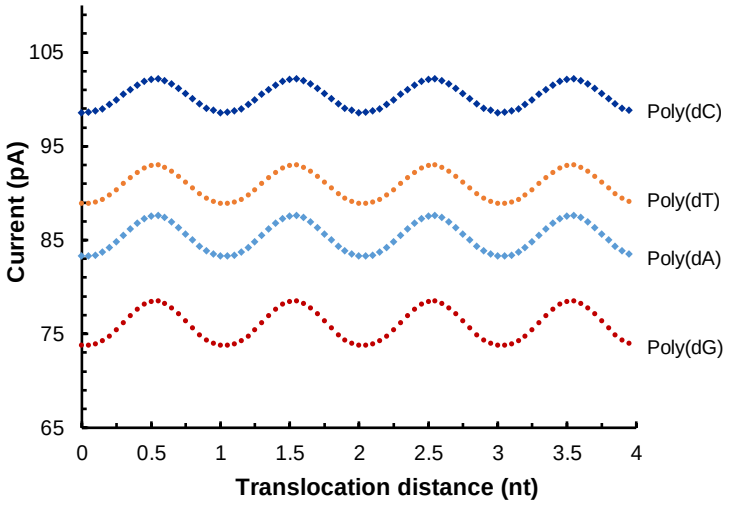
Supplementary Figure 2.4 Determination of the dimensions of the oval cross sections of deoxyribonucleosides.
The oval cross section for each nucleoside was constructed using three arcs with different radii. The dimensions of the arcs were optimized using three constraints: cross section area, current level and current signal variations due to the lateral movements from the nanopore center.



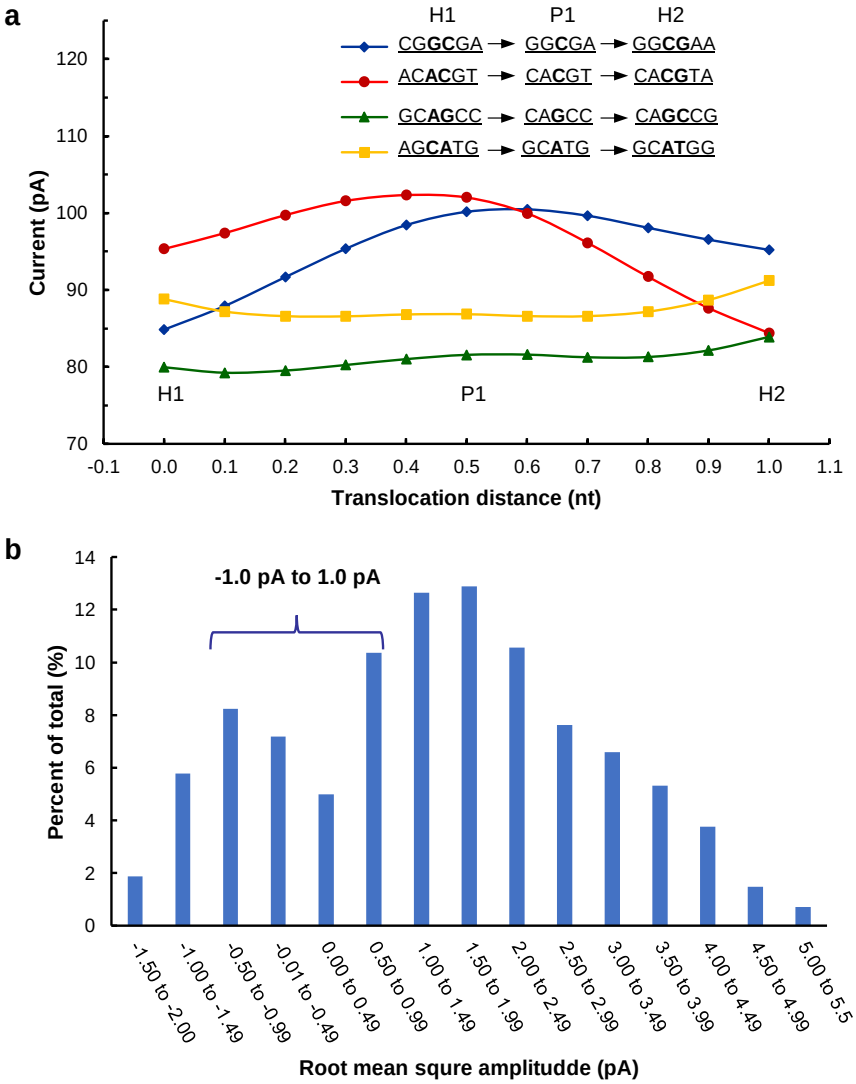
Supplementary Figure 2.5 Current variations due to lateral movements from nanopore center and the dimensions of models.
(a) Illustration of the lateral movements along the long axis or short axis of an ellipse inside a nanopore. **(b)** Changes in current level due to movements of the center of mass of an elliptical cylinder with cross section of 3.8 Å (short axis) x 4.2 Å (long axis) away from nanopore center. **(c)** Changes in current level when the cross section is morphed from a circle with 4 Å radius to elliptical cross sections of various long axis length while the total cross sectional area is maintained the same. **(d)** Standard deviations of current levels due to lateral movements of elliptical cylinders with various long axis lengths inside the nanopore. The cross section areas of the ovals were kept the same as the sectional area of the cylinder used to represent the deoxyribonucleoside (dG in this example).



Supplementary Figure 2.6 Modeling of deoxyribonucleosides as oval cylinders.
Show are the cross sections of the oval cylinders used to model the DNA nucleosides (dA, dC, dG and dT).

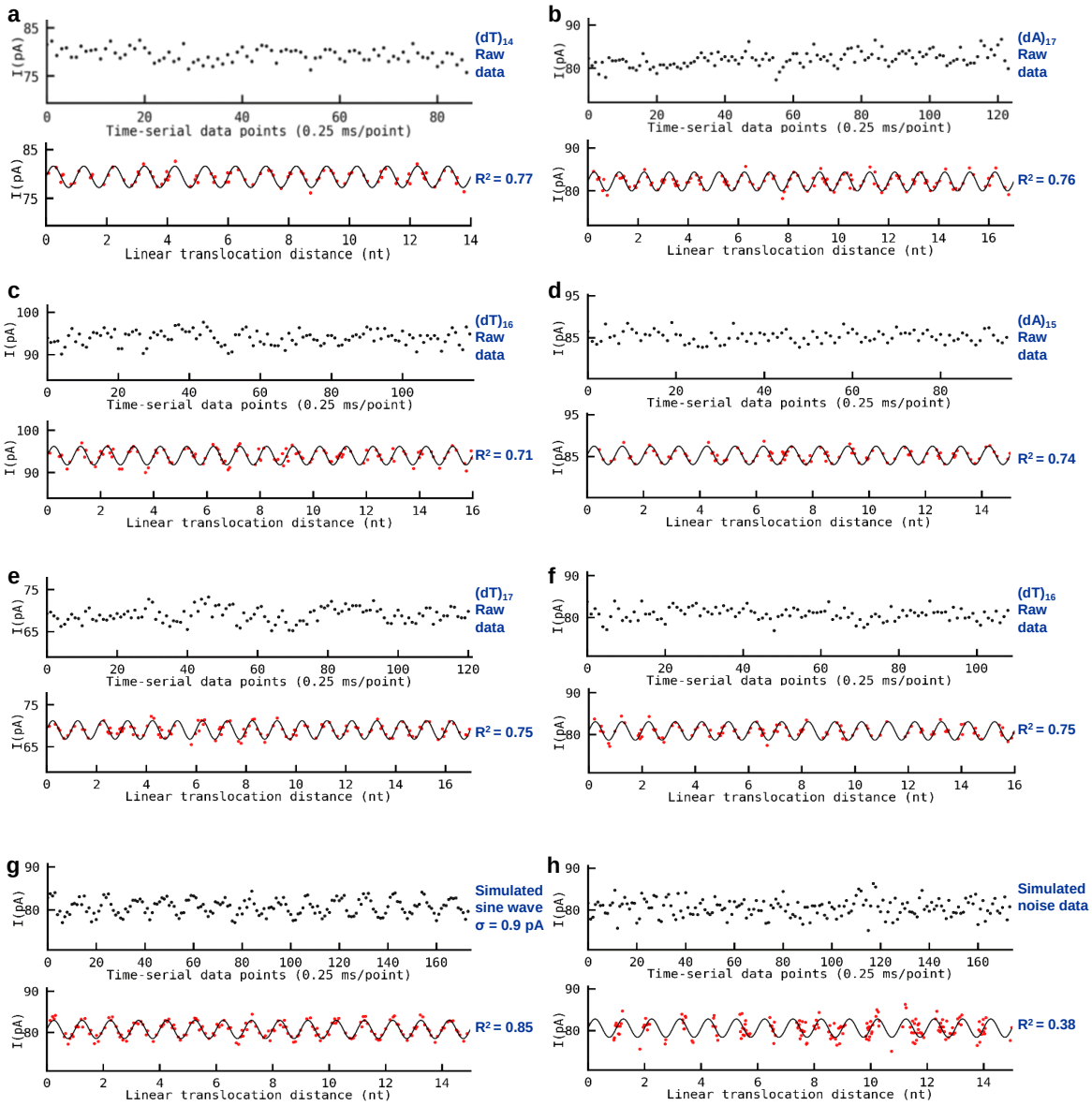


Supplementary Figure 2.7 Sinusoid profiles of nucleotide-to-nucleotide translocation of DNA homopolymers through mutant CsgG nanopore.
The homopolymers were modeled as circular cylinders using the dimensions from Supplementary Tables 2 and 3. The nucleotide-to-nucleotide translocation was modeled by moving the DNA along the nanopore vertical axis in twenty equal steps.



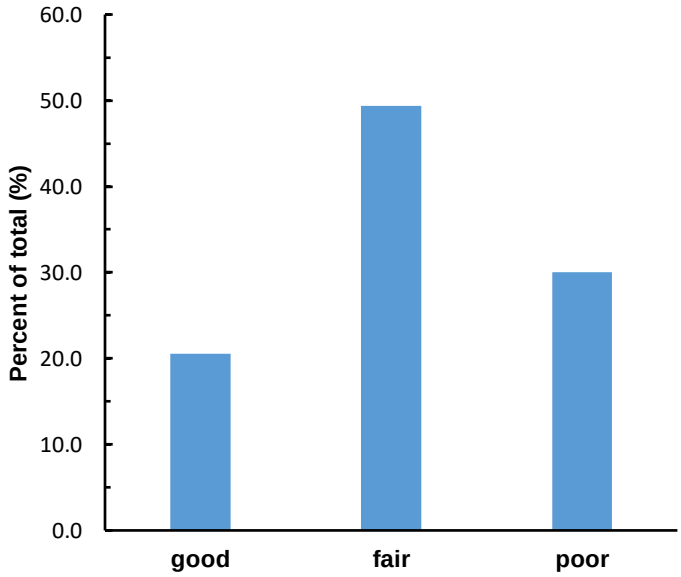
Supplementary Figure 2.8 Sinusoid profiles of nucleotide-to-nucleotide DNA translocation through a mutant CsgG nanopore and distribution of RMS amplitudes of the transition events.

The translocation from one hexamer (H1) to the next hexamer (H2) transits through a pentamer (P1). The current levels of the hexamers and pentamers, which were modeled as circular cylinders, were determined using FEA and the PNP equation system by positioning the circular cylinder models of the DNA at eleven evenly spaced positons along the vertical axis of the mutant CsgG nanopore. **(a)** Current profiles of non-homopolymer transition events. For examples are shown. The profiles of about 80% of the 4096 hexamers are similar to the top two profiles (red and blue). The remaining 20% of the hexamers have profiles similar to the bottom two profiles (yellow and green). **(b)** Distribution of RMS sine wave amplitudes. The hexamer to hexamer transition events follow a sine wave pattern and the RMS of the sine wave amplitude of each transition events was computed by dividing the peak-to-peak amplitude by $2\sqrt{2}$. The peak-to-peak amplitude was calculated as the difference between the current level of the pentamer and the mean of the current levels of the two corresponding hexamers.

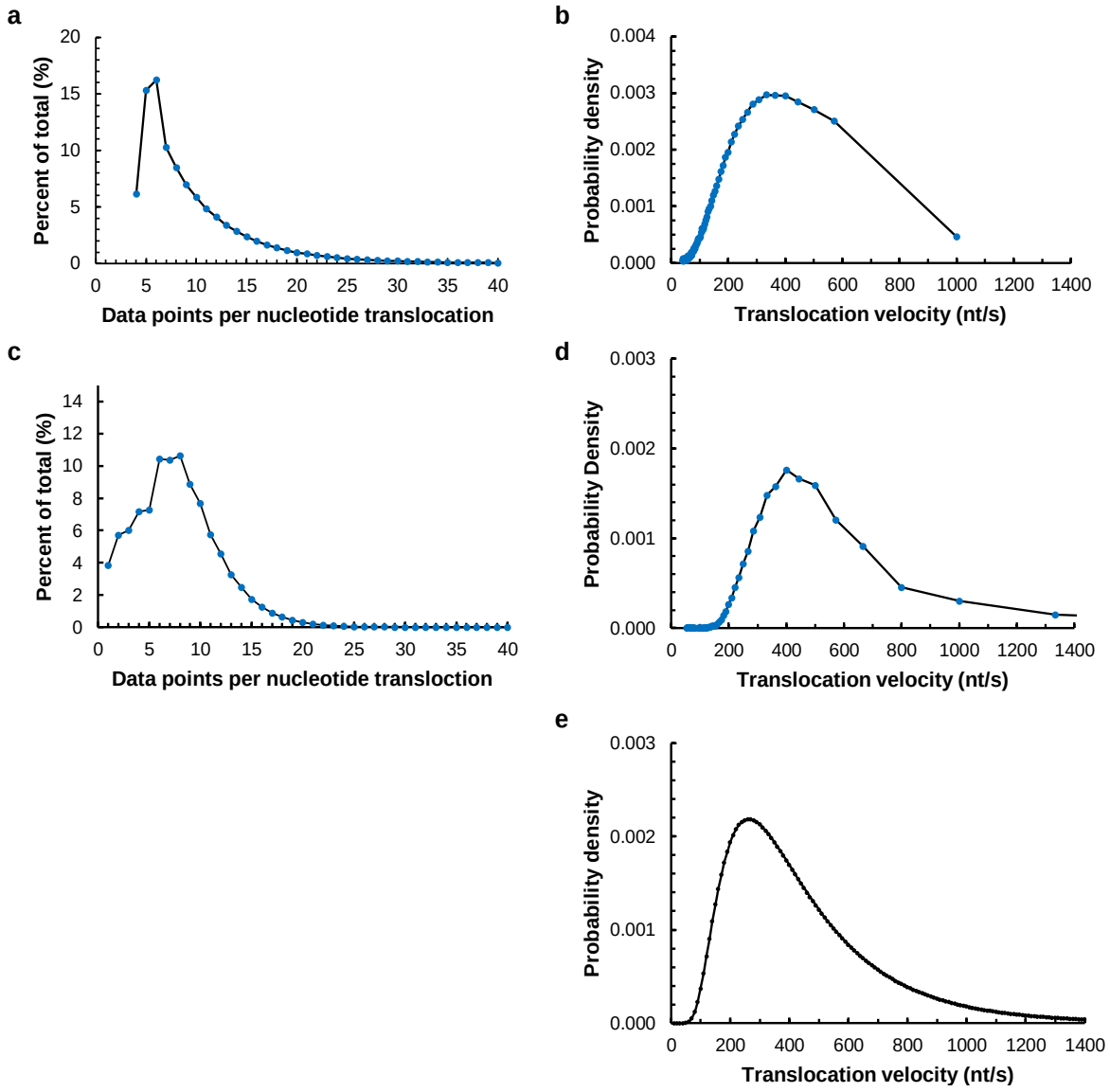


Supplementary Figure 2.9 Mapping of current signals to sine wave functions and decoding of homopolymers.

The raw experimental time-serial data and simulated data have evenly spaced data points. The temporal data points are unwarped and mapped to a sine wave function with equally spaced data point in the spatial domain. **(a)-(f)**: Six examples in which there are at least one data point mapped to 5 or more of the 6 equally divided segments of each sine wave period. **(g)** Mapping of simulated sine wave data. The data were simulated using a RMS amplitude of 2.2 pA and a standard deviation of 0.9 pA for signal variation. **(h)** Mapping of simulated random noises. The data were simulated using random signal variations with a standard deviation of 2.0 pA.

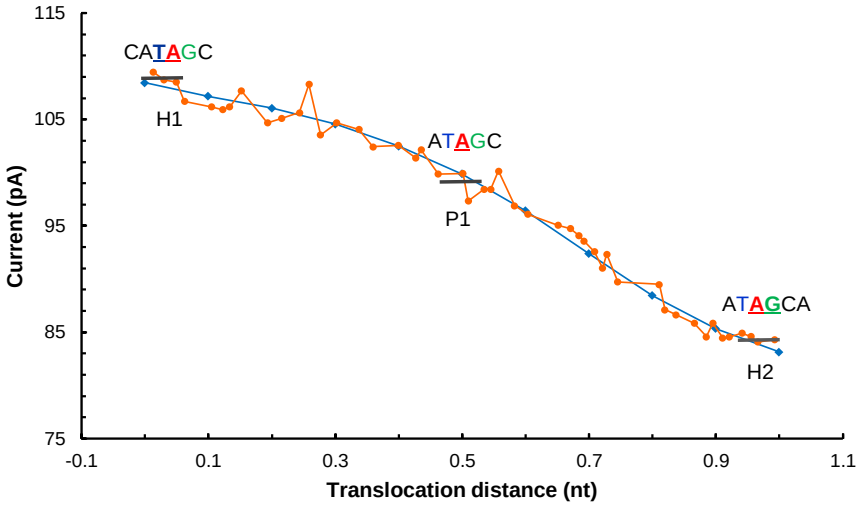


Supplementary Figure 2.10 Statistics of mapping current signals to sine wave functions. The results are based on the mapping of 25622 experimental signal profiles of 20-40-base long homopolymers that has been sequenced using the MinION R9.4 Chemistry. The quality of the mapping is classified into three categories based on the data points mapped to 6 equally spaced sub-segments of each sine wave period: good (≥ 1 data point in at least 5 sub-segments), fair (≥ 1 data point in 3 or 4 sub-segments), and poor (2 or fewer sub-segments have data points).

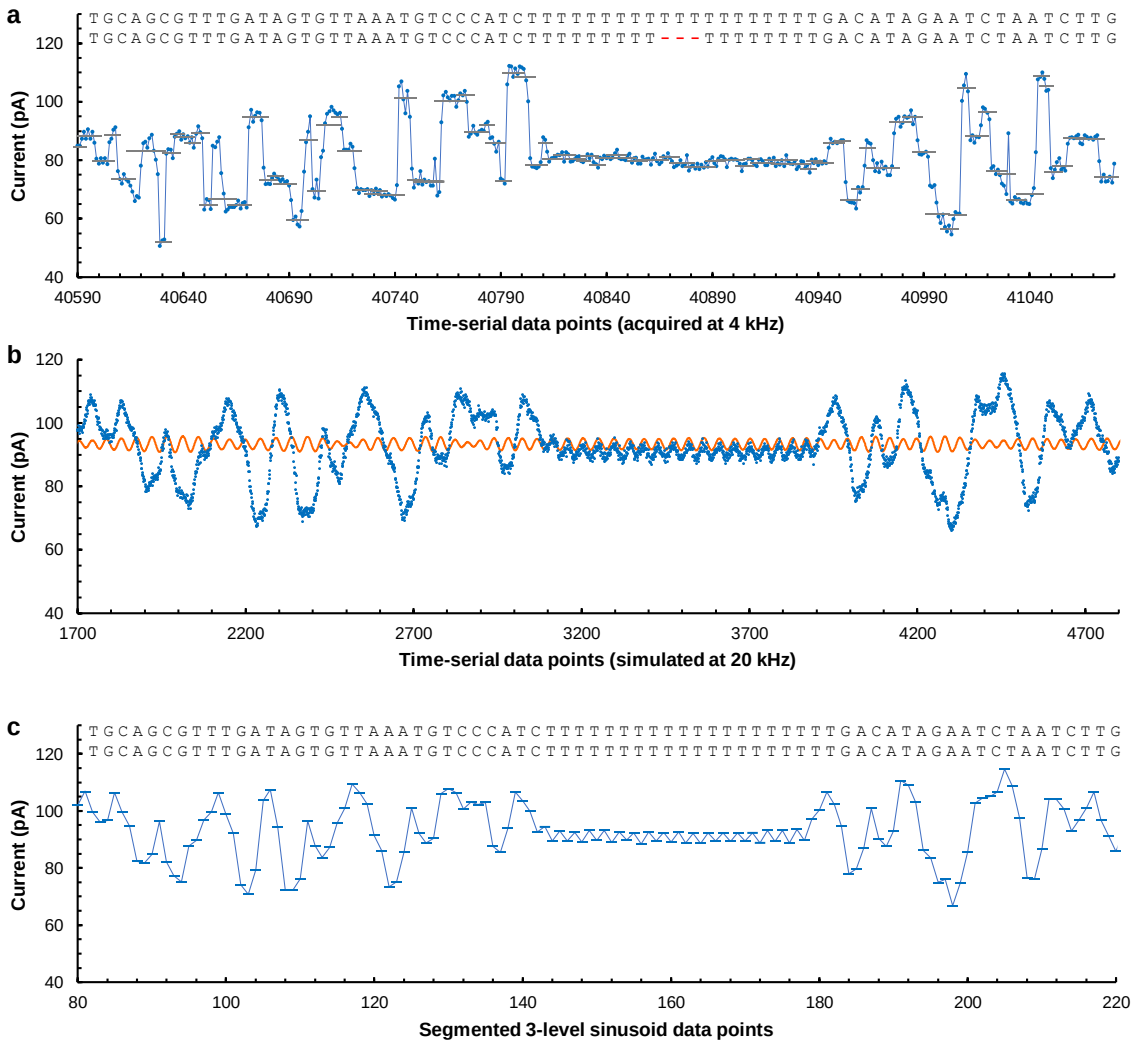


Supplementary Figure 2.11 Distributions of measured data points and velocity of nucleotide translocation.

(a) Distribution of data points per nucleotide translocation. The distribution was determined using experimental sequencing data covering 1 million bases in human genome(33). A few obvious outliner data points were removed. **(b)** Distribution of translocation velocity derived from (**a**). The mean (μ) and standard deviation (σ_v) of the velocity are determined to be about 450 nt/s and 239 nt/s, respectively. **(c)** Distribution of data points per nucleotide translocation of homopolymers. The distribution was determined using sequencing data of 25622 homopolymers (20-40 bases long)(33). **(d)** Distribution of translocation velocity derived from (**c**). The mean and standard deviation of the velocity are determined to be about 415 nt/s and 236 nt/s, respectively. **(e)** The inverse Gaussian distribution function used to simulate the translocation velocity variation. The mean and standard variation of translocation velocity are 450 nt/s and 239 nt/s, respectively.

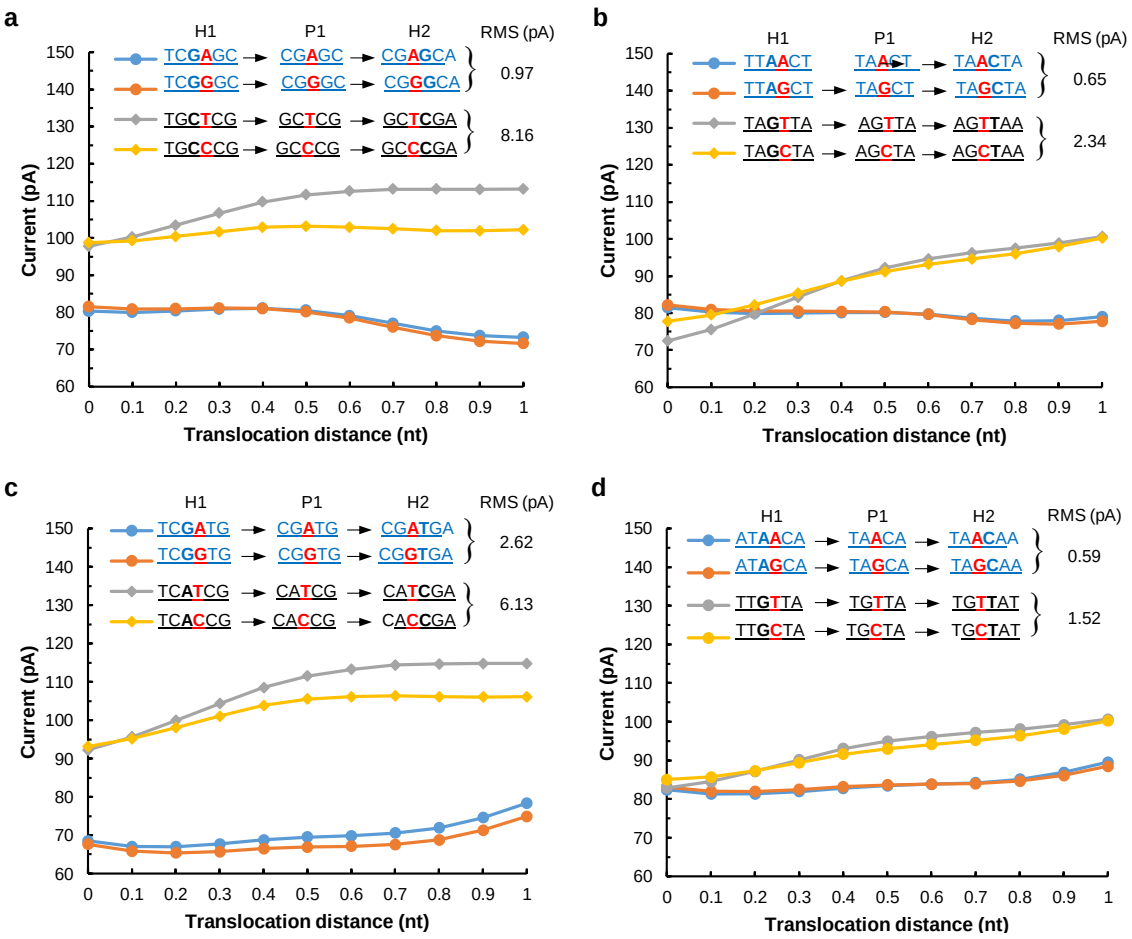


Supplementary Figure 2.12 Simulated current profile of a nucleotide-to-nucleotide translocation event.
The translocation from CATAGC to ATAGCA transits through a pentamer ATAGC (Fig. 2.1a). The distribution of data points in the spatial domain was simulated at 20 kHz using the computed current levels of the DNA models positioned at eleven equally spaced locations along the vertical axis of a mutant CsgG nanopore (Fig. 1a). The variation of translocation velocity was simulated using an inverse Gaussian distribution function with a mean of 450 nt/s and a standard variation of 239 nt/s. Signal variation due to lateral movements and measurement noise was simulated using a Gaussian distribution with a standard deviation of 0.9 pA. Model current levels are shown in blue. The simulated data points (about 44 data points per nucleotide translocation) are shown in orange. The three short horizontal bars indicate the current levels corresponding to the two hexamers and the pentamer.



Supplementary Figure 2.13 Simulated nanopore sequencing and sequence decoding using a 3-state HMM.

(a) Experimental nanopore sequencing data. The raw data (blue dots) obtained from Jain et al(33) were segmented using Metrichor (horizontal bars) and decoded using Guppy V2.1.3. The decoded sequence (bottom) is aligned to the reference sequence (top). The decoded sequence has one deletion error. **(b)** Simulated current profile. The same reference sequence corresponding to experimental data in (**a**) was used for the simulation at a bandwidth of 20 kHz. The nucleotide-to-nucleotide translocation was assumed to follow a sinusoid pattern in the spatial domain but follow an inverse Gaussian function in the temporal domain with a mean of 450 nt/s and standard deviation of 239 nt/s. Signal variations due to lateral thermal motions were simulated using a standard deviation of 0.9 pA. A bandpass filter (400-500 kHz) was applied and the filtered data (offset by the DC component of the signal) is shown (orange sine wave). **(c)** Sequence decoding using a 3-state HMM. The simulated current profile was segmented into current levels by mapping to sine wave segments of the filtered data (orange). The segmented data represented as 3-state HMM was decoded using a Viterbi algorithm. The decoded sequence (bottom) is aligned to the reference sequence (top). In this case, the sequence is decoded with 100% accuracy.



Supplementary Figure 2.14 Elimination of ambiguous basecalls by sequencing the complementary strands.

For 63 bases out of the 140 million bp sequenced at 20x depth, the consensus basecalls of the two complementary strands do not agree with one another. In all cases, the sequence of one of the complementary strands is essentially indistinguishable from another sequence, leading to the calling of one of the bases as two possible bases with almost equal probability. In these cases, to resolve the ambiguity, the RMS of the Euclidean distance between the current profiles of the two putative sequences for both complementary strands are calculated. The strand with a larger distance is assigned with a higher quality score. Shown are four examples, in which the top strand is assigned a lower quality score due to the presence of another sequence with almost identical current profile. **(a)** In this case, the consensus call of the template strand is 5'-TCG**G**GCA-3' while the consensus call of the reverse complementary is 5'-TGC**T**CGA-3'. Thus, the consensus basecalls of the two strands do not agree with one another. The RMS of the Euclidean distances between the profiles of 5'-TCG**A**GCA-3' and 5'-TCG**G**GCA-3', and between the profiles of 5'-TGC**C**CGA-3'and 5'-TGC**T**CGA-3') were calculated to be 0.97 pA and 8.2 pA, respectively. Therefore, the consensus basecall of the reverse complementary strand has a higher quality score and is used for basecalling. **(b)-(d):** another three examples. In all 63 cases, the bases are called correctly (identical to the corresponding bases in the reference sequences).

2.8 Acknowledgements

This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health (5R01GM126013 to X.H.).

This work, in part, is currently being prepared for submission for publication of the material by Wenxu Zhang, and Xiaohua Huang. The dissertation author was the primary investigator and author of this material.

Chapter 3 Nanopore Single-molecule Protein Identification

3.1 Abstract

A physical method for identification and counting of single protein molecules has remained unattainable. We investigated the feasibility of using the patterns of lysine residues in single protein molecules that are acquired with a nanopore as fingerprints for identification. We developed an algorithm for mapping the nanopore signal profiles to the reference lysine profiles in the human proteome database and found that both full-length proteins and protein fragments can be identified with high accuracy.

3.2 Introduction

Proteins are the workhorses of living systems. Complete cataloging and accurate quantification of this physiologically important class of molecules are essential for proteome-scale elucidation of their functions, and the development of precise protein-based diagnosis and treatment of human diseases. With recent advances in protein mass spectrometry, whole proteome analysis is getting closer to becoming a reality(49). For example, the SWATH-MS method allows for label-free identification and quantification of thousands of proteins with large dynamic ranges(49, 50). Microfluidics-based immunoassays can also be used for sensitive analysis and digital counting of proteins(51, 52). However, these methods have limitations. Mass spectrometry methods require a relatively large amount of sample and expensive instruments for analysis, while immunoassays are indirect methods that rely on antibodies with high affinity and specificity to the protein analytes.

3.3 Lysine patterns as protein fingerprint

Unlike nucleic acids, no molecular machinery exists in nature that can amplify protein molecules. Therefore, exhaustive characterization of whole proteomes requires physical methods

capable of direct identification and counting of single protein molecules. Several recent reports have described attempts at single-molecule protein identification using fluorescence microscopy in combination with Edman degradation(53), molecular motors for threading protein molecules(30), or solid-state nanopore(54), and by electron tunneling(55). Here, we describe a nanopore-based method for identification of single protein molecules (Fig. 3.1). Most human proteins contain many lysine residues distributed along the long primary sequences (Supplementary Fig. 3.1a, b). In our method, a protein analyte is identified by measuring its lysine profile using a nanopore and mapping it to a reference proteome database. (Fig. 3.1a, b).

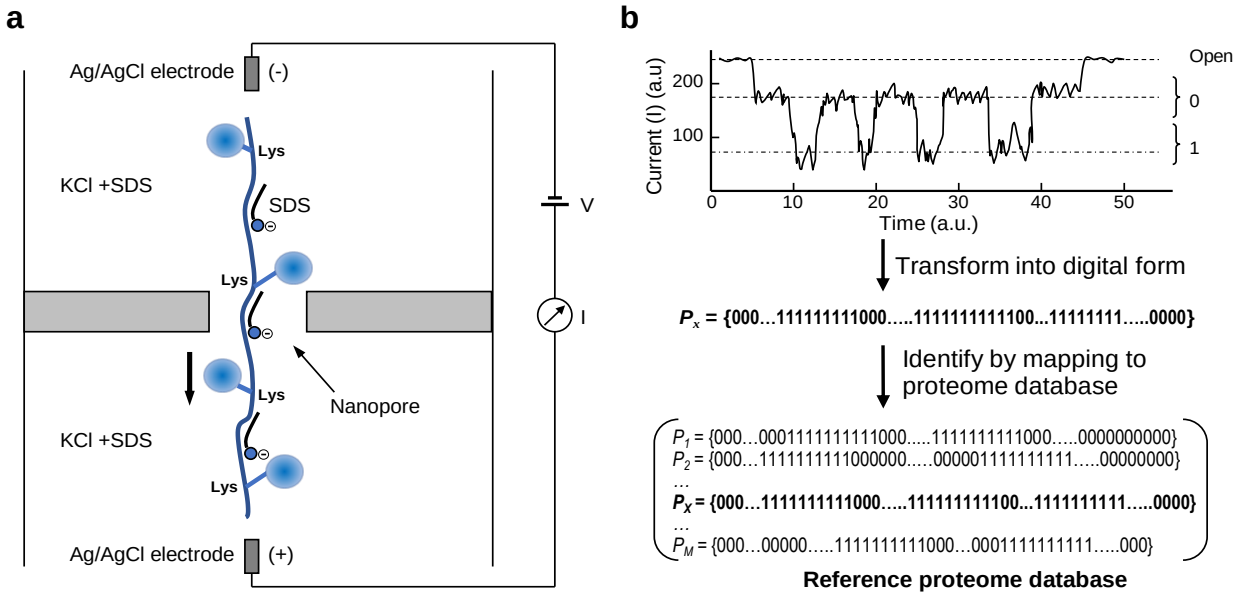


Figure 3.1 Nanopore-based single-molecule protein identification.
(**a)** Nanopore measurements of lysine profiles. A fully denatured protein with chemical labels on lysine residues is electrophoretically translocated through a nanopore and the ionic current blockage profile is acquired. Not to scale. (**b)** Strategy for computational identification. The measured current profile is transformed into a digital form and mapped to the references in the proteome database to identify the protein analyte. a.u: arbitrary unit.

We used computational simulations to investigate the feasibility and performance of our method. Several assumptions were made about potential experimental implementation. First, a reference proteome database is available. Due to technological advances in DNA sequencing and

protein mass spectrometry, proteomes of thousands of organisms have become avilable(56). Second, a method is available for persistent and unidirectional translocation of denatured protein molecules through a nanopore. Experimentally, this can be realized by using an anionic detergent such as sodium dodecyl sulfate (SDS) to simultaneously denature the protein molecules and impart a uniform charge density along the linear polypeptides (Fig. 3.1a). Commonly employed in SDS-PAGE methods for electrophoretic separation of proteins, SDS molecules bind to linear polypeptides at a ratio of about one SDS molecule per two amino acid residues, mostly through hydrophobic interactions independent of ionic strength and the composition of the proteins(57, 58). Third, the current amplifier has the bandwidth required for length measurements with sufficient accuracy. Fourth, the lysine residues are labeled with high efficiency and the chemical label is sufficiently large such that the level of current blockage by the labeled lysine is greater than those by unlabeled amino acid residues. Due to the reactive amine group on the side chain, lysine residues can be labeled with many chemical moieties almost quantitatively using coupling agents commonly employed in protein conjugation chemistry. Since the current blockage level of the labeled lysine residues is easily distinguishable from other unlabeled amino acids, the analog current blockage profiles can be transformed into a digital binary format (Fig. 3.1b).

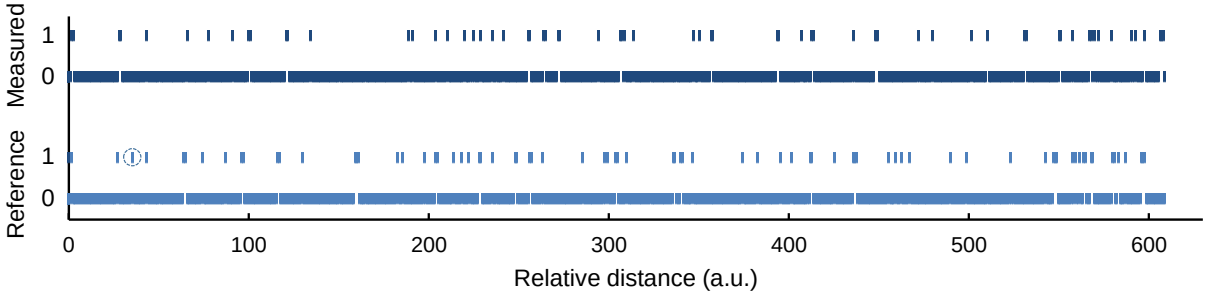
We used the UniProt human proteome database (Release 2018_11) as the reference(59). The theoretical lysine profiles of all proteins in the database were generated with the assumption of perfect labeling efficiency and distance measurements. To account for the possibility of a polypeptide translocating through the nanopore from either N-terminus to C-terminus or *vice versa*, two reference profiles were generated for each protein. Experimentally, the chemical labeling efficiency is unlikely to be unity, and the translocation process is stochastic with potentially large fluctuations in translocation velocity, and thus large uncertainties in distance

measurements. Therefore, in our computational experiments, we simulated current blockage profiles under nine different conditions. The simulated profiles with data streams of binary bits 0 and 1 were reformatted into numerical lysine distance profiles (Fig. 3.2a).

3.4 Mapping between lysine patterns and proteins

To investigate whether lysine patterns are sufficiently unique for accurate protein identification, we developed an algorithm for mapping each measured signal profile to the theoretical profile of every reference sequence in the database (see Methods for detail). To minimize the computations required, the algorithm was implemented using dynamic programming and a set of *a priori* criteria to direct the process. Briefly, the quality score of the alignment was calculated analytically based on the labeling probability and distance information. The alignments were performed roughly along the diagonal portion of the matrix to find the path with the highest mapping quality score (Fig. 3.2b). For each signal profile, the algorithm output a short list of reference sequences ranked by mapping quality scores (Fig. 3.2c). The protein analyte was identified as the sequence with the highest score.

a



b

Mapping Probability		Reference lysine distance profile											
		$(D_{1,1},D_{1,2})$	$(D_{2,1},D_{2,2})$	$(D_{3,1},D_{3,2})$	$(D_{4,1},D_{4,2})$	$(D_{5,1},D_{5,2})$	$(D_{6,1},D_{6,2})$	$(D_{7,1},D_{7,2})$	$(D_{8,1},D_{8,2})$	$(D_{9,1},D_{9,2})$	$(D_{10,1},D_{10,2})$	$(D_{11,1},D_{11,2})$	$(D_{12,1},d_{12,2})$
Measured lysine distance profile	$(d_{1,1},d_{1,2})$	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$	$P_{1,4}$	$P_{1,5}$							
	$(d_{2,1},d_{2,2})$		$P_{2,2}$	$P_{2,3}$	$P_{2,4}$	$P_{2,5}$	$P_{2,6}$						
	$(d_{3,1},d_{3,2})$			$P_{3,3}$	$P_{3,4}$	$P_{3,5}$	$P_{3,6}$	$P_{3,7}$	$P_{3,8}$				
	$(d_{4,1},d_{4,2})$				$P_{4,4}$	$P_{4,5}$	$P_{4,6}$	$P_{4,7}$	$P_{4,8}$	$P_{4,9}$			
	$(d_{5,1},d_{5,2})$					$P_{5,5}$	$P_{5,6}$	$P_{5,7}$	$P_{5,8}$	$P_{5,9}$	$P_{5,10}$		
	$(d_{6,1},d_{6,2})$						$P_{6,6}$	$P_{6,7}$	$P_{6,8}$	$P_{6,9}$	$P_{6,10}$	$P_{6,11}$	$P_{6,12}$
	$(d_{7,1},d_{7,2})$							$P_{7,7}$	$P_{7,8}$	$P_{7,9}$	$P_{7,10}$	$P_{7,11}$	$P_{7,12}$
	$(d_{8,1},d_{9,2})$								$P_{8,8}$	$P_{8,9}$	$P_{8,10}$	$P_{8,11}$	$P_{8,12}$
	$(d_{9,1},d_{9,2})$									$P_{9,9}$	$P_{9,10}$	$P_{9,11}$	$P_{9,12}$
	$(d_{10,1},d_{10,2})$										$P_{10,10}$	$P_{10,11}$	$P_{10,12}$

c

Profile	Alignment												Score
Measured	$(d_{1,1},d_{1,2})$	$(d_{2,1},d_{2,2})$	-	$(d_{3,1},d_{3,2})$	$(d_{4,1},d_{4,2})$	$(d_{5,1},d_{5,2})$	-	$(d_{6,1},d_{6,2})$	$(d_{7,1},d_{7,2})$	$(d_{8,1},d_{8,2})$	$(d_{9,1},d_{9,2})$	$(d_{10,1},d_{10,2})$	$P_{10,12}$
Reference	$(D_{1,1},D_{1,2})$	$(D_{2,1},D_{2,2})$	$(D_{3,1},D_{3,2})$	$(D_{4,1},D_{4,2})$	$(D_{5,1},D_{5,2})$	$(D_{6,1},D_{6,2})$	$(D_{7,1},d_{7,2})$	$(D_{8,1},D_{8,2})$	$(D_{9,1},D_{9,2})$	$(D_{10,1},D_{10,2})$	$(D_{11,1},D_{11,2})$	$(D_{12,1},D_{12,2})$	

Figure 3.2 Protein identification by mapping to a proteome database.
(**a**) Measured and reference profiles. As an example, the simulated and reference profiles of human serum albumin are shown. The measured profile is simulated with 85% lysine labeling efficiency and 20% CV in distance measurements. One of the unlabeled lysine residues is indicated by a circle. (**b**) Illustration of mapping process. Dynamic programming is used to implement an algorithm to compute the best quality score of mapping the signal profile to a theoretical reference profile from the proteome database. (**c**) Alignment and identification. The protein analyte is identified as the sequence in the reference database with the highest mapping quality score.

We first assessed the upper bounds of identification accuracy by mapping the signal profiles of all human proteins with perfect labeling efficiency and length measurements. Essentially all proteins that contain at least one lysine residue, 99.4% of the 20,416 human proteins in the database, were uniquely identified (Fig. 3.3a, and Supplementary Table 3.1). The

remaining 0.6% proteins failed to be identified due to the absence of lysine residues in their sequences (Supplementary Fig. 3.1a).

Next, we investigated identification accuracy using the profiles simulated with nine different combinations of three chemical labeling efficiencies (85%, 95% and 100%) and three coefficients of variation (CV) in distance measurements (0%, 20% and 50%). Two thousand proteins were randomly selected from the human proteome for the computational experiments. We found that most proteins can be identified with high accuracy (Fig. 3.3a and Supplementary Table 3.1). The identification accuracy reaches almost 99% if the labeling efficiency is 95% and CV of lysine distance measurements is 20%. Even with a labeling efficiency of only 85% and a large coefficient of variation of 50% in distance measurements, about 95% of the proteins are uniquely identified (Fig. 3.3a). When the labeling efficiency is increased from 85% to 95%, the identification accuracy improves appreciably (up to ~3% increase) while further increase in labeling efficiency from 95% to 100% results in almost negligible improvement in identification accuracy (<0.3% increase). The variation in distance measurements has similar influence on accuracy. At 85% labeling efficiency, identification accuracy improves appreciably (up to 3%) when the CV of distance measurements decreases from 50% to 20%. A very small percentage of proteins cannot be identified correctly primarily because either they have lysine patterns similar to other proteins in the database (Supplementary Table 3.1) or they contain fewer than 5 lysine residues (Supplementary Table 3.2). Their identities can be narrowed down to several candidates by ranking their mapping scores (Supplementary Table 3.1).

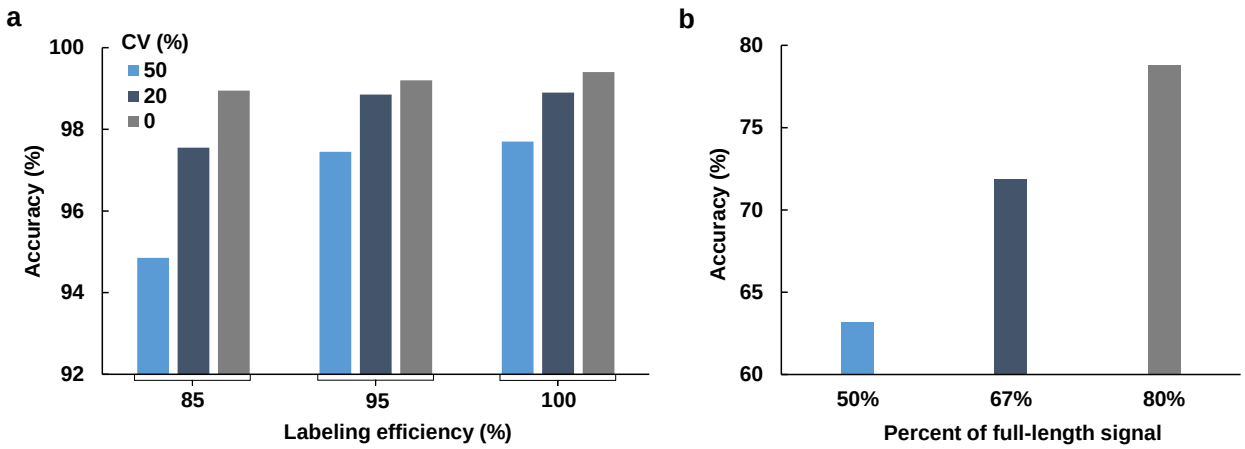


Figure 3.3 Identification of full-length proteins and fragmented proteins using lysine profiles.
(a) Accuracies of full-length protein identification. The simulations were performed under nine different combinations of the indicated labeling efficiencies and coefficients of variation in distance measurements. Two thousand proteins randomly selected from the human proteome were used for the simulations. **(b)** Accuracies of protein fragment identification. For each group, one thousand fragments were randomly selected from full-length signal profiles simulated with 85% labeling efficiency and 20% CV in distance measurements.

Due to degradation, fragments of proteins are commonly present in biological samples. Certain proteins such as histone proteins also contain chemically modified lysine residues(60). Thus, we also investigated whether protein fragments can be identified. We randomly selected portions of simulated full-length signal profiles and used our algorithm to determine the identities of the protein fragments. Depending on their lengths, about 60 to 80% of the fragments were identified correctly (Fig. 3.3b). The relatively lower accuracy is primarily due to the presence of too low of a number of labeled lysine residues in the fragments for unique identification (Supplementary Fig. 3.2a-c). However, up to 98% of the protein fragments can be identified if the fragments contain 7 or more labeled lysine measurements (Supplementary Fig. 3.2d).

3.5 Methods

3.5.1 Overview of approach and general methods

In this work, we aim to demonstrate the feasibility of a nanopore-based method for accurate protein identification using computational simulations. The pattern of lysine residues in the primary sequence of a given protein analyte is acquired by measuring the ionic current blockage profile associated with the translocation of the fully denatured and labeled protein through a nanopore (Figure 3.1a). The lysine residues are labeled with a relatively large chemical moiety so that the level of current blockage of the labeled lysine is distinguishable from those of unlabeled amino acid residues. The protein analyte is then identified by matching the measured lysine pattern to a reference proteome database using a dynamic alignment algorithm (Figure 3.1b).

One major challenge in nanopore detection or sequencing is the measurement of small ionic current levels (10's pA to nA) accompanying the translocation of biomolecules through a nanopore, which is usually rapid and stochastic due to the uneven charge distribution and thermal motions of the molecules. Unlike DNA or RNA, protein molecules do not carry periodic charges along the linear polypeptides. The persistent and unidirectional electrophoretic translocation of a fully denatured protein with some degree of uniformity in velocity requires a means to impart a uniform density of charge along the linear polypeptides. Experimentally, this potentially can be realized by using SDS to simultaneously denature the proteins and render a uniform charge density onto the linear polypeptide, with about 1 charge (sulfate on the SDS) per 2 amino acid residues (Figure 3.1a).

In our simulations, we assume that the translocation of each amino acid residue on the linear polypeptide is independent with a mean velocity, and an average of 10 data points ($\mu = 10$)

are acquired per amino acid at a fixed frequency. Therefore, the relative distance (d) along the linear polypeptide is linearly proportional to the number of data points in the current measurements ($d = \mu * (\text{number of amino acid residues})$). Due to the stochastic fluctuations in translocation velocity, the measured relative distance has a variation, which is assumed to be random and follows a Gaussian distribution with a mean (μ) and a specified standard deviation (σ). Since the level of ionic current blockage of labeled lysine residues is very different from those of the unlabeled amino acid residues, the measured ionic current data can be transformed into digital binary format (Figure 3.1b). Each signal data point associated with a non-labeled amino acid is represented by the binary bit 0 while that for a labeled lysine residue is represented by the binary bit 1. Therefore, each signal profile consists of a stream of binary bits 0 and 1.

Even though chemical labeling of lysine residues on proteins are usually very efficient and rapid using standard protein conjugation chemistry such as nucleophilic substitution reaction between the primary amine group on lysine and N-hydroxysuccinimide-activated chemical compounds, the labeling efficiency is very unlikely to be 100%. In addition, some lysine residues on certain proteins such as histones are already chemically modified. Experimentally, 95% labeling efficiency is very likely achievable. Nevertheless, we also evaluated our method using a relatively low labeling efficiency of 85%.

The UniProt human proteome database release 2018_11 was used as the reference. The simulations and dynamic alignment algorithm were performed using a desktop computer and custom programs written in Python (version 3.6). Since the denatured labeled linear polypeptides can be translocated through the nanopore in both directions, either from N-terminus to C-terminus or *vice versa* (C to N and N to C), two profiles were generated for the proteins or protein fragments. The theoretical reference data set contains both profiles (C to N and N to C)

of each of the 20,416 curated protein sequences in the human proteome database. A random number generator was used to select the proteins from the database for simulations. The direction of translocation was assumed to be random.

3.5.2 Simulations of ionic current profiles and lysine patterns

The theoretical reference data set was constructed using binary current signal levels by representing each lysine residue with 10 consecutive binary bit 1's and each of the other amino acid residues with 10 consecutive 0's. For the computational experiments, the probability of any lysine residue being chemically labeled was modeled using the specified labeling efficiency. The variation in distance measurements was modeled using a Gaussian distribution with the mean of 10 ($\mu = 10$) and the specified relative standard variation (RSD) or coefficient of variation (CV). Two thousand proteins were randomly selected from the human genome database for the simulations. The nanopore current profiles were simulated using nine different combinations of three labeling efficiencies (85%, 95% and 100%) and three coefficients of variation in distance measurements ($CV = \sigma/\mu = 0\%$, 20 % or 50%, corresponding to $\sigma = 0$, 2 or 5 for $\mu = 10$).

Protein identification based on partial sequences or fragments was simulated by randomly choosing a signal window covering 50%, 67% or 80% of the full-length profiles. For each of the three partial length percentages, one thousand fragments were randomly selected from the data set simulated with 85% labeling efficiency and 20% CV in distance measurements.

3.5.3 Dynamic alignment algorithm for mapping and protein identification

We developed an alignment algorithm to map the experimental signal profiles to the theoretical reference profiles in the proteome database to identify the protein analytes (Figure 3.3b, c). Given the experimental binary current profile *s* of the protein analyte that contains the information on the number of measured labeled lysine residues and the separation distances

between them due to non-lysine residues, we want to find the profile of reference sequence l in the proteome database that maps to the signal s with the highest probability score.

3.5.3.1 Data formatting

First, the binary data streams of the profiles were reformatted into distance measurements, which are defined as the numerical counts of the binary bits in the signal data stream. Each experimental binary signal profile is reformatted into a distance profile *s*:

$$s = \{(d_{1,1}, d_{1,2}), (d_{2,1}, d_{2,2}), \dots, (d_{i,1}, d_{i,2}), \dots, (d_{M,1}, d_{M,2})\},$$

where $d_{i,1}$ is the measured distance of non-labeled amino acid residues preceding the particular lysine residue i^{th} in the measured signal profile, $d_{i,2}$ is the measured distance of the i^{th} labeled lysine, and M corresponds to the last segment of the signal. If the last segment does not end with a labeled lysine, it is considered to have a null labeled lysine with a length of 0, for example, $(d_{M,1}, 0)$. Similarly, the binary data streams of the theoretical reference profiles were also reformatted into distance profile *l*:

$$l = \{(D_{1,1}, D_{1,2}), (D_{2,1}, D_{2,2}), \dots, (D_{j,1}, D_{j,2}), \dots, (D_{N,1}, D_{N,2})\},$$

where $D_{j,1}$ is the distance of non-lysine amino acid residues preceding the particular lysine residue j^{th} on the linear polypeptide, $D_{j,2}$ is the distance of the j^{th} lysine, and N corresponds to the last segment of the sequence. If the last segment does not end with a lysine, it is considered to have a null lysine with a length of 0, for example, $(D_{N,1}, 0)$.

In special cases where there are multiple consecutive lysine residues in the reference sequence, these lysine residues are considered to be separated by a distance of 0. For example, in our simulations, the distance of the j^{th} lysine in sub sequence $K_{j-1}K_j$ is represented by $(D_{j,1}, D_{j,2}) = (0,10)$, since 10 is the mean number of bits per amino acid. By doing so, we ensure that $N \geq M$ and our algorithm will not collapse.

3.5.3.2 Algorithm

Each measured signal profile s was aligned to every theoretical reference profile l in the proteome database (L). Let $p(l, s;\ \varphi)$ be the joint probability of reference profile l and measured signal profile s, which is parameterized by a mapping function φ. The best mapped reference profile can then be represented as:

$$l_0 = \underset{l \in L}{\operatorname{argmax}} \max_{\varphi}\ p(l, s; \varphi) = \underset{l \in L}{\operatorname{argmax}} \max_{\varphi} \left(\prod_{i=1}^{M} \Theta_i \prod_{j=1}^{N} E_j \right),$$

where Θ_i and E_j are defined as follows.

Θ_i is the alignment quality score that is calculated by mapping the signal $(d_{i,1}, d_{i,2})$ to sub lengths $\{(D_{g,1}, D_{g,2}), (D_{g+1,1}, D_{g+1,2}), \ldots, (D_{g+k,1}, D_{g+k,2})\}$ of the theoretical reference profile:

$$\Theta_i = \Theta_i^{g,\ g+k} = \exp\left(-\frac{\left(d_{i,1} - \sum_{h=g}^{g+k-1}(D_{h,1} + D_{h,2}) - D_{g+k,1}\right)^2}{2\left(\sum_{h=g}^{g+k-1}(D_{h,1} + D_{h,2}) + D_{g+k,1}\right)\sigma^2}\right) \exp\left(-\frac{(d_{i,2} - D_{g+k,2})^2}{2D_{g+k,2}\sigma^2}\right)$$

E_j is a weighing factor which is determined by the mapping and labeling efficiency:

$E_j = p_l^x (1 - p_l)^{1-x}$,

where p_l is the labeling efficiency, and $x = 1$ if the j^{th} lysine on the reference aligns to a labeled lysine in the measured signal profile, otherwise $x = 0$.

For mapping the signal profiles with perfect distance measurements, standard deviation (σ) is set to a very small value to prevent division with zero. To find the best map between a given measured signal profile and each reference profile, a dynamic algorithm was illustrated as follows.

Algorithm 3.1 Map a measured signal profile to a reference profile l and compute the alignment quality score

The measured signal profile (s) and the theoretical reference profile (l) to be mapped are:

$s = \{(d_{1,1}, d_{1,2}), (d_{2,1}, d_{2,2}), \dots, (d_{i,1}, d_{i,2}), \dots, (d_{M,1}, d_{M,2})\}$,
$l = \{(D_{1,1}, D_{1,2}), (D_{2,1}, D_{2,2}), \dots, (D_{j,1}, D_{j,2}), \dots, (D_{N,1}, D_{N,2})\}$.

Let map φ_m^n be the best alignment between sub signal length and sub reference sequence length: $s' = \{(d_{1,1}, d_{1,2}), \dots, (d_{m,1}, d_{m,2})\}$ and $l' = \{(D_{1,1}, D_{1,2}), \dots, (D_{n,1}, D_{n,2})\}$, and p_m^n be the quality score of this alignment. And let $\Theta_i^{g,\ g+k}$ be the alignment quality score of mapping the measured signal $(d_{i,1}, d_{i,2})$ to the sub length $\{(D_{g,1}, D_{g,2}), (D_{g+1,1}, D_{g+1,2}), \dots, (D_{g+k,1}, D_{g+k,2})\}$ of the reference profile.

Initialize mapping quality score of path φ_n^1 by calculating

$$p_1^n = \Theta_1^{1,n} = \exp\left(-\frac{\left(d_{1,1} - \sum_{h=1}^{n-1}(D_{h,1} + D_{h,2}) - D_{n,1}\right)^2}{2\left(\sum_{h=1}^{n-1}(D_{h,1} + D_{h,2}) + D_{n,1}\right)\sigma^2}\right)\exp\left(-\frac{(d_{1,2} - D_{n,2})^2}{2D_{n,2}\sigma^2}\right)$$

For $i = 2$ to M

For $j = i$ to N

$$p_i^j = \max_{k=1,\dots j-i+1}\left(p_{i-1}^{j-k}\Theta_i^{j-k+1,j}\prod_{h=1}^{k}E_{j-k+h}\right)$$

End

End

Assign quality score of the best alignment between measured signal profile and theoretical reference profile as p_M^N (= $P_{M,N}$).

3.5.3.3 Implementation using dynamic programming

The measured signal profile s was mapped to every theoretical reference profile l in the proteome database. The algorithm was implemented using dynamic programming to minimize the computations required. During the alignment, a set of *a priori* criteria was used to determine whether to proceed with the alignment process, and the direction and the number of lysine distances to compute the probability scores. The criteria are as follows: 1) The alignment is limited to only reference profiles that contain a number of lysine residues equal to or greater than, but less than 2 times the number of labeled lysine distance measurements in the signal

profile. The reference profiles must also have a total length between 50% and 150% of the measured length of the signal profile; 2) After mapping a particular labeled lysine to the sub lengths of the reference profile, paths that are more than five orders of magnitude less probable than the most probable path are terminated. This allows paths that have up to five skipped lysine residues to continue forward if their distance measurements align; 3) The alignment is terminated if the joint probability score drops below the highest score of previously aligned reference sequences by more than five orders of magnitude. This criterion only applies when at least one reference profile has already been aligned to the measured profile with a quality score.

For each measured signal profile of the protein analyte, the algorithm outputs a short list of reference profiles ranked by mapping quality scores. The protein analyte is identified as the reference profile that mapped with the highest score (Figure 3.3c). For proteins that are not identified correctly using the highest probability mapping score, their identities can also be narrowed down by examining the list of reference proteins ranked by mapping quality scores.

3.3.3.4 Identification of protein fragments

To map the signal profiles of protein fragments, some constraints used to minimize computations in the full-length alignment were removed. First, the mapping is not limited by the signal profile's total length or its number of measured labeled lysine distances. Second, the mapping of the initial distance signal $(d_{1,1}, d_{1,2})$ is not limited to only the initial several distances in the reference since it is unknown where the fragment derives from in the full-length sequence. The mapping is expanded to align to any length $(D_{j,1}, D_{j,2})$ in the reference profile. Third, similarly, the last signal length $(d_{M,1}, d_{M,2})$ is not required to align with the lengths at the end of the reference profile. The removal of these restraints allow the partial signal fragment to be mapped to any consecutive regions of the reference profile.

3.6 Conclusions

In summary, we described a nanopore-based method for single-molecule protein identification. We showed that full-length proteins can be identified with high accuracy, up to 99%, depending on labeling efficiency and variation in distance measurements. Protein fragments containing 7 or more lysine residues can also be identified with up to 98% accuracy. As compared to other reported methods which make use of multiple amino acid residues for identifications(61-64), our method is simpler to implement experimentally. A labeling efficiency of 95% is very likely achievable. Whether relative distances can be measured with sufficient consistency and accuracy using nanopores remains to be explored. However, nanopore DNA sequencing has become a reality(10). Given that our method requires the measurements of only two readily distinguishable current levels of current blockage, one by unlabeled amino acid side chains and the other by much larger labeled lysine residues, instead of the thousands of levels for nanopore DNA sequencing(10), our method potentially can be implemented using a similar platform for accurate proteome-scale identification and counting of single protein molecules. We envision that our method can also be implemented using a nanopore device integrated into a microfluidic platform capable of low-loss biomolecule processing to enable digital analysis of both RNA and protein molecules in single cells(65, 66).

3.7 Supplementary materials

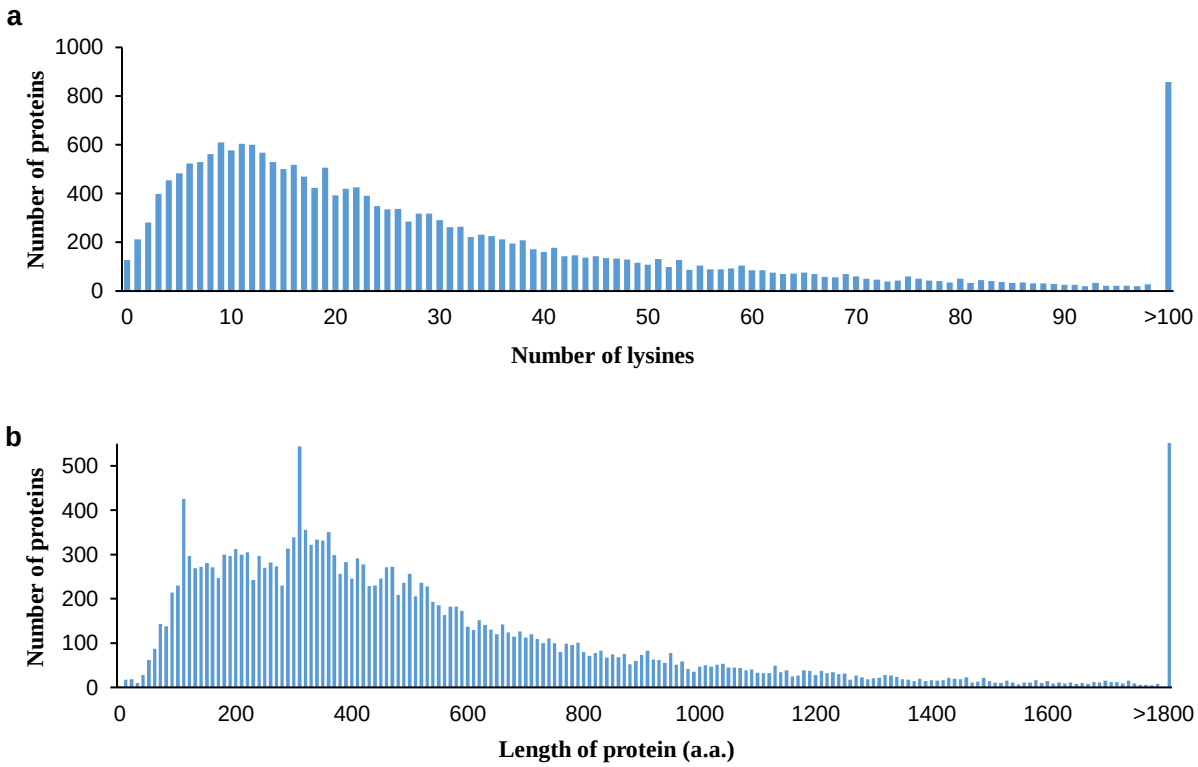
Supplementary Table 3.1. Accuracies of full-length protein identifications. The accuracies are calculated based on a protein analyte being identified uniquely as the reference sequence with the highest mapping score (1st one in the ranking by mapping quality score), or within the top 3 or top 10 sequences (1st-3rd or 1st-10th ones in the ranking). The results are based on the simulations of 2,000 proteins randomly selected from the human proteome database.

Labelling efficiency (%)	Variation in distance measurement (CV or %RSD)	Identification accuracy (%)		
		Matched to reference sequence(s) ranked by quality score		
		1st	1st – 3rd	1st - 10th
85	0	99.8	100	100
	20	97.6	99.0	99.2
	50	94.8	97.6	98.7
95	0	99.8	100.0	100
	20	98.8	99.4	99.5
	50	97.4	99.0	99.3
100	0	100	100	100
	20	98.9	99.4	99.4
	50	97.7	99.2	99.6

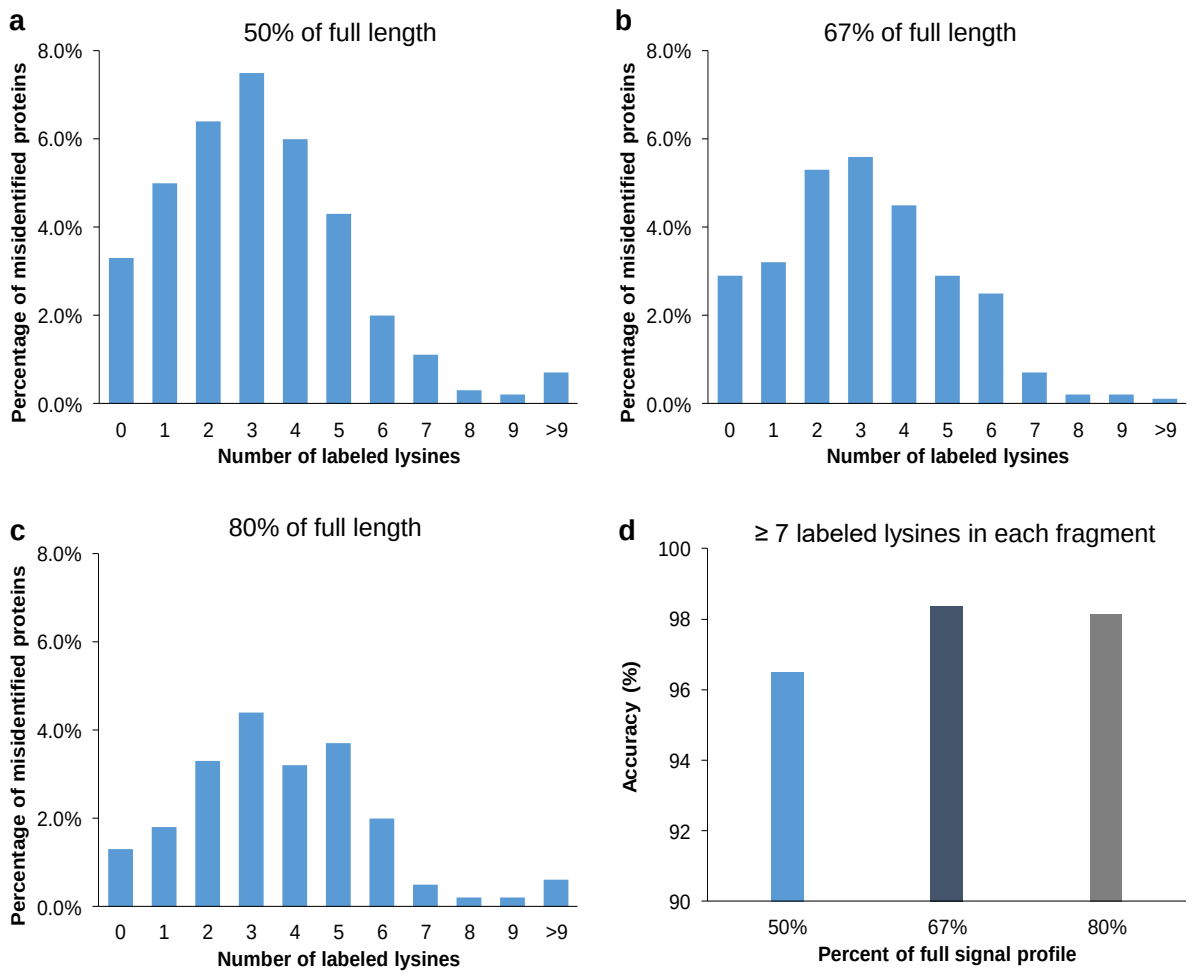
Supplementary Table 3.2. Statistics on misidentified full-length proteins. The total percent of proteins misidentified is the percent of proteins not identified correctly based on mapping quality score (Supplementary Table 3.1).

Labelling efficiency (%)	Variation in distance measurement (CV or %RSD)	Percent of proteins misidentified			
		Total (%)[a]	Percentage of total (%)[b]		
			< 5 Lys	5-10 Lys	> 10 Lys
85	0	0.2	25	0	75
	20	2.4	71	10	18
	50	5.2	62	25	13
95	0	0.2	0	67	33
	20	1.2	83	13	4
	50	2.6	72	14	14
100	0	0.0	0	0	0
	20	1.1	73	4	23
	50	2.3	76	11	13

[a]Percentage of all proteins simulated. [b]Percentage of the misidentified proteins



Supplementary Figure 3.1 Distribution of lysine residues and lengths of human proteins. There are 20,416 sequences in the UniProt human proteome database release 2018_11. **(a)** Distribution of number of lysine residues. The mode and median of lysine residues of all human proteins are 9 and 21, respectively. On average, lysine residues make up 5.2% of the amino acid residues. About 120 proteins do not contain any lysine residue. **(b)** Distribution of lengths. The average length of human proteins is about 552 residues long.



Supplementary Figure 3.2 Identification of protein fragments.
Distribution of errors as a function of the number of labeled lysine distance measurements for fragments containing the indicated percentage of full-length signal profile: **(a)** 50%; **(b)** 67%; **(c)** 80%. **(d)** Accuracies of identification of fragments containing 7 or more labeled lysine distance measurements. The results are based on the simulations of 1,000 fragments randomly picked from full-length profiles of 2,000 proteins randomly selected from the human proteome database.

3.8 Acknowledgements

This work was supported in part by a grant from the National Institute of General Medical Sciences of the National Institutes of Health (5R01GM126013 to X.H.).

This work, in part, is currently being prepared for submission for publication of the material by Sylvia Liang, Wenxu Zhang, and Xiaohua Huang. Sylvia Liang and the dissertation author were the primary investigators and authors of this material.

Chapter 4 Opto-Electrical System for Nanopore Creation by Controlled Dielectric Breakdown

4.1 Abstract

The ability to fabricate solid-state nanopore quickly and reliably is crucial for nanopore sensing. Nanopore fabrication using cleanroom techniques usually require a lot of labor, time, and expensive instruments. Controlled dielectric breakdown has been demonstrated as a fast and straightforward method to produce stable and functional pores. Here we developed an opto-electrical system for nanopore creation with controlled dielectric breakdown. With the addition of an optical subsystem, it is feasible to monitor the nanopore formation in real-time. We found that the optical detection is more sensitive than electrical measurement during controlled dielectric breakdown.

4.2 Introduction

Nanopore sensing shows great promise for label-free, single-molecule sequencing. It relies on the electrophoretically driven translocation of biomolecules through nanoscale pores embedded in thin insulating membranes, such as a protein channel embedded in a lipid bilayer(43, 67) or a nanopore on a solid-state membrane(68-70). While protein nanopores have already been used in commercial sequencer(71), solid-state nanopores remain extremely attractive due to their superior mechanical stability, scalability, and adjustable surface properties(72).

The solid-state nanopore could be fabricated using transmission electron microscope(12), ion-beam sputtering(70) and focused ion beam(69). However, these techniques are usually involved with complex and high-cost instruments. Furthermore, those nanopores are usually fabricated under a vacuum environment with cleanroom techniques. Nanopore sequencing

experiments are conducted in an aqueous solution, and nanopores fabricated under vacuum environments could behavior significantly different under the aqueous solution.

Controlled dielectric breakdown (CDB) delivers an alternative strategy to fabricate nanopores(68, 73). A sustainable leakage current is observed through the membrane when a high electric field is constantly applied to thin solid-state membranes. The nanopore is created when a sudden irreversible increase in leakage current observed(68).

However, the nanopore could sometimes be created without a sudden change in the current level from our experiments. Ca^{2+} sensitive dyes have been demonstrated as a powerful tool to probe nanopore(74-76). In this work, we developed an opto-electrical system that could demonstrate nanopore creation using CDB with Ca^{2+} sensitive dyes. An electron multiplying charged-coupled device (EMCCD) was used to monitor the CDB in real-time. Furthermore, we used an avalanche photodiode (APD) to detect the fluorescence signal from the nanopore region. This system gives us the capability for synchronized optical and electrical measurement in future single-molecule nanopore translocation experiments.

4.3 Fowcell for optical imaging

First, we designed and built the flowcell for optical imaging. The detailed configuration of the flowcell is illustrated (Figure 4.1). The flowcell could be either made by polycarbonate (PC) or polytetrafluoroethylene (PTFE). When the SiN_x membrane was illuminated by 488 nm laser, we could observe that the laser leaked through the membrane when the SiN_x membrane was aligned with the laser since the membrane is only 12 nm thick, which is less than its penetration depth. Compared to PTFE, PC-based flowcell is transparent and easy to align. However, it does not show chemical resistance to acids or alkalis solution compared with PTFE. At the same time, the epoxy binding to PC is hard to remove.

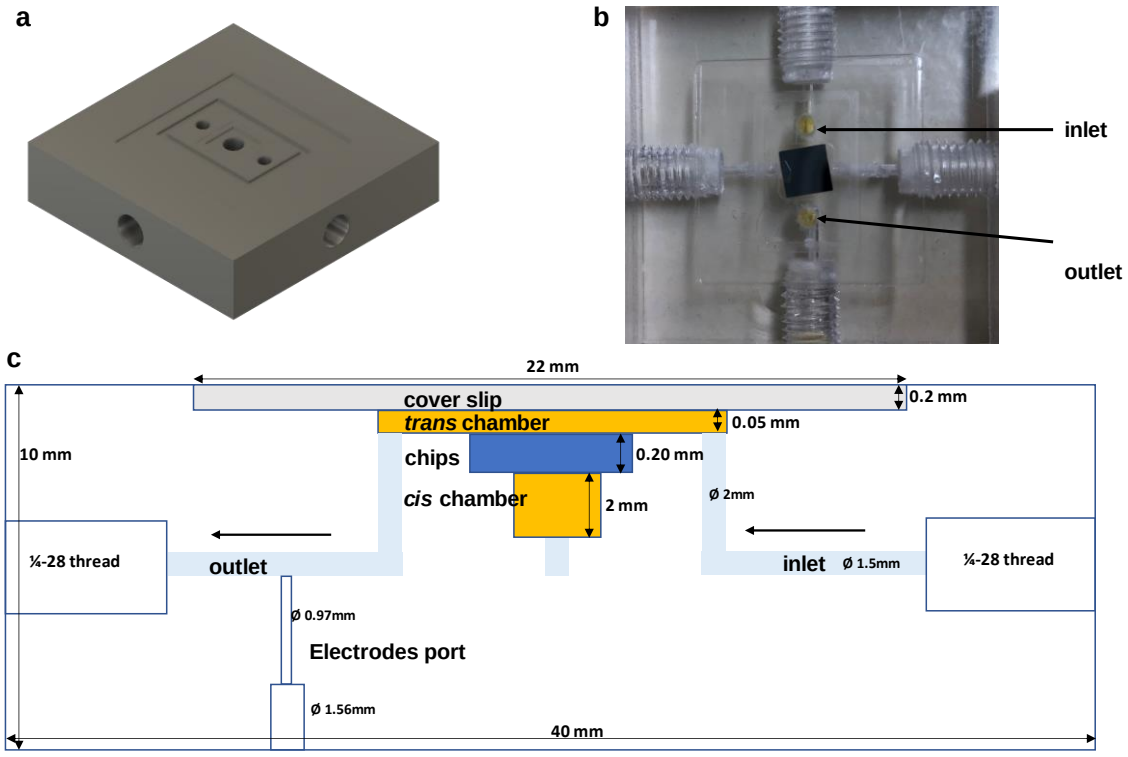


Figure 4.1 Flowcell for fluorescence imaging.
(a) design of flowcell for optical imaging. **(b)** a flowcell made with polycarbonate. **(c)** detailed outline of flowcell from the front side. The inlet, outlet and electrode port for the *cis* chamber are not illustrated.

4.4 Real-time CDB monitor

Then, we built the opto-electrical system for real-time monitor (Figure 4.2). The optical system consists of Epifluorescence, TIRF and confocal module. When the flowcell was loaded on the stage, we first observed the membrane under EMCCD using epifluorescence with white light. The SiN_x membrane was then moved to the center of EMCCD view. Then we turned off the illumination from epifluorescence module and switched to TIRF module for real-time nanopore creation. The 488 nm laser was used for illumination and Fluo-4 was used as the fluorescence molecule in our experiments. The TIRF angle was set to zero degree through

National Instrument analog output channel. A current source was used to apply the potential on the SiN_x membrane, and a multimeter was used to monitor the current level.

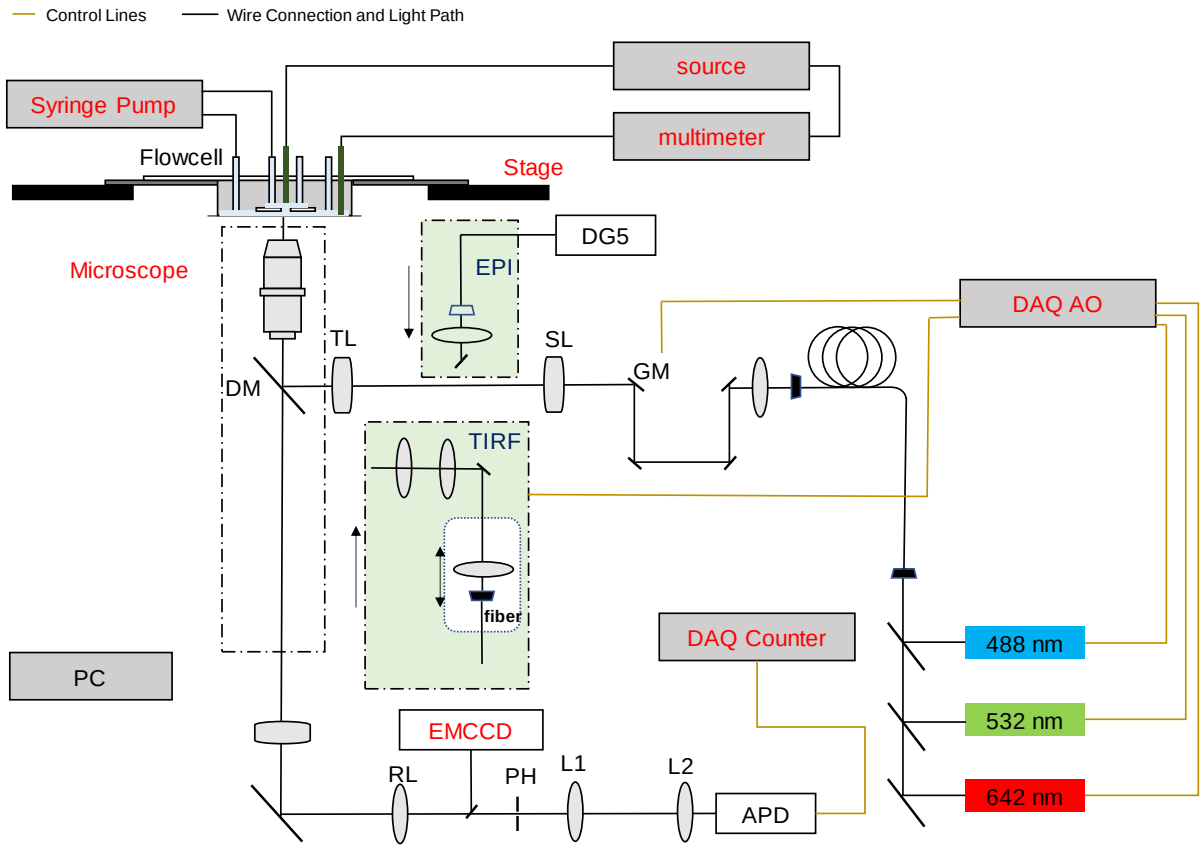


Figure 4.2 optical and electrical measurement system for monitor CDB.
A current source was used to apply potential on the SiN_x membrane, and a multimeter was used to monitor the current level in real time. EPI, confocal and TIRF modules were built in the optical system. EMCCD and APD were used to observe fluorescence signals.

The optical system was configured into TIRF to monitor CDB. The laser was focused on the back focal plane of a 40x objective and then collimated on the *trans* side of silicon membrane. The emission fluorescence signal was collected by same objective. Image was then formed on the EMCCD with 1 second exposure time. The voltage applied on the SiN_x membrane

started from 3V and ramped up 100 mV every 15 seconds. The electrode connected on the *trans* chamber was grounded and the one on the *cis* chamber was positively biased.

The *trans* side of flowcell was filled with 1M KCl, 10 mM Fluo-4 and 10 mM EDTA, the *cis* side of flowcell was filled with 1M KCl and 1M $CaCl_2$. We applied voltage on the silicon membrane and monitored current flow through the membrane (Figure 4.3).

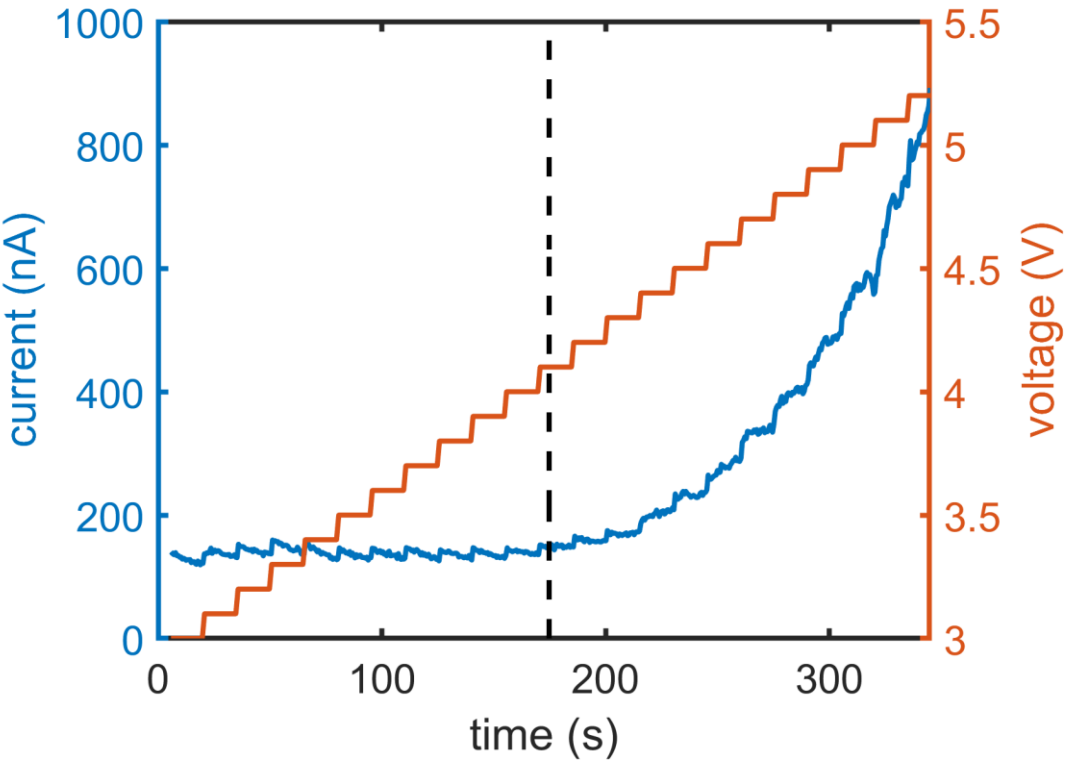


Figure 4.3 Electrical measurement during the CDB.
The red line indicated the ramped voltage applied on the SiN_x membrane. The blue line indicated the current level monitored by the multimeter. The black line indicated the time (174.7 seconds after beginning experiment) that a bright spot was observed under EMCCD image shown in Figure 4.4b.

Fluo-4 has a more than 100-fold fluorescence intensity increase upon binding Ca^{2+}. While a fluorescence enhancement was observed under EMCCD, the nanopore was created on

the SiN_x membrane since the Ca^{2+} ion flow through the nanopore (Figure 4.4). However, a sudden current increase was not observed (Figure 4.3). We concluded that optical detection is more sensitive compared to electrical measurement.

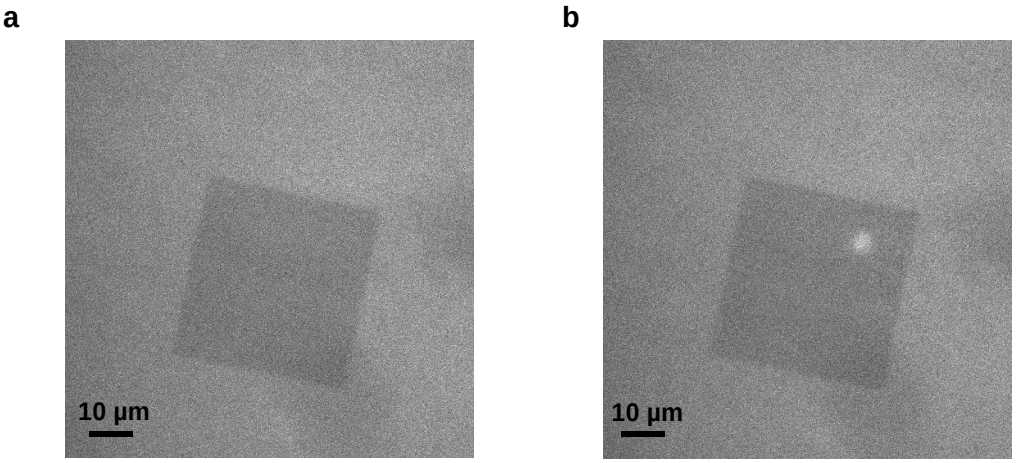


Figure 4.4 Optical monitor during CDB.
(a) Image of Silicon membrane 173.7 seconds after beginning experiment. **(b)** Image of silicon membrane 174.7 seconds after beginning experiment.

4.5 Ion transportation through nanopore

After the nanopore created with CDB, we focused the laser on the pore region and observed fluorescence signal with an APD. We reversed the voltage on the SiN_x membrane between positive and negative every 10 seconds. When positive voltage was applied on the membrane, the Ca^{2+} ions flowed into *trans* chamber through the nanopore and combined with Fluo-4 molecules. A significant fluorescence signal was emitted upon binding Ca^{2+} ions on the Fluo-4 molecules, and the APD recorded high photon counts. While the negative voltage was applied, the Ca^{2+} ions flowed back to the *cis* chamber and low photon counts were recorded by the APD (Figure 4.5).

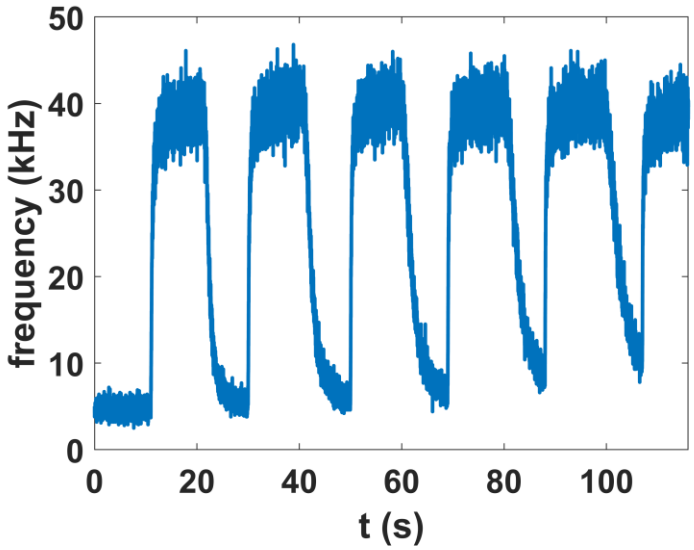


Figure 4.5 photon counts observed under APD.
The voltage polarization across the nanopore modulates the photon counts.

4.6 Method

4.6.1 Dielectric membranes

Silicon nitride (SiN_x) membranes used in experiments are commercially available (Norcada product # NBPX5004Z-AO/O-HR). Each membrane is made of 12 nm thick SiN_x and supported by 200 µm thick silicon frame. There is a 60 nm under-layer SiO_2 present between the SiNx membrane and silicon frame. The under-layer SiO_2 is not present in the membrane area. The Silicon frame is 5 × 5 mm in size and a 0.04 mm × 0.04 mm window on the backside of the Si substrate is opened. SiN_x membrane is wet in pure ethanol before mounting into flowcell.

4.6.2 Flowcell for optical imaging

A silicon chip containing a free-standing SiN_x membrane is mounted ~50 µm from a glass coverslip on a custom-made PTFE flow cell (Figure 4.1). Epoxy (Scotch-Weld Epoxy Adhesive DP420, 3M) is used to bond the silicon chip to the PTFE insert and a No. 1 glass coverslip to the outer cell. The fluidic chamber with Ca^{2+} ions is termed the *cis* chamber and the

fluidic chamber with Fluo-4 molecules is termed the *trans* chamber. Inlets and outlets for both chambers are tapped with a 10 mm depth ¼-28 threads from four sides of PTFE flowcell. The threads are used to connect with pumps and valves using nuts (Idex P-252X, Cole-Parmer) with 1/16" outer diameter tubing (Masterflex Transfer Tubing Microbore PTFE, Cole-Parmer). Top openings in the flow channel are used to insert Ag/AgCl electrodes.

4.6.3 Experiment system setup

4.6.3.1 Fluidic module

The silicon chip is mounted on a custom-made flowcell and the flowcell is mounted on a closed-loop XY stage (Ludl BioPrecision2) with 100 nm accuracy. The stage is on top of a microscope objective. The inlets and outlets of flowcell are connected to syringe pumps and valves (Cavro). Reagents are pulled from valves to flowcell with the syringe pump.

The syringe pump and valve are connected to a computer via serial communication (RS-232). NI-VISA library is used to establish communication between computer and instruments.

4.6.3.2 Optical module

A 488 nm laser diode is used for excitation. The laser diode is installed on a mount (TCLDM9, Thorlabs) and operated by a controller (ITC1xx, Thorlabs). Emitted light is collected and collimated using a broadband optical fiber with transmission efficiency greater than 60% (kineFLEX). The laser either transmits through confocal module or TIRF module.

In the confocal module, the laser passes the galvo mirror (62xxH series, Galvanometer Scanners, Cambridge Technology), scan lens (CSL-SL, Thorlabs), tube lens (TTL 165-A, Thorlabs), a microscope objective, and focuses on the SiNx membrane. In the TIRF module, the laser couples to the TIRF slider (Zeiss), passes a -20 mm (ACN127-020-A[e], Thorlabs) and 25 mm (AC127-025-A[d], Thorlabs) focal length lens separated by 9.9 mm. The TIRF angle could be

adjusted by varying the voltage through National Instrument analog output channels. The laser is focused on the back focal plane of microscope objective and collimated illuminating the SiNx membrane after it.

The emission light is then collected by the objective and filtered using an appropriate long-pass filter. In the TIRF mode, the emitted light is focused on EMCCD camera (ANDOR, iXon 888/897). In the confocal module, the light passes through a 100 µm pinhole (P100D, Thorlabs), collimates using a 60 mm focal length achromatic doublet (AC254-060-A, Thorlabs) and then focuses on an avalanche photodiode (SPCM-AQRH-45, Excelitas) using a 30 mm focal length achromatic doublet (AC254-030-A, Thorlabs). APD is controlled by National Instrument analog output channels and EMCCD connect directly to computer through PCIe controller card.

4.6.3.3 Electrical module

The two chambers filled with liquid electrolyte are electrically connected to a multimeter (Keithley 2010) and a current source (Keithley 6221) to perform CDB. The current source is used as a constant voltage source by setting a high compliance current. Both the multimeter and current source connect to computer with serial communication.

4.6.3.4 System control

A MATLAB package is developed to monitor the electrical and optical signal simultaneously.

4.6.4 Flowcell wetting protocols

The flowcell is washed with DI water, pure ethanol, DI water and 1M KCl. If leakage current has not been observed with a low voltage applied on the SiN_x chip, multiple washes with pure ethanol will be performed.

4.6.5 Fluorescence detection

We use Fluo-4 fluorescence molecules as Ca^{2+} sensitive dyes in our experiments. Pentapossium Fluo-4 (F-14200) salts were purchased from Life Technologies and stored in 10 mM with Milli-Q water at -20 degree until use. The *trans* chamber is filled with 1M KCl, 10 mM EDTA and 10 mM Fluo-4. The *cis* chamber is filled with 1M KCl and 1M $CaCl_2$.

4.7 Conclusions and discussion

In summary, we developed an opto-electrical system to monitor the CDB in real-time with synchronized optical and electrical measurement. A flowcell was designed and built for optical imaging. We developed optical system consisting of epifluorescence, TIRF and confocal modules. The fluorescence signals were further collected by EMCCD or APD. We also used a synchronized electrical measurement to monitor current levels. We wrote a program to control and synchronize the whole system.

We could monitor nanopore creation using fluorescence signal from Ca^{2+} sensitive dyes. However, a sudden irreversible current increase might not be observed when nanopore is created by CDB. We conclude that the optical detection has a higher sensitivity compared to current monitor.

This work laid the foundation for future single-molecule protein identification experiments using high-speed optical detection. we could apply the concept of Förster resonance energy transfer (FRET) on this system. The biological nanopore is labeled with acceptor fluorophores. The protein molecule is denatured, and then certain amino acids are labeled with donor fluorophores. When this protein molecule is translocated through nanopore, an APD could be used to monitor the fluorescence photon signal with MHz bandwidth. The fluorescence

photon signals encode the identity of protein. The system has the potential to achieve single-molecule protein identification in future.

4.8 Acknowledgement

This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health (5R01GM126013 to X.H.).

This work, in part, is currently being prepared for submission for publication of the material by Wenxu Zhang and Xiaohua Huang. The dissertation author was the primary investigator and author of this material.

Chapter 5 Theoretical Feasibility of Nanopore Protein Sequencing

5.1 Abstract

Frederick Sanger first sequenced a protein almost 70 years ago(77-79). However, protein sequencing by Edman degradation(9) and mass spectrometry, the only methods currently available, requires substantial effort, large quantities of protein, and expensive instruments. Here we showed the theoretical feasibility of direct single-molecule protein sequencing using a nanopore. We modeled and numerically calculated the current blockade of all 1.28×10^9 heptamers of 20 natural amino acid residues through nanopores that are potentially realizable experimentally. We simulated sequencing by decoding the computed blockade profiles of proteins from a human proteome database using a forwards-backwards algorithm for Hidden Markov Model (HMM). Random Gaussian noises were introduced into profiles, and each signal was decoded to obtain the sequence. We found that the sequences of the proteins can be decoded with high accuracy using a nanopore with a thickness of 0.76 nm and a diameter of 0.9 nm. Twelve amino acid residues can be identified with greater than 99% accuracy, and three amino acid residues can be identified with greater than 97% accuracy while the other six amino acid residues can be identified as three pairs also with greater than 99% accuracy. In addition to demonstrating a theoretical proof of concept, this work established the computational framework for decoding sequences and provided some physical insights into the experimental realization of nanopore protein sequencing.

5.2 Introduction

Proteins carry out most functions in living systems. Quantitative understanding of biological operating systems and human diseases may require precise determination of the identity or sequence, quantity, and functions of each member of this important class of molecules

and their interaction networks. However, current technologies for protein identification and quantification remain very limited or primitive compared to genomic technologies. The chemical degradation method developed by Pehr Edman(9) in 1949 remained the only method for protein sequencing until the development of the shotgun mass spectrometry (MS) methods for high throughput protein identification and sequencing for routine proteome-scale protein analysis(80-82). However, the determination of protein sequences from the mass spectra of mixtures of fragmented peptides by database searching or *de novo* sequencing is still computationally demanding and error-prone(83, 84). Moreover, both the chemical degradation and MS methods require a substantial number of molecules for analysis. Since proteins cannot be amplified *in vitro,* single-molecule protein sequencing is not feasible using these methods.

The basic idea of sequencing biopolymers using nanopores, conceived initially by George Church and David Deamer more than 20 years ago(10), is to use a nanopore as a sensor to measure the chemo-physical properties (e.g., blockage of ionic current flow) of monomeric units along the linear polymer chain while the polymer is being translocated through the pore. Recently, nanopore DNA sequencing has become a reality even though the nucleobases of DNA consist of a pair of purine bases and a pair of prymidine bases with minimal differences in size and shape(14, 85). As a comparison, the 20 natural amino acid residues of proteins are very diverse in size, shape, and other properties. The difference could be as dramatic as having one hydrogen atom on a glycine versus 18 atoms on a hydrophobic tryptophan side chain. We reason that the nanopore technology is perhaps better suited for protein sequencing. Bailey et al(86) has demonstrated the feasibility of identifying free amino acids using protein nanopores. Akeson et al(87) has reported the controlled unfolding and translocation of protein through protein nanopores. Recently, Wilson et al has also explored the feasibility of protein sequencing using

idealized graphene nanopores and molecular dynamic simulations(88). Considering that the spacing between the neighboring amino acid residues is much smaller than that of the adjacent bases in DNA, it is not apparent that nanopore protein sequencing could be realized.

Here we investigated the theoretical feasibility of direct single-molecule protein sequencing using nanopores. First, we calculated the current blockade of specific combinations of heptamer with 20 natural amino acid residues through a nanopore using finite element analysis (FEA) and the Poisson-Nernst Planck (PNP) equation system. Current blockades of other heptamers were then derived to obtain the entire dataset with all the possible 1.28×10^9 heptamers. We then simulated the current blockade profiles of all proteins from a human proteome database. Finally, we investigated the feasibility of decoding amino acid sequences from the profiles.

5.3 Method

5.3.1 General introduction

Molecular Dynamics (MD)(89, 90), Brownian Dynamics (BD)(91, 92), and Poisson-Nernst-Planck (PNP) theory(93, 94) are wildly used to simulate ionic transportation in narrow channels. MD and BD methods can provide realistic results but at the cost of computing resources.

Here we used FEA method with COMSOL software to study the current blockage by protein molecule. The nanopore used in our study was simplified as a simple cylinder (Figure 5.1). The scale potential inside the fluidic system follows the Poisson equation with proper boundary conditions.

$$\nabla^2 \Phi(r, \phi, z) = -\frac{\rho}{\varepsilon_0 \varepsilon_r}$$

$$\Phi(z = 0) = 0$$

$$\Phi(z = L) = V_0$$

$$\frac{\partial \Phi}{\partial n}|_S = \frac{\sigma}{\varepsilon_0 \varepsilon_r}$$

where r is the radial coordinates, ϕ is the azimuth, and z is the height, respectively, defined in the cylindrical coordinate system; $\rho = F \times (c_{K^+} - c_{Cl^-})$ is the space charge density; F is the faraday constant; σ is the surface charge density on the nanopore; ε_0 is the vacuum permittivity; ε_r is the relative permittivity of transportation medium, $z = 0$ is the bottom surface of nanopore; $z = L$ is the top surface of nanopore; and S is the boundary of the fluidic system. We assumed the cylindrical surface is no charge.

When a linear single-stranded DNA is electrophoretically driven through a protein nanopore such as α-hemolysin, the characteristic resident time scale is about 1 μs per base. The free diffusion of K^+ and Cl^- ions ($D \approx 1.0 \times 10^{-9}$ m^2 s^{-1}) over a length scale of about 1 nm takes only about 0.5 ns, much smaller than the time required to translocate a nucleotide through the nanopore. The translocation of a polypeptide through a nanopore is very likely much slower than that of a single-stranded DNA molecule. Therefore, the concentrations of the ions can be assumed to reach a steady equilibrium state during the translocation of the polypeptide, and the flux of the ions through the nanopore can be calculated using the steady-state assumption.

The nanopore constriction is assumed to be neutral in our model. However, the side chains of the amino acid residues lining the surface of the protein nanopore constriction site may carry a charge or dipole depending on the pH and solvent used. The relative permittivity of water in nano confinements is more than ten times smaller compared to bulk freely diffusing water due to the significant decrease of rotational freedom of water dipoles near surfaces(46). The Debye

length (in angstrom) of an ionic solution inside the nanopore can be estimated using the Debye-Hückel theory. For a 400 mM KCl in a nanopore, the Debye length is calculated to be 0.765 Å, which is less than 10% of the diameter of a 1 nm nanopore.

5.3.2 Nanopore and protein COMSOL model

Multiple amino acid residues close to the nanopore constriction site contribute to ionic current blockades. Therefore, we constructed the physical models of the amino acid heptamers inside the nanopores and solved the PNP equations using COMSOL Multiphysics software. Two modules were employed: the electrostatics module for solving the Poisson equation and the transport of diluted species module for solving the Nernst-Planck equation. The nanopore was modeled as a hollow cylinder with axial symmetry (Figure 5.1, 5.2) and the model was solved in the steady-state.

To reduce computations required for calculating the ionic flux through a 3D nanopore, we used a 2D model to compute the intensity of the ionic flux and then integrated 360 degrees around the cylindrical axis to obtain the total flux.

The 2D model of the device used for simulation is shown in Figure 5.1. The device consists of a cylinder with a radius of 100nm and a height of 200 nm, which is divided into two chambers by a 0.76 nm thick membrane (about two times the linear length of a peptide bond) with a 0.9 nm nanopore in the center. In addition, we also modeled two different cylindrical nanopores, one with the same diameter but a greater thickness (1.14 nm), and one with the same thickness but a greater diameter (1.1 nm).

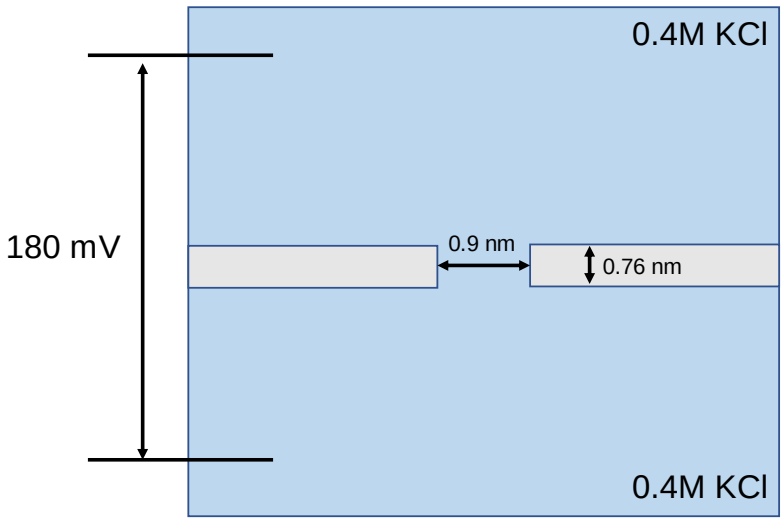


Figure 5.1 2D model of nanopore device
The two cylindrical chambers (in blue) were filled with 0.4 M KCl solution and were separated by a 0.76 nm thick membrane (in gray). A 180-mV potential was applied to the two chambers.

Protein residues were modeled as cylinders, and the height of each residue was assumed as 3.8Å for all amino acids. The cylinder radius for each amino acid residue was calculated based on the experimentally determined hydrodynamic volume(95) (Table 5.1). Based on the initial modeling of homopolymers, we found that for each residue at the vertical center of the nanopore, up to six adjacent residues (three above and three below) contribute to current blockades. Therefore, we modeled current blockades of heptamers inside the nanopore. The voltage applied across the nanopore was set at 180 mV, and the concentration of KCl was set to 0.4 M. The surface of the nanopore and amino acid residues were treated as neutral for the study. Proper boundary conditions were applied. The electric field distribution of the nanopore was shown in Figure 5.2. The current blockades were calculated by taking the ionic flux of both K^+ and Cl^- into account

$$B = 1 - \frac{I_B}{I_O} = 1 - \frac{I_{B^+} + I_{B^-}}{I_{O^+} + I_{O^-}} = 1 - \frac{\int ds(J_{B^+} + J_{B^-})}{\int ds(J_{O^+} + J_{O^-})}$$

where I_B is the current when the heptamer blocks the nanopore; I_O is current of open nanopore; I_{B^+}, $I_{O^+}$and I_{B^-}, $I_{O^-}$are ionic currents of K^+ and Cl^- with/without amino acid residues inside nanopore; J_{B^+}, J_{O^+} and J_{B^-}, J_{O^-} are ionic current densities of K^+ and Cl^- with/without amino acid residues inside the nanopore.

Table 5.1 Amino acid volume and radius information

Residues	Volume(Å^3)	Radius(Å)	Residues	Volume(Å^3)	Radius(Å)
Gly(G)	63.8	2.31	*Gln(Q)*	148.1	3.52
Ala(A)	92.7	2.79	*His(H)*	156	3.61
Ser(S)	92.8	2.79	*Lys$^+$(K)*	162.4	3.69
Asp$^-$(D)	105.2	2.97	*Met(M)*	167	3.74
Cys(C)	114.2	3.09	*Ile(I)*	167.5	3.75
Thr(T)	119.7	3.17	*Leu(L)*	170.8	3.78
Asn(N)	120.5	3.18	*Arg$^+$(R)*	184.9	3.94
Pro(P)	129.3	3.29	*Phe(F)*	194.4	4.04
Glu$^-$(E)	132.3	3.33	*Tyr(Y)*	197.4	4.07
Val(V)	142.7	3.46	*Trp(W)*	231.1	4.4

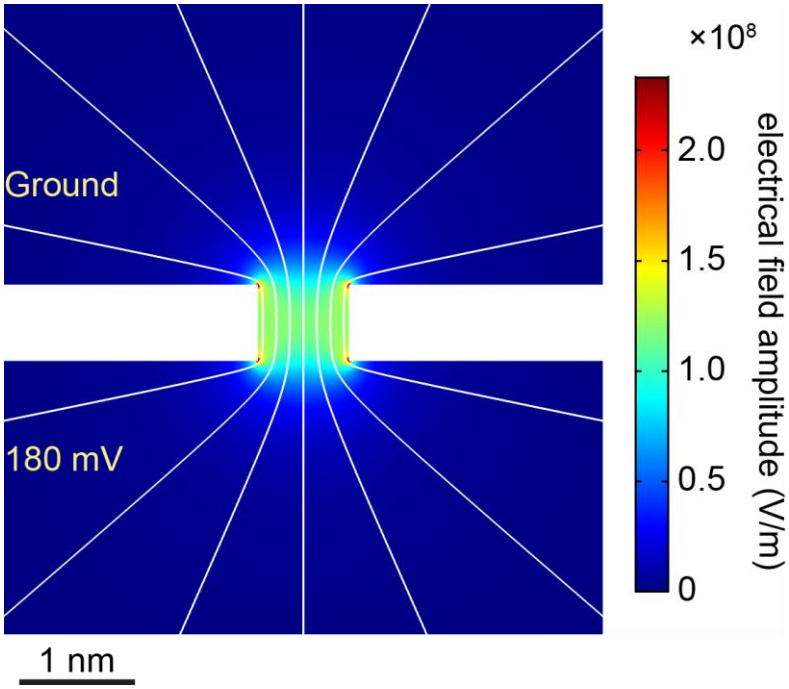


Figure 5.2 Electric field distribution through a nanopore
The applied voltage was 180 mV, and the concentration of the KCl electrolyte solution was 0.4 M. The white lines represented the electric field lines and the color represented the electric field intensity.

5.3.3 Current blockades of protein k-mers

As expected, the current blockades of any given amino acid in the center of the nanopore do not have a fixed value. We modeled the current blockages of triplets, pentamers, and heptamers of amino acid homopolymers to evaluate how many neighboring residues contribute to current blockades. Three amino acids, glycine, valine, and tryptophan were used in this study considering the various size of amino acids. We calculated the influence of the current blockades by adding two additional residues (one above and one below) of three different sizes (glycine, valine, and tryptophan) into triplets/pentamers/heptamers models. The results are listed in Tables 5.2–5.4. The addition of two larger residues increased the current blockage above noise level (0.2%) for pentamers consisting of small or even medium residues. For heptamers, the addition

of two more residues had little influence on the current blockages of most combinations except the ones consisting of the small residues such as glycine (Table 5.4).

Table 5.2 Current blockades of triplets and the influence of two additional nearest neighboring residues on their blockades for a nanopore of 0.76 nm in thickness and 0.9 nm in diameter.

	% blockade	% blockade (% change as compared to triplet*)		
Triplet X	Triplet X	G + X + G	V + X + V	W + X + W
GGG	17.2	17.7 (**2.91%**)	20.2 (**17.44%**)	23.8 (**38.37%**)
VVV	45.4	45.5 (**0.22%**)	45.8 (**0.88**%)	47.0 (**3.52%**)
WWW	92.0	92.0 (0.00%)	92.0 (0.00%)	92.0 (0.00%)

Table 5.3 Current blockades of pentamers and the influence of two additional nearest neighboring residues on their blockades for a nanopore of 0.76 nm in thickness and 0.9 nm in diameter.

	% blockade	% blockade (% change as compared to pentamer*)		
Pentamer X	Pentamer X	G + X + G	V + X + V	W + X + W
GGGGG	17.7	17.8 (**0.56%**)	18.1 (**2.26%**)	18.6 (**5.08%**)
VVVVV	45.8	45.9 (**0.22%**)	45.9 (**0.22%**)	46.0 (**0.44%**)
WWWWW	92.0	92.0 (0.00%)	92.0 (0.00%)	92.0 (0.00%)

Table 5.4 Current blockades of heptamers and the influence of two additional nearest neighboring residues on their blockades for a nanopore of 0.76 nm in thickness and 0.9 nm in diameter.

	% blockade	% blockade (% change as compared to heptamer*)		
Heptamer X	Heptamer X	G + X + G	V + X + V	W + X + W
GGGGGGG	17.8	17.9 (**0.56%**)	17.9 (**0.56%**)	18.0 (**1.12%**)
VVVVVVV	45.9	45.9 (0.00%)	45.9 (0.00%)	46.0 (**0.22%**)
WWWWWWW	92.0	92.0 (0.00%)	92.0 (0.00%)	92.0 (0.00%)

5.3.4 Modeling and computations of reference data set of all amino acid heptamers

In theory, we must model all 1.28×10^9 ($= 20^7$) heptamers in the nanopore to obtain the reference dataset. But it is unrealistic computationally (it would take more than 100 years to compute even with our powerful workstation). It is also infeasible to acquire the reference dataset for all heptamers experimentally. However, we cannot decode the amino acid sequences from experimental data without the reference dataset. Therefore, we modeled the current

blockades of a small but representative subset, and then completed the rest of the dataset by prediction. The prediction accuracy was validated by modeling a random set. Then modeled current blockades were compared with predicted ones to ensure the error was smaller than a preset threshold. The amino acid heptamers were represented by $X_{-3}X_{-2}X_{-1}X_0X_1X_2X_3$, with X_{-3} to X_3 being any of the 20 natural amino acids and X_0 being the residue at the center of the nanopore constriction site. A 3D matrix with 8000 × 400 × 400 entries could fully describe the heptamer dataset (Figure 5.3). Each value in the matrix represented the current blockade for a specific amino acids heptamer. Any entry in this matrix was associated with a set of unique index (i, j, k), where $1 \le i \le 8000$, $1 \le j \le 400$ and $1 \le k \le 400$. The index *i* determined the centered three amino acids, X_{-1}, X_0 and X_1. The index *j* determined the amino acids X_{-2} and X_2. The index *k* determined the amino acids X_{-3} and X_3.

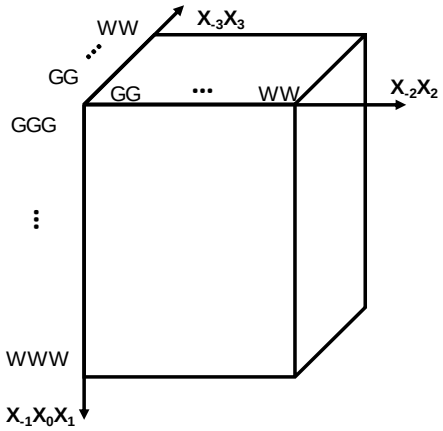


Figure 5.3 Heptamer 3D reference dataset
A 3D matrix with 8000 × 400 × 400 entries could fully describe the heptamer dataset. The index of the matrix uniquely determined the amino acid heptamers, and the value of each entry represented the current blockades.

Intuitively, the residues at the center of the nanopore constriction site and its two nearest neighbors very likely dominant the current blockage. The ability to accurately predict the current

blockades of the heptamers requires high accuracy in modeling the triplet at the center of the nanopore. Therefore, we first fixed the amino acids on the sides of the heptamer and modeled current blockages for varied centered triplets. Then, we figured out how the current blockades changed with different amino acids on the side of the heptamer. The relation was used to predict unknown heptamer current blockades.

5.3.4.1 Heptamer dataset for $GGX_{-1}X_0X_1GG$

First, we modeled a selected set of the complete 8000 (= 20^3) triplets of amino acids in the context of the polypeptide backbone, that is, by fixing the X_{-3}, X_{-2}, X_2 and X_3 as glycine in the heptamers. In the heptamer 3D reference dataset, the dataset for $GGX_{-1}X_0X_1GG$ is present in the left column of the front surface and this partial set could be then reshaped into another 3D matrix (Figure 5.4a). Three indexes of this new 3D matrix determined amino acid species for X_{-1}, X_0, and X_1 (Figure 5.4a).

Since there are always a pair of heptamers that are symmetrical (with a C2 symmetry) and have the same current blockage level (e.g., GGDSTGG and GGTSDGG), we assumed that the radius of X_{-1} is equal to or smaller than X_1.

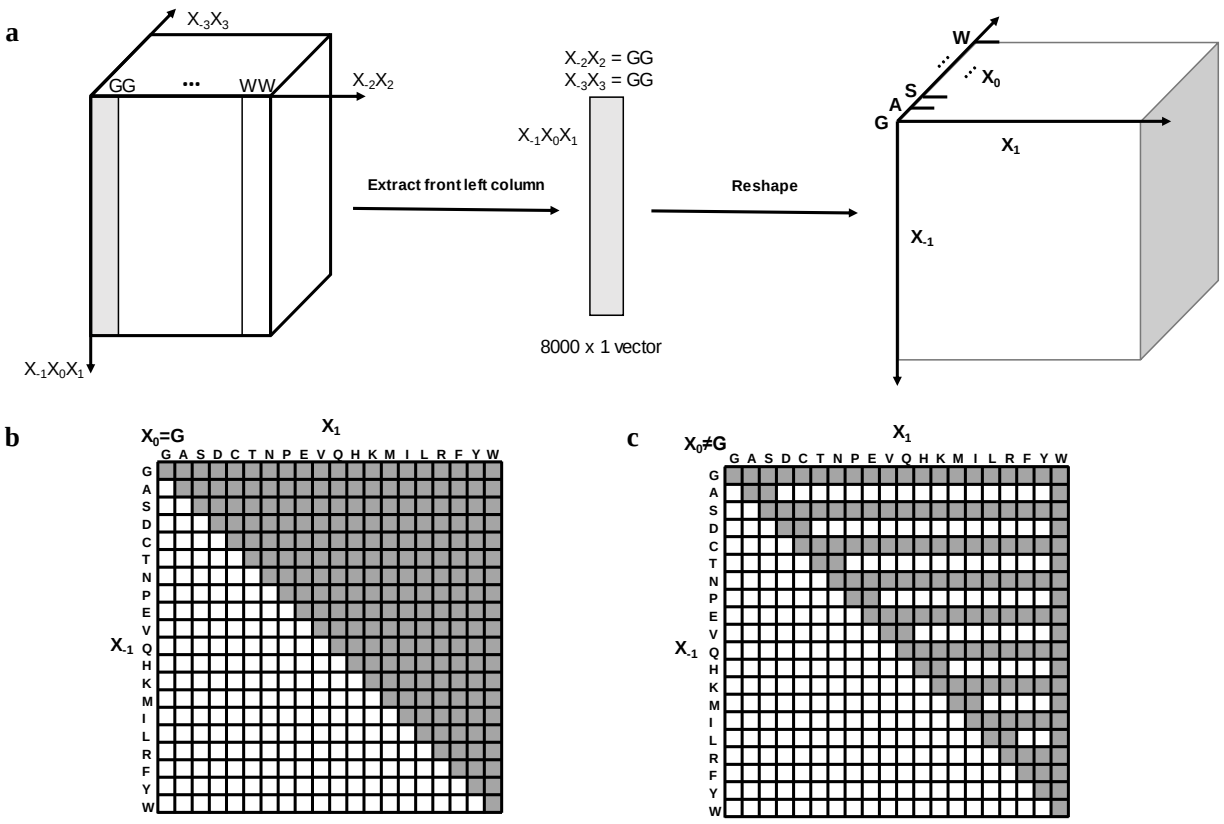


Figure 5.4 Heptamer dataset generation for GGX$_{-1}$X$_0$X$_1$GG
a) GGX$_{-1}$X$_0$X$_1$GG heptamer dataset 3D matrix representation. The front left column of the whole heptamer dataset represents heptamer with sides four amino acids as glycine. The 8000 × 1 vector could be further reshaped into a new 3D matrix with 20 × 20 × 20 entries. **b)** first layer in the GGX$_{-1}$X$_0$X$_1$GG heptamer dataset (X$_0$ is glycine). The gray heptamers were modeled. **c)** layer in GGX$_{-1}$X$_0$X$_1$GG heptamer dataset if amino acid in the center is not glycine. The gray heptamers were modeled.

We divided twenty amino acids into two categories: 1) class 1 contains the amino acids G, S, C, N, E, Q, K, I, R, and Y; 2) class 2 contains the amino acids A, D, T, P, V, H, M, L, F, and W.

If X$_0$ is glycine, all possible combinations of X$_{-1}$ and X$_1$ were modeled (Figure 5.4b). If X$_0$ is not glycine, we modeled heptamers in which the following criteria were met: 1) X$_{-1}$ and X$_1$ are the same amino acids; 2) the radius of X$_1$ is slightly larger than the radius of X$_{-1}$; 3) X$_{-1}$ belongs to amino acids in class 1; 4) X$_1$ is tryptophan (Figure 5.4c).

Next, current blockades for unmodeled $GGX_{-1}X_0X_1GG$ heptamers were predicted by linear regression and error correction (Procedure 5.1). We assumed that X_1 and X_1 influenced current blockades in a similar pattern for various X_0. Since current blockades were modeled for heptamer where X_0 is glycine, current blockades should vary similarly for other amino acids in the center. Then, an error correction step increased prediction accuracy by taking more models into accounts.

Procedure 5.1 The procedure to compute $GGX_{-1}X_0X_1GG$ heptamer reference dataset

Notation: X_{-1b} belongs to amino acids in class 1 and is slightly smaller than X_{-1}. X_{-1u} belongs to amino acids in class 1 and is slightly larger than X_{-1}

For X_0 is not glycine:

For X_{-1} belongs to class 2 amino acids and X_1 with a radius larger than X_{-1}:

Calculated linear regression coefficient

$$\alpha = \frac{B(GGX_{-1u}X_0X_1GG) - B(GGX_{-1b}X_0X_1GG)}{B(GGX_{-1u}GX_1GG) - B(GGX_{-1b}GX_1GG)}$$

Estimated linear regression current level

$$esti(GGX_{-1}X_0X_1GG) = B(GGX_{-1b}X_0X_1GG) + \alpha \times (B(GGX_{-1}X_0X_1GG) - B(GGX_{-1b}X_0X_1GG))$$

Estimated correction coefficient

$$\beta = \frac{esti(GGX_{-1}X_0X_1GG) - esti(GGX_{-1}X_0X_{-1u}GG)}{esti(GGX_{-1}X_0WGG) - esti(GGX_{-1}X_0X_{-1u}GG)}$$

Estimated correction term

$$corr(GGX_{-1}X_0X_1GG) = corr(GGX_{-1}X_0X_{-1u}GG) + \beta \times (corr(GGX_{-1}X_0WGG) - corr(GGX_{-1}X_0X_{-1u}GG))$$

where

$corr(GGX_{-1}X_0WGG) = B(GGX_{-1}X_0WGG) - esti(GGX_{-1}X_0WGG)$ and
$corr(GGX_{-1}X_0X_{-1u}GG) = B(GGX_{-1}X_0X_{-1u}GG) - esti(GGX_{-1}X_0X_{-1u}GG)$

Then predicted current blockade

$$pred(GGX_{-1}X_0X_1GG) = esti(GGX_{-1}X_0X_1GG) + corr(GGX_{-1}X_0X_1GG)$$

End

End

5.3.4.2 Heptamer dataset for $GX_{-2}X_{-1}X_0X_1X_2G$

In the heptamer 3D reference dataset, the dataset for $GX_{-2}X_{-1}X_0X_1X_2G$ is present as the front surface (Figure 5.5a). For each fixed $X_{-1}X_0X_1$, we first modeled and computed heptamers where X_{-2} and X_2 were individually selected from glycine, valine, and tryptophan. Then, a polynomial fitting was applied to predict all other combinations of X_{-2} and X_2.

We first modeled a partial of heptamer dataset for $GX_{-2}X_{-1}X_0X_1X_2G$ where X_{-2} and X_2 were one of glycine, valine, and tryptophan. The heptamer dataset $GX_{-2}X_{-1}X_0X_1X_2G$ could be reshaped to a new 3D matrix for fixed X_{-2} and X_2. Each layer of this matrix represented heptamer with fixed X_0. Each column in this layer described heptamer with fixed X_0 and X_1. Each row in this layer represented the heptamer model with fixed X_0 and X_{-1}. We assigned class 3 amino acids to contain G, E, L, and W. Class 4 amino acids include G, C, V, I, and W. Heptamers, where X_{-1}, X_1 belong to class 3 amino acids, X_0 belong to class 4 amino acids were modeled. (Figure 5.5a)

We then predicted heptamer $GX_{-2}X_{-1}X_0X_1X_2G$, where X_{-2} and X_2 were individually selected from glycine, valine, and tryptophan (Procedure 5.2). The case where both X_{-2} and X_2 were glycine was discussed in the previous section. Current blockades are not linearly related to the volume of amino acids. A large amino acid has a large current blockage. If the size of amino acid increases, the current blockage increases. However, it increases less for a larger amino acid in the center of the nanopore. For $GX_{-2}X_{-1}X_0X_1X_2G$ with fixed centered triplets, the current blockade is higher when both X_{-2} and X_2 are larger amino acids compared to glycine. And the current blockade difference caused by X_{-2} and X_2 is related to volume of centered triplets. The current blockade difference is smaller if centered triplets are larger. We assumed the current

blockade difference had a polynomial relation with the volume of triplets in the center (Figure 5.5b) or the current blockade of triplets with glycine on the edge of heptamer (Figure 5.5c).

Procedure 5.2 The procedure to compute $GX_{-2}X_{-1}X_0X_1X_2G$ heptameter reference dataset for X_{-2} and X_2 selected from glycine, valine, and tryptophan

For X_{-2} and X_2 selected from glycine, valine, and tryptophan and cannot be glycine at the same time:

For X_0 is one of amino acids in class 3 and X_{-1} is one of amino acids in class 4:

Calculated current blockades difference ΔB for all X_1 belongs to amino acids in class 4 by

$$\Delta B = B(GX_{-2}X_{-1}X_0X_1X_2G) - B(GGX_{-1}X_0X_1GG)$$

Fit the relation between ΔB and Volume of X_1 with a polynomial function, estimated current blockades difference, and predicted current blockades for X_1 does not belong to amino acids in class 4.

For X_0 is one of amino acids in class 3 and X_1 is one of 20 amino acids:

Calculated current blockades difference ΔB for all X_{-1} belongs to amino acids in class 4 by

$$\Delta B = B(GX_{-2}X_{-1}X_0X_1X_2G) - B(GGX_{-1}X_0X_1GG)$$

Fit the relation between ΔB and Volume of X_{-1} with a polynomial function, estimated current blockades difference, and predicted current blockades for X_{-1} does not belong to amino acids in class 4.

For X_0 does not belong to amino acids in class 3:

Calculated current blockades difference ΔB for all X_0 belongs to amino acids in class 3 by

$$\Delta B = B(GX_{-2}X_{-1}X_0X_1X_2G) - B(GGX_{-1}X_0X_1GG)$$

Fit the relation between ΔB and $B(GGX_{-1}X_0X_1GG)$ with a polynomial function, estimated current blockades difference, and predicted current blockades for X_0 does not belong to amino acids in class 3.

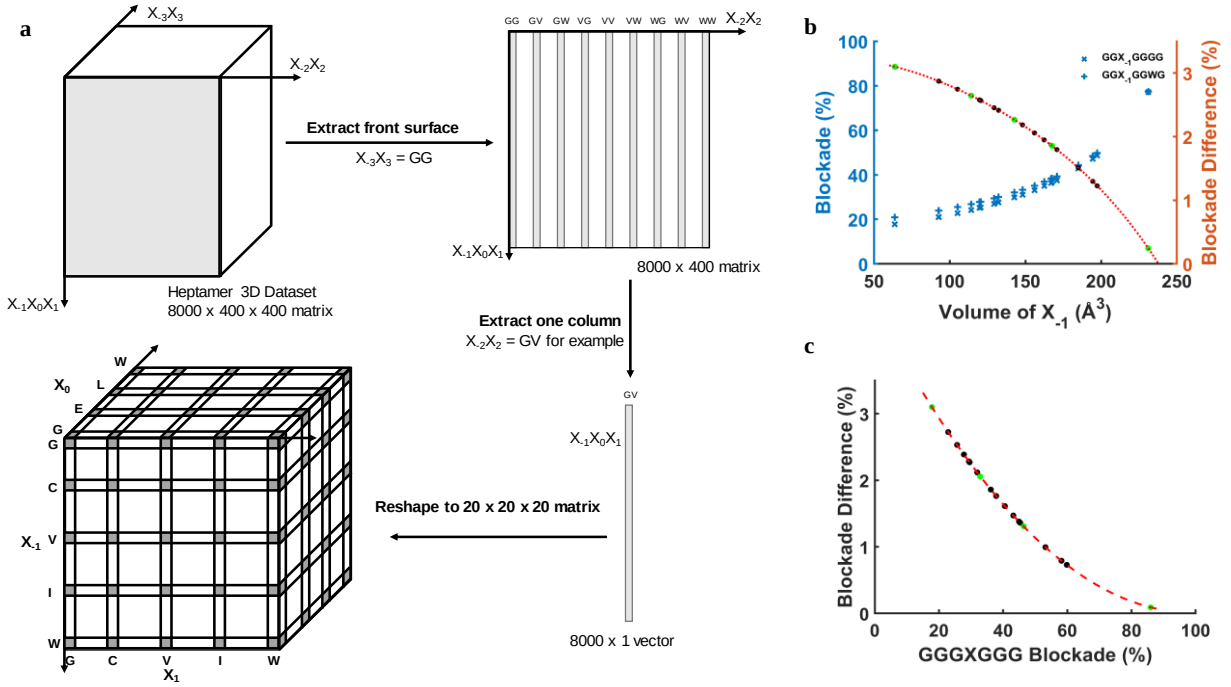


Figure 5.5 Heptamer dataset generation for GX$_{-2}$X$_{-1}$X$_0$X$_1$X$_2$G with X$_{-2}$ and X$_2$ selected from glycine, valine, and tryptophan

a, GX$_{-2}$X$_{-1}$X$_0$X$_1$X$_2$G heptameter reference dataset for X$_{-2}$ and X$_2$ selected from glycine, valine, and tryptophan. **b,** Heptamer dataset generation for varied X$_{-1}$. The relation between current blockades difference and Volume of X$_{-1}$ was first to fit with a polynomial function (red dot curve), and then current blockades for X$_{-1}$ not in class 4 were predicted (blue dot). **c,** Heptamer dataset generation for varied X$_0$ by polynomial fitting. The relation between current blockades difference and current blockades for GGGXGGG was fit with a polynomial function.

Next, a polynomial fitting was applied to predict all other combinations of X$_{-2}$ and X$_2$. For heptamer GX$_{-2}$X$_{-1}$X$_0$X$_1$X$_2$G with fixed X$_{-1}$X$_0$X$_1$, we modeled or predicted X$_{-2}$, X$_2$ as a combination of glycine, valine, and tryptophan. We calculated the current blockade difference between GX$_{-2}$X$_{-1}$X$_0$X$_1$X$_2$G and GX$_{-2}$GGGX$_2$G. Then, we fit a polynomial relation between the current blockade difference and current blockade for heptamer GX$_{-2}$GGGX$_2$G. Lastly, we estimated current blockades for all GX$_{-2}$X$_{-1}$X$_0$X$_1$X$_2$G (Procedure 5.3).

Procedure 5.3 The procedure to compute $GX_{-2}X_{-1}X_0X_1X_2G$ heptameter reference dataset

For fixed triplet $X_{-1}X_0X_1$ in the centered:

For X_2 belongs to one of glycine, valine, or tryptophan:

Calculated current blockades difference ΔB for all X_{-2} belongs to glycine, valine, or tryptophan by

$$\Delta B = B(GX_{-2}X_{-1}X_0X_1X_2G) - B(GX_{-2}GGGX_2G)$$

Fit the relation between ΔB and $B(GX_{-2}GGGX_2G)$ with a polynomial function, estimated current blockades difference, and predicted current blockades for X_{-2} does not belong to one of glycine, valine, and tryptophan.

End

For X_{-2} belongs to one of 20 amino acids:

Calculated current blockades difference ΔB for all X_2 belongs to glycine, valine, or tryptophan by

$$\Delta B = B(GX_{-2}X_{-1}X_0X_1X_2G) - B(GX_{-2}GGGX_2G)$$

Fit the relation between ΔB and $B(GX_{-2}GGGX_2G)$ with a polynomial function, estimated current blockades difference, and predicted current blockades for X_2 does not belong to one of glycine, valine, and tryptophan.

End

End

Followed the similar procedure, we also generated the heptamer dataset for $WX_{-2}X_{-1}X_0X_1X_2W$. We first generated dataset for $WGX_{-1}X_0X_1GW$, $WGX_{-1}X_0X_1VW$, $WGX_{-1}X_0X_1WW$, $WVX_{-1}X_0X_1VW$, $WVX_{-1}X_0X_1WW$, $WWX_{-1}X_0X_1WW$ by Procedure 5.2. We then used the Procedure 5.3 to generate the heptamer dataset for $WX_{-2}X_{-1}X_0X_1X_2W$. It is worth mentioning that we also use Procedure 5.2 to generate heptamer reference dataset for $WX_{-2}X_{-1}X_0X_1X_2W$. By applying this change, we further reduced the number of models.

5.3.4.3 Heptamer dataset for $X_{-3}X_{-2}X_{-1}X_0X_1X_2X_3$

We modeled heptamer $X_{-3}GGGGGX_3$ for all possible X_{-3} and X_3, and then predicted all heptamer dataset by

$$\begin{aligned} &esti(X_{-3}X_{-2}X_{-1}X_0X_1X_2X_3) \\ &= B(GX_{-2}X_{-1}X_0X_1X_2G) + \gamma \times (B(X_{-3}GGGGGX_3) - B(GGGGGGG)) \end{aligned}$$

where γ is the correction coefficient and could be expressed as

$$\gamma = \frac{B(WX_{-2}X_{-1}X_0X_1X_2W) - B(GX_{-2}X_{-1}X_0X_1X_2G)}{B(WGGGGGW) - B(GGGGGGG)}$$

5.3.4.4 Validation of Heptamer Dataset Generation

We randomly modeled 3000 protein heptamers and calculated current blockades. Then calculated current blockades were compared with predicted ones to prove our reference dataset generation strategy was correct. The average absolute offset between predicted data and modeled data was 0.02% (Figure 5.6). The variance is much lower than the noise level of 0.2%.

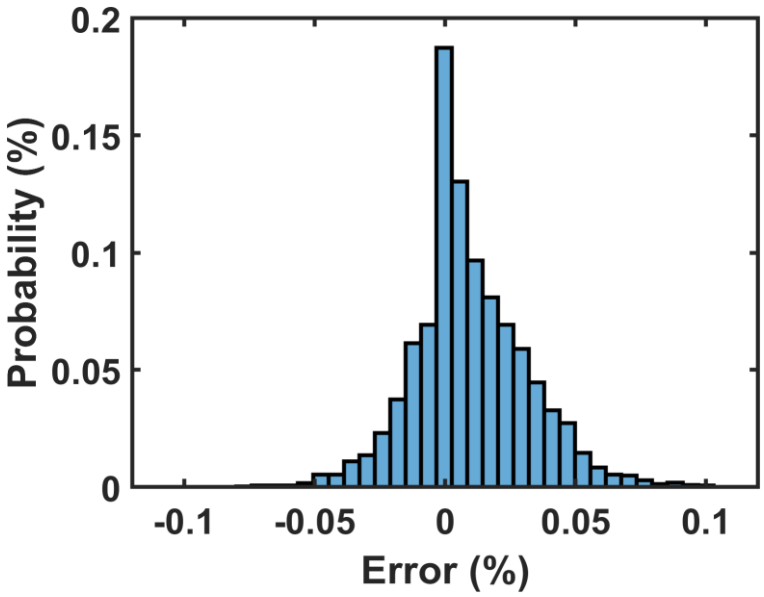


Figure 5.6 Heptamer prediction accuracy
This figure showed the absolute difference between predicted signals and modeled signals for 3000 randomly selected heptamers.

5.3.5 Generation of synthetic current profiles

Synthetic current profiles were generated from the Hidden Markov Model. Current Blockades $B_1, B_2, \ldots, B_{t-1}, B_t, B_{t+1}, \ldots, B_T$ were observed at different time points and blockade level B_t was uniquely determined by hidden heptamer state H_t (Figure 5.7).

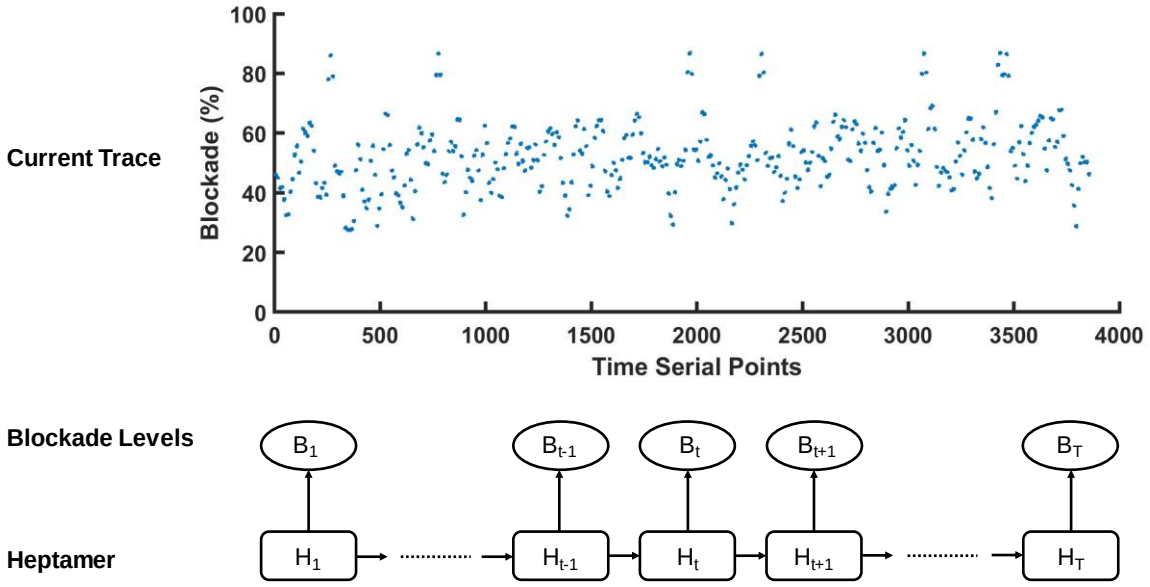


Figure 5.7 Synthetic signal generation
This figure showed how synthetic current trace was generated from protein sequence. The protein sequence was first represented as heptamers chain and then current blockades for those heptamers were extracted. Random noise was added to the current blockades to simulate the real experiment current trace.

An alteration, as small as 0.2%, of the blockade results in a change of 1.0 pA, which technically can be detected using a commercially available amplifier such as the Axopatch 200B., There are usually N data points in real measurements $B_{t,1}$, $B_{t,2}$,…,$B_{t,N}$ instead of one single measurement B_t corresponding to a heptamer state H_t. We used N = 10 for our data analysis. If one measurement follows a normal distribution with variance equals to σ, then the mean of N measurements will follow a normal distribution with a variance equals to $\frac{\sigma}{\sqrt{N}}$.

5.3.6 Pseudo-Heptamer decoding algorithms

We developed a Pseudo-Heptamer decoding algorithms to determine the amino acid sequences from simulated current blockade profiles. In the forwards-backwards algorithms(96), probability of heptamer H_t corresponding to the observation B_t at time point t was estimated by

$$p(H_t = K|B_{1:T}) = p(H_t = K|B_{1:t}, B_{t+1:T}) \propto p(H_t = K|B_{1:t})p(B_{t+1:T}|H_t = K) = \alpha_t(K)\beta_t(K)$$

where $\alpha_t(K) = p(H_t = K|B_{1:t})$ is the filtered marginals given previous message and $\beta_t(K) = p(B_{t+1:T}|H_t = K)$ is the conditional likelihood of future evidence given the hidden state at time t. $\alpha_t(K)$ and $\beta_t(K)$ are recursively calculated from the beginning and the tail of the signal by

$$\alpha_t(K) = p(H_t = K|B_{1:t}) = p(H_t = K|B_t, B_{1:t-1}) \propto p(B_t|H_t = K)p(H_t = K|B_{1:t-1})$$

$$= p(B_t|H_t = K)\sum_J p(H_t = K|H_{t-1} = J)p(H_{t-1} = J|B_{1:t-1}) = \psi_t(K)\sum_J \Psi(J,K)\alpha_{t-1}(J)$$

$$\beta_{t-1}(K) = p(B_{t:T}|H_{t-1} = K) = \sum_J p(B_{t+1:T}|H_t = J)p(B_t|H_t = J)p(H_{t-1} = K|H_t = J)$$

$$= \sum_J \beta_t(J)\,\psi_t(J)\Psi(J,K)$$

Where $\psi_t(K) = p(B_t|H_t = K)$ is the local evidence at time t and $\Psi(J,K) = p(H_t = K|H_{t-1} = J)$ is the transition matrix.

However, we adapted a heptamer model in hidden states, and there were in total 1.28 x 10^9 (20^7) unique states. The transition matrix $\Psi(J,K)$ contains 20^7 x 20^7 entries. $\psi_t(K)$, $\alpha_t(K)$ and $\beta_t(K)$ are all 1 x 20^7 vector. This sheer amount of memory and computations required to find the right paths through the hidden states would not be feasible even with a supercomputer in the traditional forwards-backwards algorithms. However, in our case, the current blockade signals of most heptamers are predominantly contributed by the center amino acids, with little contribution from the amino acids at the edges of heptamers as shown earlier (Tables 6.2 – 6.4). We could use this information to reduce the complexity and search space required by the forwards-backwards algorithms.

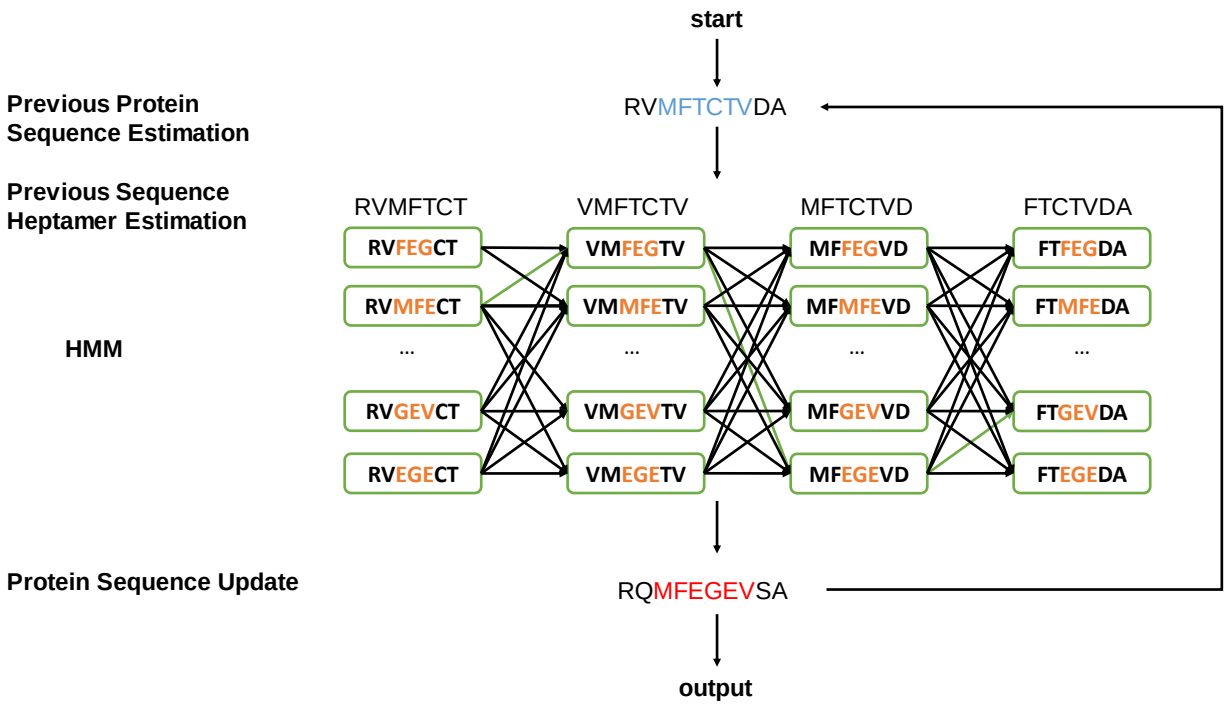


Figure 5.8 Sequence update procedure
The four amino acids on the side of heptamer were fixed for the estimation in the next step. The partial heptamer dataset was used for inference in each update. After a few updates, the sequence output would be used as a decoding sequence.

Instead of searching through the vast hidden states of all 20^7 heptamers by brute force, we initiated the search through only the heavily weighted 20^3 hidden states of the triplets in the centers of the heptamers (Figure 5.8). The probability for all 20^3 triplets was updated using the identity of the amino acids determined from the previous step. For $\psi_t(K)$, we only need to calculate a 1 x 20^3 vector instead of a 1 x 20^7 vector. At time t, the roughly estimated heptamer for blockade level B_t was $X_{-3}X_{-2}X_{-1}X_0X_1X_2X_3$. At the update step, we only calculated $\psi_t(K)$ for heptamer $X_{-3}X_{-2}X'_{-1}X'_0X'_1X_2X_3$ with the center triplets $X'_{-1}X'_0X'_1$ as variables

$$\psi_t(K) = \psi_t(X_{-3}X_{-2}X'_{-1}X'_0X'_1X_2X_3) \propto exp(-\frac{(R(K) - B_t)^2}{2\sigma^2})$$

Where R(K) is the reference blockade level for heptamer $X_{-3}X_{-2}X'_{-1}X'_{0}X'_{1}X_{2}X_{3}$ and σ is the noise level used in our calculations.

For $\Psi(J,K)$, the allowed transition from heptamer K ($X_{-3a}X_{-2a}X_{-1a}X_{0a}X_{1a}X_{2a}X_{3a}$) to J ($X_{-3b}X_{-2b}X_{-1b}X_{0b}X_{1b}X_{2b}X_{3b}$) only requires $X_{0a}X_{1a} = X_{-1b}X_{0b}$ in Pseudo-Heptamer decoding algorithms. It is defined as

$$\Psi(J,K) = \begin{cases} 1/20, & X_{0a} = X_{-1b} \text{ and } X_{1a} = X_{0b} \\ 0, & otherwise \end{cases}$$

The amino acids X_{-3a}, X_{-2a}, X_{2a}, X_{3a}, X_{-3b}, X_{-2b}, X_{2b} and X_{3b} were determined from the previously estimated sequence. Transition matrix was reduced to 20^3 x 20^3 entries by applying a loose requirement on the transition matrix.

Algorithm 5.1 Pseudo-Heptamer decoding algorithms

Input: blockade level $B_1, B_2, \ldots, B_{t-1}, B_t, B_{t+1}, \ldots, B_T$; Heptamer dataset for $X_{-3}X_{-2}X_{-1}X_0X_1X_2X_3$
The nth estimation of sequence was recorded as $S_n = (A_1^n, \ldots, A_{t-1}^n, A_t^n, A_{t+1}^n, \ldots, A_T^n)$ using amino acids and $P_n = (H_1^n, \ldots, H_{t-1}^n, H_t^n, H_{t+1}^n, \ldots, H_{T-6}^n)$ using heptamer where $H_t^n = A_t^n A_{t+1}^n \ldots A_{t+6}^n$, $\alpha_t^n(K) = p(H_t^n = K | B_{1:t})$ is the filtered marginals given previous message and $\beta_t^n(K) = p(B_{t+1:T} | H_t^n = K)$ is the conditional likelihood of future evidence given the hidden state at time t.

Initialize the estimation amino acid sequence as $S_0 = (V, \ldots V, V, \ldots V)$

For n = 1: preset constant integral

Initialize

$\alpha_1^n(K) = p(H_1^n = K | B_1) = \psi_t(H_1^n = K) = p(H_1^n = A_1^{n-1} A_2^{n-1} A_3^n A_4^n A_5^n A_6^{n-1} A_7^{n-1} | B_1)$ for all combination of $A_3^n A_4^n A_5^n$

For t = 2: T

$$\alpha_t^n(J) = p(H_t^n = A_t^{n-1} A_{t+1}^{n-1} A_{t+2}^n A_{t+3}^n A_{t+4}^n A_{t+5}^{n-1} A_{t+6}^{n-1} | B_{1:t}) = \psi_t(H_t^n = J) \sum_K \alpha_{t-1}^n(J) \Psi(J, K)$$

End

Initialize $\beta_{T-6}^n(K) = p(\phi | H_{T-6}^n = K) = 1$

For t = T – 7: 1

$$\beta_t^n(J) = p(B_{t+1:T} | H_t = A_t^{n-1} A_{t+1}^{n-1} A_{t+2}^n A_{t+3}^n A_{t+4}^n A_{t+5}^{n-1} A_{t+6}^{n-1}) = \sum_K \psi_{t+1}(H_{t+1}^n = K) \beta_{t+1}(K) \Psi(J, K)$$

End

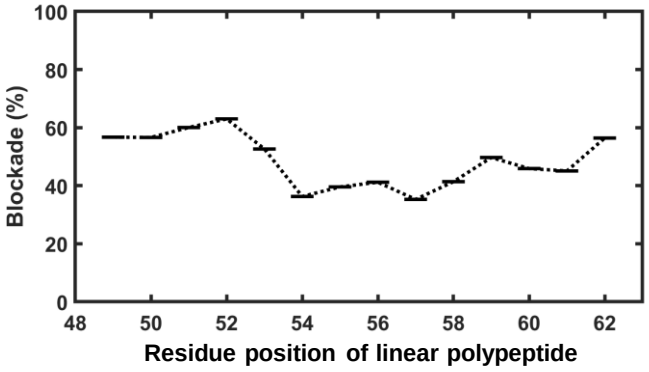
For t = 1: T - 6

$$p(H_t^n = J | B_{1:T}) = \alpha_t^n(J) \beta_t^n(J)$$

Extract center amino acid from H_t^n and generate sequence S_n

End

For the initial estimation, all amino acids in sequence were assumed as valine. Multiple rounds of estimation are applied to precisely determine four amino acids on the edge of heptamer (Figure 5.9). When the side four amino acids could be determined with higher confidence, $\psi_t(K)$ could be determined more accurately. We sampled 5000 protein sequences from protein dataset and tested accuracy from different numbers of updates. In this work, we updated the sequence three times.



Position	53	54	55	56	57	58
Initial	V	V	V	V	V	V
1st update	T	C	T	V	D	A/S
2nd update	E	G	E	V	A/S	A/S
3rd update	E	G	E	V	A/S	A/S
Ground Truth	E	G	E	V	A/S	A/S

Figure 5.9 Sequence determination through multiple updates decoding process
The accuracy of the sequence was improved by having a better estimation of four amino acids on the side of the heptamer.

5.4 Results

Recognizing that even for an atomically thin nanopore, multiple residues along the polypeptide chain contribute to the current blockade, we chose to model the current blockade of amino acid heptamers with the center residue positioned in the center of the nanopore (Figure 5.10a, b). For simplicity, amino acid residues were modeled as a cylinder. To calculate the current blockade using FEA, a 180-mV potential was applied across the *cis* and *trans* chambers of the nanopore containing a solution of 0.4 M KCl. The current blockades of some specific heptamers were determined by numerically solving the PNP equations and other heptamer current blockades were determined by prediction based on those particular heptamers results. The current blockade values of 1.28×10^9 possible heptamer combinations were estimated for a

0.76 nm thick, 0.9 nm diameter nanopore, a 0.76 nm thick, 1.1 nm diameter nanopore and a 1.14 nm thick, 0.9 nm diameter nanopore (Figure 5.10c-e).

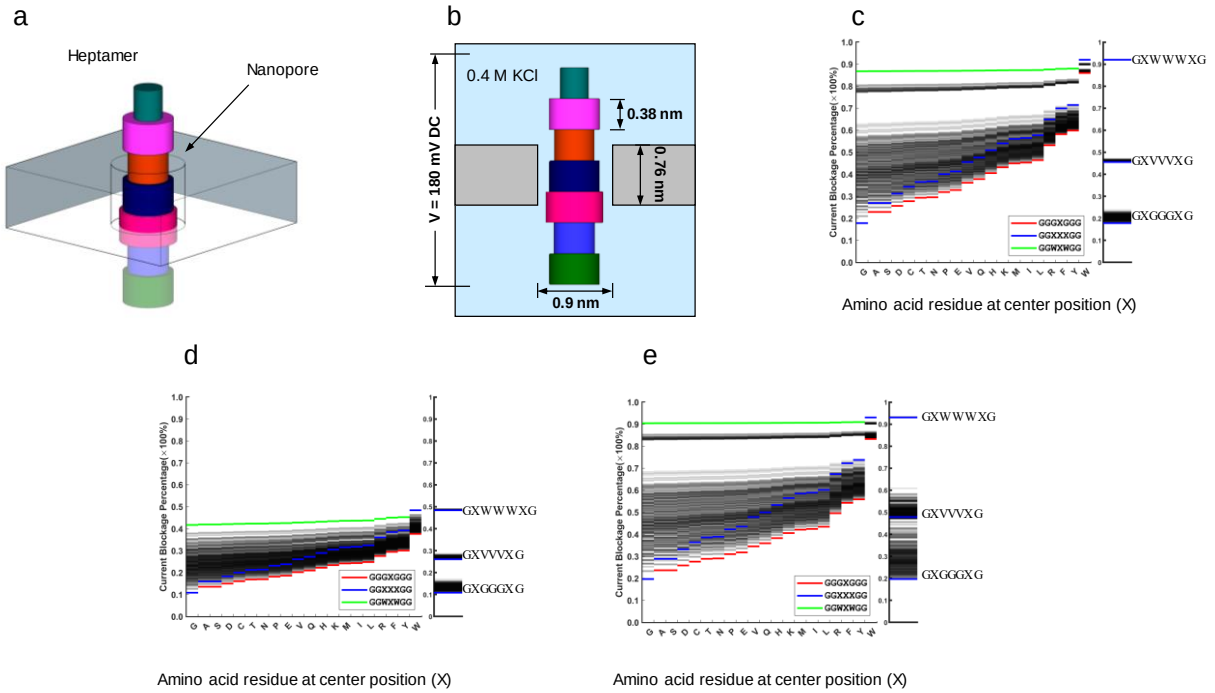


Figure 5.10 Modeling and computation of current blockade of amino acid heptamers **a,** model showed a heptamer in a nanopore; **b,** cross-section of model (a) to compute the current blockade. **c, d, e,** Current blockade of all 8000 heptamers with two Gly's in both sides through a 0.76 nm thick and 0.9 nm diameter nanopore (c), a 0.76 nm thick and 1.1 nm diameter nanopore (d) and a 1.14 nm thick and 0.9 nm diameter nanopore (e). The shaded lines cover current blockades of all 8000 heptamers with two Gly's on each side. The residues on the horizontal axis (X) are the center residues in the heptamer. The lower boundary (red) is the current blockade of GGGXGGG heptamer, while the upper boundary (green) is the current blockade of GGWXWGG heptamer. The horizontal line (blue) is the current blockade of the heptamer with the homomeric triplet (XXX) at the center and two G at both ends. The right bar shows the 400 extended heptamer data for corresponding triplets in the middle and a Gly in both ends.

As expected, the current blockade of any given amino acid in the center of the nanopore does not have a discrete value, but is highly influenced by its six nearest residues (Figure 5.12c-e). The distribution of the current blockages of all the 20^6 heptamers with a fixed amino acid residue in the middle is very broad. For heptamer with a smaller residue such as Gly at the

center, the current blockage varies greatly, ranging from 17.73% for the heptamer with two Gly's as neighbors (GGGGGGG) to 86.79% for the heptamer with two Trp's as the neighbors (GGWGWGG) for 0.76 nm thick and 0.9 nm diameter nanopore. As the size of the center residue in the heptamer increases, the distribution becomes narrower. We also noticed that larger residues make a more dominant contribution to the blockade. Due to the more dominant contribution from larger neighboring residues, the distribution is much broader for heptamer with a smaller residue in the center. The distribution of current blockades through the 1.14 nm thick nanopore is similar to that of 0.76 nm thick nanopore. (Figure 5.10c-e).

For heptamer with a small triplet in the middle, it is also highly influenced by amino acids in the side of heptamer. For example, consider the heptamer with GGG in the middle, the current blockage varies from 17.73% for the heptamer with two Gly's as neighbors (GGGGGGG) to 23.71% for the heptamer with two Trp's as the neighbors (GWGGGWG). If the middle triplet is larger, amino acids at two ends of heptamer influence less to the blockade. The influence of another two amino acids on a triplet is shown in table 5.2.

The current flow through an open nanopore (0.9 nm diameter, 0.76 nm thick) is 465.1 pA and the current blockade of amino acid heptamers is in the range of about 17.7% to 92.0%. An alteration of as small as 0.2% of the blockade results in a change of 1.0 pA, which technically can be detected with precision at high bandwidth by commercially available amplifiers such as Axopatch 200B. However, due to the extensive overlaps among the current blockade signals of heptamer, it is not feasible to determine the identity of the heptamer directly from the current blockade profile of the protein.

To investigate the theoretical feasibility of nanopore protein sequencing, we used the current blockade of the heptamers to simulate the current blockade profile for all 63324 proteins

sequences from the proteome database, which consists of no shorter than ten amino acids and contain no unknown amino acids, and investigated the feasibility of decoding the sequences from the profiles using MATLAB (R2021a, The MathWorks). We used the UniProt human proteome database (Release 2018_11) as the reference(59). We used a Hidden Markov model with pseudo-heptamer decoding algorithms to search for possible heptamers (Figure 5.11).

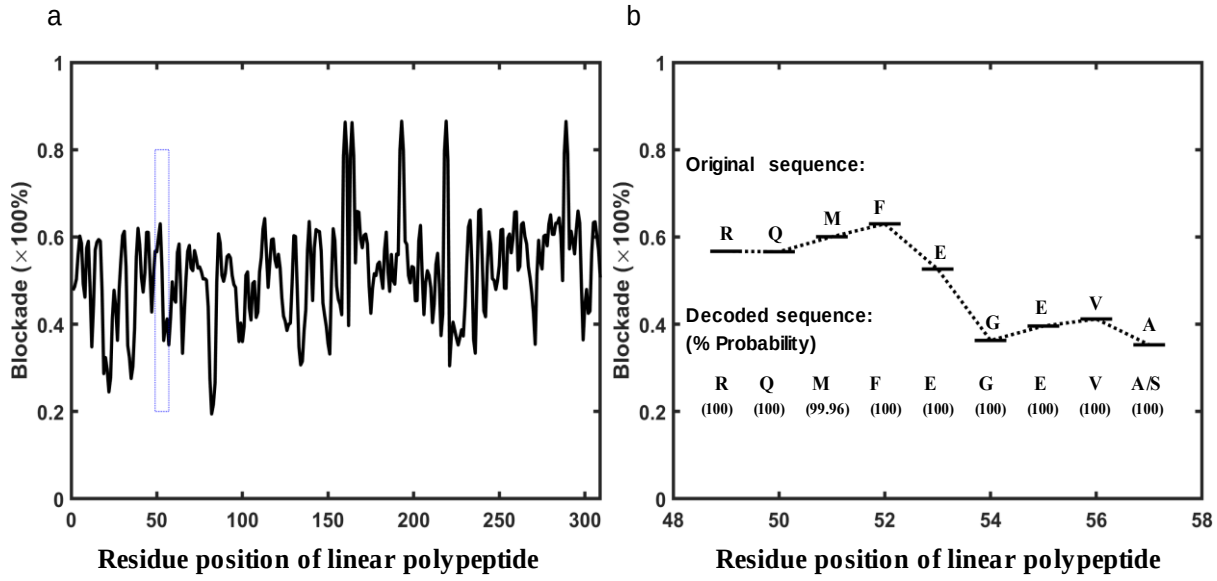


Figure 5.11 Computed current blockade profile and sequence decoding process
a, Current blockades profile of a protein. **b,** sequence decoded from synthetic current profiles with probabilities shown for residue position from 49 to 57.

The accuracy in decoding each amino acid was shown in table 5.5. Twelve amino acids could be decoded with more than 99% accuracy and three amino acids with more than 97% accuracy. If A/S, T/N and M/I are group together, it could also be decoded with more than 99% accuracy.

Table 5.5 Accuracy of decoding all natural amino acids in proteomics database

Original sequence	G	A	S	D	C	T	N	P	E	V	Q	H	K	M	I	L	R	F	Y	W
Decoded sequence (% probability)	G (99.74)	A/S (99.87)		D (99.08)	C (97.83)	T (84.84)	N (71.88)	P (98.98)	E (99.23)	V (99.79)	Q (99.75)	H (99.76)	K (99.90)	M (93.87)	I (98.31)	L (99.91)	R (100)	F (99.98)	Y (99.97)	W (100)
						N (14.81)	T (27.70)							I (5.62)						

5.5 Conclusions and discussion

In summary, we used cylinder models to represent amino acids which enabled us to perform FEA on an ideal nanopore systems. We determined that seven amino acids centered in the nanopore influence the current blockades for a 0.76 nm thick and 0.9 nm diameter nanopore. We first modeled partial of the heptamer dataset and then predicted the whole dataset. The prediction methods could achieve very high accuracy. It indicated that experimental determination of current level for whole heptamer dataset could be feasible. We then generated synthetic current blockades signals for all proteins in human proteome database. A Pseudo-Heptamer decoding algorithms was developed to decode the synthetic signals. Lastly, we concluded that the sequences of the proteins can be decoded with high accuracy using a thin nanopore. Twelve amino acid residues can be identified with greater than 99% accuracy, and three amino acid residues can be identified with greater than 97% accuracy while the other six amino acid residues can be identified as three pairs also with greater than 99% accuracy.

We have demonstrated protein sequence is feasible using a thin nanopore system with electrical measurement. For simplicity, the peptide backbone and the amino acid residues are modeled as a featureless rigid cylinder positioned in the center of the nanopore. In reality, a denatured polypeptide molecule is flexible and the amino acid side chains have specific shape and physical properties such as hydrophobicity and charged characteristics. The polypeptide backbone is unlikely to stay in the center of the nanopore either. These factors undoubtedly influence the current blockade of the heptamers. The effect could result in better resolution or

separation between the amino acids with similar side chain volumes, such as Ala and Ser. The effect could also broaden the current blockade of the heptamers, resulting in more overlap between the current blockade signals.

5.6 Acknowledgement

This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health (5R01GM126013 to X.H.).

This work, in part, is currently being prepared for submission for publication of the material by Wenxu Zhang and Xiaohua Huang. The dissertation author was the primary investigator and author of this material.

Chapter 6 Future work

6.1 Protein fingerprinting with a nanopore

Protein fingerprinting is the first step towards protein sequencing. It might be achieved either by optical or electrical detection.

In the electrical detection, specific amino acids, cysteine for example, on the protein are labeled. Since the effective size of the labeled cysteine is significantly larger than all other unmodified amino acids, the blockade of cysteine is much larger than all other amino acids. The induced current trace could be treated as a binary time serial signal, which contains fingerprinting of specific proteins. The signal could be decoded with properly developed algorithms. In optical detection, we could apply the concept of Förster resonance energy transfer (FRET) to nanopore sequencing. The protein nanopore, labeled with several acceptor fluorophores, is hybridized on the solid-state nanopore by chemical linkage. The targeted protein is denatured, and certain amino acids are labeled with donor fluorophores. When the target protein is translocated through the nanopore, a significant fluorescence signal will be emitted when the labeled amino acid is closed to the pore region and low fluorescence signal will be recorded when no labeled or missed labeled amino acid is translocated through the nanopore. We could record fluorescence signals at the MHz level with APD, which is comparable to protein translocation. It is theoretically feasible to retrieve fluorescence signals which encode the information about labeled amino acids. This information could be used for protein fingerprinting.

6.2 Protein nanopore sequencing

Neither Edman degradation nor mass spectrometry for protein sequencing offers a single-molecule level sensitivity. If nanopore protein sequencing can be achieved, there are significant

advantages over other current protein sequencing technology considering the cost, portability, and single-molecule sensitivity.

However, several challenges need to be tackled before the full experimental realization of nanopore protein sequencing. First, amino acids are varied in charge properties, and uniform translocation of protein molecules through a nanopore under the electric field cannot be simply achieved. Second, no natural motor with the ability to ratchet amino acids one-by-one exists for protein molecules. Third, amino acids are smaller and more diverse compared to nucleotides. It poses challenges to nanopore design and algorithm development. A higher quality nanopore must be proposed to distinguish twenty different amino acids compared to four nucleotides.

Even with many technical challenges, it would revolutionize proteomics research if nanopore protein sequencing could be achieved. Quantitative understanding of biological systems and human diseases is possible with protein sequencing in single-molecule levels.

6.3 Ultra-accurate nanopore DNA sequencing

Accurate of nanopore has rapidly improved over the last ten years by changing the biological nanopore, improving the basecall algorithm, etc. However, it is still suffering from low consensus accuracy and the systematic error could not be eliminated by improving sequencing depth. A better nanopore geometry and a better translocation control unit are possible to solve the problem.

However, the naïve and straightforward method for nanopore sequencing is to sequence DNA directly without the help of helicase or polymerase. The use of helicase and polymerase limits sequencing throughput and introduces translocation instability. The DNA free

translocation speed is 1 μs/base instead of more than 2 ms/base with an enzyme. It is promising to see that throughput could be improved thousands of times higher with the enzyme replaced.

At the same time, the capability to sequence DNA methylation is limited. When the pattern between multiple k-mers is clearly separated, it is much simpler to incorporate DNA methylation and DNA damage analysis.

If all those proposals could be achieved in the future, we will see a portable single-molecule DNA sequencer with unprecedented sequencing speed and capability.

Reference

1. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345-53.

2. Nelson DL, Cox MM, Lehninger AL. Lehninger principles of biochemistry. Seventh edition. ed. New York, NY

Houndmills, Basingstoke: W.H. Freeman and Company ;

Macmillan Higher Education; 2017. xxxiv, 1172, AS34, G20, I45 pages p.

3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-7.

4. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. Science. 2003;299(5607):682-6.

5. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133-8.

6. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20(11):631-56.

7. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.

8. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15(3):201-6.

9. Edman P. A method for the determination of amino acid sequence in peptides. Arch Biochem. 1949;22(3):475.

10. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol. 2016;34(5):518-24.

11. Li J, Stein D, McMullan C, Branton D, Aziz MJ, Golovchenko JA. Ion-beam sculpting at nanometre length scales. Nature. 2001;412(6843):166-9.

12. Storm AJ, Chen JH, Ling XS, Zandbergen HW, Dekker C. Fabrication of solid-state nanopores with single-nanometre precision. Nat Mater. 2003;2(8):537-40.

13. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. P Natl Acad Sci USA. 1996;93(24):13770-3.

14. Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, Pavlenok M, Niederweis M, Gundlach JH. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nat Biotechnol. 2012;30(4):349-53.

15. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4(4):265-70.

16. Akeson M, Branton D, Kasianowicz JJ, Brandin E, Deamer DW. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. Biophys J. 1999;77(6):3227-33.

17. Meller A, Nivon L, Brandin E, Golovchenko J, Branton D. Rapid nanopore discrimination between single polynucleotide molecules. Proc Natl Acad Sci U S A. 2000;97(3):1079-84.

18. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. Nat Biotechnol. 2012;30(4):344-8.

19. Derrington IM, Craig JM, Stava E, Laszlo AH, Ross BC, Brinkerhoff H, Nova IC, Doering K, Tickman BI, Ronaghi M, Mandell JG, Gunderson KL, Gundlach JH. Subangstrom single-molecule measurements of motor proteins using a nanopore. Nat Biotechnol. 2015;33(10):1073-5.

20. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods. 2015;12(4):351-6.

21. Carson S, Wanunu M. Challenges in DNA motion control and sequence readout using nanopore devices. Nanotechnology. 2015;26(7):074004.

22. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12(8):733-5.

23. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018.

24. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouedraogo N, Afrough B, Bah A, Baum JH, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-Sanchez A, Carter LL, Doerrbecker J, Enkirch T, Dorival IGG, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallash E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Trina R, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner D, Pollakis G, Hiscox JA, Matthews DA, O'Shea MK, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, Woelfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keita S, Rambaut A, Formenty P, Gunther S, Carroll MW. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228-32.

25. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, Oliveira G, Robles-Sikisaka R, Rogers TF, Beutler NA, Burton DR, Lewis-Ximenez LL, de Jesus JG, Giovanetti M, Hill SC, Black A, Bedford T, Carroll MW, Nunes M, Alcantara LC, Jr., Sabino EC, Baylis SA, Faria NR, Loose M, Simpson JT, Pybus OG, Andersen KG, Loman NJ. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017;12(6):1261-76.

26. Faria NR, Kraemer MUG, Hill SC, Goes de Jesus J, Aguiar RS, Iani FCM, Xavier J, Quick J, du Plessis L, Dellicour S, Theze J, Carvalho RDO, Baele G, Wu CH, Silveira PP, Arruda MB, Pereira MA, Pereira GC, Lourenco J, Obolski U, Abade L, Vasylyeva TI, Giovanetti M, Yi D, Weiss DJ, Wint GRW, Shearer FM, Funk S, Nikolay B, Fonseca V, Adelino

TER, Oliveira MAA, Silva MVF, Sacchetto L, Figueiredo PO, Rezende IM, Mello EM, Said RFC, Santos DA, Ferraz ML, Brito MG, Santana LF, Menezes MT, Brindeiro RM, Tanuri A, Dos Santos FCP, Cunha MS, Nogueira JS, Rocco IM, da Costa AC, Komninakis SCV, Azevedo V, Chieppe AO, Araujo ESM, Mendonca MCL, Dos Santos CC, Dos Santos CD, Mares-Guia AM, Nogueira RMR, Sequeira PC, Abreu RG, Garcia MHO, Abreu AL, Okumoto O, Kroon EG, de Albuquerque CFC, Lewandowski K, Pullan ST, Carroll M, de Oliveira T, Sabino EC, Souza RP, Suchard MA, Lemey P, Trindade GS, Drumond BP, Filippis AMB, Loman NJ, Cauchemez S, Alcantara LCJ, Pybus OG. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science. 2018;361(6405):894-9.

27. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods. 2017;14(4):411-3.

28. Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, Poon P, Goh YT, Yap PML, Chooi JY, Chng WJ, Ng SB, Thiery A, Goh WSS, Goke J. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat Biotechnol. 2021.

29. Alfaro JA, Bohlander P, Dai M, Filius M, Howard CJ, van Kooten XF, Ohayon S, Pomorski A, Schmid S, Aksimentiev A, Anslyn EV, Bedran G, Cao C, Chinappi M, Coyaud E, Dekker C, Dittmar G, Drachman N, Eelkema R, Goodlett D, Hentz S, Kalathiya U, Kelleher NL, Kelly RT, Kelman Z, Kim SH, Kuster B, Rodriguez-Larrea D, Lindsay S, Maglia G, Marcotte EM, Marino JP, Masselon C, Mayer M, Samaras P, Sarthak K, Sepiashvili L, Stein D, Wanunu M, Wilhelm M, Yin P, Meller A, Joo C. The emerging landscape of single-molecule protein sequencing technologies. Nat Methods. 2021;18(6):604-17.

30. Restrepo-Perez L, Joo C, Dekker C. Paving the way to single-molecule protein sequencing. Nat Nanotechnol. 2018;13(9):786-96.

31. Nivala J, Marks DB, Akeson M. Unfoldase-mediated protein translocation through an alpha-hemolysin nanopore. Nat Biotechnol. 2013;31(3):247-50.

32. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008;26(10):1146-53.

33. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A,

Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338-45.

34. de Lannoy C, de Ridder D, Risse J. The long reads ahead: de novo genome assembly using the MinION. F1000Res. 2017;6:1083.

35. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019;20(1):129.

36. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19(1):90.

37. Sarkozy P, Jobbagy A, Antal P. Calling Homopolymer Stretches from Raw Nanopore Reads by Analyzing k-mer Dwell Times. Embec & Nbc 2017. 2018;65:241-4.

38. Noakes MT, Brinkerhoff H, Laszlo AH, Derrington IM, Langford KW, Mount JW, Bowman JL, Baker KS, Doering KM, Tickman BI, Gundlach JH. Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage. Nat Biotechnol. 2019.

39. Goyal P, Krasteva PV, Van Gerven N, Gubellini F, Van den Broeck I, Troupiotis-Tsailaki A, Jonckheere W, Pehau-Arnaudet G, Pinkner JS, Chapman MR, Hultgren SJ, Howorka S, Fronzes R, Remaut H. Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. Nature. 2014;516(7530):250-3.

40. Oxford Nanopore Technologies. kmer_models [Available from: https://github.com/nanoporetech/kmer_models.

41. Craig JM, Laszlo AH, Brinkerhoff H, Derrington IM, Noakes MT, Nova IC, Tickman BI, Doering K, de Leeuw NF, Gundlach JH. Revealing dynamics of helicase translocation on single-stranded DNA using high-resolution nanopore tweezers. Proc Natl Acad Sci U S A. 2017;114(45):11932-7.

42. Hartel AJW, Ong P, Schroeder I, Giese MH, Shekar S, Clarke OB, Zalk R, Marks AR, Hendrickson WA, Shepard KL. Single-channel recordings of RyR1 at microsecond resolution in CMOS-suspended membranes. Proc Natl Acad Sci U S A. 2018;115(8):E1789-E98.

43. Butler TZ, Pavlenok M, Derrington IM, Niederweis M, Gundlach JH. Single-molecule DNA detection with an engineered MspA protein nanopore. Proc Natl Acad Sci U S A. 2008;105(52):20647-52.

44. Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, Craig JM, Langford KW, Samson JM, Daza R, Doering K, Shendure J, Gundlach JH. Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol. 2014;32(8):829-33.

45. Oxford Nanopore Technologies. [Available from: https://github.com/nanoporetech.

46. Fumagalli L, Esfandiar A, Fabregas R, Hu S, Ares P, Janardanan A, Yang Q, Radha B, Taniguchi T, Watanabe K, Gomila G, Novoselov KS, Geim AK. Anomalously low dielectric constant of confined water. Science. 2018;360(6395):1339-42.

47. Cuervo A, Dans PD, Carrascosa JL, Orozco M, Gomila G, Fumagalli L. Direct measurement of the dielectric polarization properties of DNA. Proc Natl Acad Sci U S A. 2014;111(35):E3624-30.

48. Gopinadhan K, Hu S, Esfandiar A, Lozada-Hidalgo M, Wang FC, Yang Q, Tyurnina AV, Keerthi A, Radha B, Geim AK. Complete steric exclusion of ions and proton transport through confined monolayer water. Science. 2019;363(6423):145-8.

49. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016;537(7620):347-55.

50. Gillet LC, Leitner A, Aebersold R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. Annu Rev Anal Chem. 2016;9:449-72.

51. Barbee KD, Hsiao AP, Roller EE, Huang X. Multiplexed protein detection using antibody-conjugated microbead arrays in a microfabricated electrophoretic device. Lab Chip. 2010;10(22):3084-93.

52. Cohen L, Walt DR. Highly Sensitive and Multiplexed Protein Measurements. Chem Rev. 2019;119(1):293-321.

53. Swaminathan J, Boulgakov AA, Hernandez ET, Bardo AM, Bachman JL, Marotta J, Johnson AM, Anslyn EV, Marcotte EM. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. Nat Biotechnol. 2018.

54. Wang R, Gilboa T, Song J, Huttner D, Grinstaff MW, Meller A. Single-Molecule Discrimination of Labeled DNAs and Polypeptides Using Photoluminescent-Free TiO2 Nanopores. ACS Nano. 2018;12(11):11648-56.

55. Zhao Y, Ashcroft B, Zhang P, Liu H, Sen S, Song W, Im J, Gyarfas B, Manna S, Biswas S, Borges C, Lindsay S. Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. Nat Nanotechnol. 2014;9(6):466-73.

56. UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506-D15.

57. Reynolds JA, Tanford C. Binding of dodecyl sulfate to proteins at high binding ratios. Possible implications for the state of proteins in biological membranes. Proc Natl Acad Sci U S A. 1970;66(3):1002-7.

58. Jones MN. Surfactant Interactions with Biomembranes and Proteins. Chemical Society Reviews. 1992;21(2):127-36.

59. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46(5):2699.

60. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA, Ge Y, Gunawardena J, Hendrickson RC, Hergenrother PJ, Huber CG, Ivanov AR, Jensen ON, Jewett MC, Kelleher NL, Kiessling LL, Krogan NJ, Larsen MR, Loo JA, Loo RRO, Lundberg E, MacCoss MJ, Mallick P, Mootha VK, Mrksich M, Muir TW, Patrie SM, Pesavento JJ, Pitteri SJ, Rodriguez H, Saghatelian A, Sandoval W, Schluter H, Sechi S, Slavoff SA, Smith LM, Snyder MP, Thomas PM, Uhlen M, Van Eyk JE, Vidal M, Walt DR, White FM, Williams ER, Wohlschlager T, Wysocki VH, Yates NA, Young NL, Zhang B. How many human proteoforms are there? Nat Chem Biol. 2018;14(3):206-14.

61. Swaminathan J, Boulgakov AA, Marcotte EM. A theoretical justification for single molecule peptide sequencing. Plos Comput Biol. 2015;11(2):e1004080.

62. Yao Y, Docter M, van Ginkel J, de Ridder D, Joo C. Single-molecule protein sequencing through fingerprinting: computational assessment. Phys Biol. 2015;12(5):055003.

63. van Ginkel J, Filius M, Szczepaniak M, Tulinski P, Meyer AS, Joo C. Single-molecule peptide fingerprinting. Proc Natl Acad Sci U S A. 2018;115(13):3338-43.

64. Ohayon S, Girsault A, Nasser M, Shen-Orr S, Meller A. Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. Plos Comput Biol. 2019;15(5).

65. Walsh MT, Hsiao AP, Lee HS, Liu Z, Huang X. Capture and enumeration of mRNA transcripts from single cells using a microfluidic device. Lab Chip. 2015;15(14):2968-80.

66. Lee HS, Chu WK, Zhang K, Huang X. Microfluidic devices with permeable polymer barriers for capture and transport of biomolecules and cells. Lab Chip. 2013;13(17):3389-97.

67. Qing Y, Ionescu SA, Pulcu GS, Bayley H. Directional control of a processive molecular hopper. Science. 2018;361(6405):908-12.

68. Kwok H, Briggs K, Tabard-Cossa V. Nanopore fabrication by controlled dielectric breakdown. PLoS One. 2014;9(3):e92880.

69. dela Torre R, Larkin J, Singer A, Meller A. Fabrication and characterization of solid-state nanopore arrays for high-throughput DNA sequencing. Nanotechnology. 2012;23(38):385308.

70. Li J, Stein D, McMullan C, Branton D, Aziz MJ, Golovchenko JA. Ion-beam sculpting at nanometre length scales. Nature. 2001;412(6843):166-9.

71. Carter JM, Hussain S. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. Wellcome Open Res. 2017;2:23.

72. Dekker C. Solid-state nanopores. Nat Nanotechnol. 2007;2(4):209-15.

73. Waugh M, Briggs K, Gunn D, Gibeault M, King S, Ingram Q, Jimenez AM, Berryman S, Lomovtsev D, Andrzejewski L, Tabard-Cossa V. Solid-state nanopore fabrication by automated controlled breakdown. Nat Protoc. 2020;15(1):122-43.

74. Anderson BN, Assad ON, Gilboa T, Squires AH, Bar D, Meller A. Probing solid-state nanopores with light for the detection of unlabeled analytes. ACS Nano. 2014;8(11):11836-45.

75. Gilboa T, Zrehen A, Girsault A, Meller A. Optically-Monitored Nanopore Fabrication Using a Focused Laser Beam. Sci Rep. 2018;8(1):9765.

76. Gilboa T, Zvuloni E, Zrehen A, Squires AH, Meller A. Automated, Ultra-Fast Laser-Drilling of Nanometer Scale Pores and Nanopore Arrays in Aqueous Solutions. Advanced Functional Materials. 2020;30(18):1900642.

77. Sanger F, Tuppy H. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J. 1951;49(4):463-81.

78. Sanger F, Tuppy H. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J. 1951;49(4):481-90.

79. Sanger F. Frederick Sanger - Nobel Lecture: The Chemistry of Insulin1958. Available from: http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1958/sanger-lecture.html

80. Lu B, Xu T, Park SK, Yates JR, 3rd. Shotgun protein identification and quantification by mass spectrometry. Methods Mol Biol. 2009;564:261-88.

81. Beck M, Claassen M, Aebersold R. Comprehensive proteomics. Curr Opin Biotechnol. 2011;22(1):3-8.

82. Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem. 2011;80:273-99.

83. Claassen M, Reiter L, Hengartner MO, Buhmann JM, Aebersold R. Generic comparison of protein inference engines. Mol Cell Proteomics. 2011.

84. Liu X, Sirotkin Y, Shen Y, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, Pevzner PA. Protein identification using top-down spectra. Mol Cell Proteomics. 2011.

85. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17(1):239.

86. Boersma AJ, Bayley H. Continuous stochastic detection of amino acid enantiomers with a protein nanopore. Angew Chem Int Ed Engl. 2012;51(38):9606-9.

87. Nivala J, Marks DB, Akeson M. Unfoldase-mediated protein translocation through an alpha-hemolysin nanopore. Nat Biotechnol. 2013.

88. Wilson J, Sloman L, He Z, Aksimentiev A. Graphene Nanopores for Protein Sequencing. Adv Funct Mater. 2016;26(27):4830-8.

89. Aksimentiev A, Schulten K. Imaging alpha-hemolysin with molecular dynamics: ionic conductance, osmotic permeability, and the electrostatic potential map. Biophys J. 2005;88(6):3745-61.

90. Yoo J, Aksimentiev A. Improved Parametrization of Li+, Na+, K+, and Mg2+ Ions for All-Atom Molecular Dynamics Simulations of Nucleic Acid Systems. The Journal of Physical Chemistry Letters. 2012;3(1):45-50.

91. Tian P, Smith GD. Translocation of a polymer chain across a nanopore: A Brownian dynamics simulation study. The Journal of Chemical Physics. 2003;119(21):11475-83.

92. Comer J, Aksimentiev A. Predicting the DNA sequence dependence of nanopore ion current using atomic-resolution Brownian dynamics. J Phys Chem C Nanomater Interfaces. 2012;116(5):3376-93.

93. Cervera J, Schiedt B, Ramírez P. A Poisson/Nernst-Planck model for ionic transport through synthetic conical nanopores. Europhysics Letters (EPL). 2005;71(1):35-41.

94. Kosińska ID, Goychuk I, Kostur M, Schmid G, Hänggi P. Rectification in synthetic conical nanopores: A one-dimensional Poisson-Nernst-Planck model. Physical Review E. 2008;77(3):031131.

95. Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. Structure. 1994;2(7):641-9.

96. Murphy KP. Machine learning : a probabilistic perspective. Cambridge, MA: MIT Press; 2012. xxix, 1067 p. p.