Unidirectional Single-File Transport of Full-Length Proteins Through a Nanopore

Luning Yu
Northeastern University https://orcid.org/0000-0002-7947-1440
Xinqi Kang
Northeastern University
Fanjun Li
University of Massachusetts at Amherst
Behzad Mehrafrooz
University of Illinois at Urbana-Champaign
Amr Makhamreh
Northeastern University
Ali Fallahi
Northeastern University
Aleksei Aksimentiev
University of Illinois at Urbana-Champaign
Min Chen
University of Massachusetts at Amherst
Meni Wanunu (m.wanunu@northeastern.edu)
Northeastern University

Article

Keywords:

Posted Date: May 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1671261/v1

Unidirectional Single-File Transport of Full-Length Proteins Through a Nanopore

Luning Yu,[1] Xinqi Kang,[2] Fanjun Li,[4] Behzad Mehrafrooz,[6] Amr Makhamreh,[2] Ali Fallahi,[2] Aleksei Aksimentiev,[5] Min Chen,[4] Meni Wanunu[1, 2, 3]*

Department of [1]Physics, [2]Bioengineering and [3]Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States

[4]Department of Chemistry, University of Massachusetts at Amherst, Amherst, Massachusetts 01003, United States

[5]Department of Physics and [6]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States

*E-mail: wanunu@neu.edu

Abstract:

Nanopore sequencing is one of only a few methods that can potentially determine the amino acid sequence of individual protein molecules as these are passed through a pore sensor. However, mechanisms for unfolding and translocation of proteins are still unavailable to date. Here we describe a general approach for realizing unidirectional transport of full-length proteins through nanopores. We combine a chemically resistant biological nanopore platform with a high concentration guanidinium chloride buffer to achieve unidirectional, single-file protein transport that is propelled by a giant electro-osmotic effect, as revealed by molecular dynamics simulations and confirmed experimentally. Remarkably, we observed that protein velocities are uniform regardless of the protein sequence, which allows the identification and discrimination among proteins based on their electrical signatures, as well as to distinguish protein signatures by their threading orientation (N-to-C vs. C-to-N terminus). With average transport velocities of 10 μs per amino acid, our method can enable direct, enzyme-free protein fingerprinting and protein sequencing when combined with a high-resolution pore and high-speed nanopore readout.

Introduction

Living systems use complex molecules to encode and carry out their essential functions, with nucleic acids being a central information carrier across generations and within species. A tremendous level of scientific and technological efforts has been made to develop new methods for high-throughput and long-read genomic sequencing[1]. In particular, efforts over the past 20 years have led to major breakthroughs in single-molecule sequencing[2], in which individual DNA molecules are sequenced by either real-time replication monitoring[3] or by passing a DNA strand through a nanopore detector[4,5] .

While DNA serves as a blueprint for the biomolecules that a cell can produce, ultimately, cell phenotypes are difficult to predict from genomic information alone[6,7]. Transcriptome analysis provides insight into the expressed mRNA molecules, although ultimately, the proteome, which is defined as the types and quantities of cellular proteins that are produced, defines a cell's phenotype. There are $>10^4$ types of proteins expressed in a human cell, although proteome complexity extends far beyond the canonical exome: various isoforms[8,9] and post-translational modifications (PTMs)[10,11] flavor a cell's phenotype and determine its functionality. A molecular understanding of these diverse phenotypes requires quantitative methods for protein counting, sequencing, and discrimination among various isoforms and PTMs. The common approaches in mass spectrometry (MS) involve fragmentation, quantification, and model reconstruction. The high sensitivity of MS to minute protein quantities permits single-cell proteomics[12], although MS falls short of providing a complete view of proteomes because low peptide ionization rates and other limitations result in only a few percent sampling efficiencies[13,14]. This unmet need has summoned alternative approaches that can deliver a more complete single-cell proteome without extensive fragmentation.

Nanopores sequencing has recently gained substantial recognition in genomics by enabling ultralong-read of DNA and RNA molecules[15,16]. The basic principle of nanopore sensing involves passing a molecule through a nanometer-scale pore in an impermeable membrane. Although efforts have been made to adapt nanopore sensing principle to protein sequencing[17-22], the complexity coming from the higher order structures, the intramolecular interactions and the at least twenty amino acids (aa's) and their PTMs makes it challenging to decode the sequence information of proteins. Moreover, unlike DNA and RNA which have a uniformly charged phosphate backbone that results in a nearly constant force[23] when voltage is applied across the pore, the peptide backbone is not charged, which means that the transmembrane voltage does not necessarily generate a large enough force to pull the strand taut in the pore for ionic current readout. Voltage-induced[24-27] or temperature-induced[28] unfolding have been demonstrated, although most commonly, chemical denaturation has been the method of choice. To enhance the access of the unfolded protein to the nanopore constriction, sodium dodecyl sulfate (SDS)[18], urea and guanidinium chloride (GdmCl) have been used with α-hemolysin, aerolysin and narrow solid-state pores in the past studies[29-33]. Enzyme-based unfolding and translocation of large proteins has been demonstrated using ClpXP as a motor to unfold and to pull the protein through α-hemolysin[34,35]. Moreover, peptide−DNA conjugates were fed through a pore *via* ratcheting or unwinding through the MspA nanopore using a phi29 DNA polymerase[36] or a Hel308 DNA helicase[37], respectively.

Here we demonstrate an enzyme-free platform for linearization and steady transport of full-length proteins through a nanopore reader. Proved by experimental data and molecular dynamics simulations, protein transports are driven by electroosmotic flow with uniform and slow enough speeds (~10 μs/amino acid), comparing to typical nucleic acid transport times (~1 μs/nucleotide[38]). Opposite signal trends have been shown for N-terminus and C-terminus transport of the same protein. Machine-learning tools can easily classify a protein molecule in a binary mixture on-the-fly with >94% classification accuracy. Signal analysis reveals that the signal properties are reproducible, and further, are sufficient to discriminate one protein from another. With further development, our approach paves the way for single-molecule protein identification and quantification.

Results and Discussion:

Experimental setup for unfolding and transporting full-length protein through a nanopore reader

Specific chemical environments and stronger forces are necessary to linearize and transport proteins through a nanopore. In **Figure 1A** we depict a cut-away view through a flow cell that contains top (*cis*) and bottom (*trans*) buffer reservoirs, each equipped with an electrode. In the polytetrafluoroethylene (PTFE)-based fluidic cell is affixed a silicon (Si) chip that contains a central square opening (~200 μm), atop which a wedge-on-pillar (WOP) membrane support is housed. The WOP support chip comprises a 20-μm-thick SU-8 photoresist layer that houses a 100 μm diameter central aperture (details were published elsewhere[39]). We chose this design to provide a stable membrane support for high voltage tolerance (>350 mV for 100 μm diameter membrane) and long experiments (~15 hours)[40]. The membranes are made by pipetting a lipid or polymer solution in organic solvent over the aperture, followed by repetitive bubbling of air over the film to thin the membrane down to a bilayer, which is determined by monitoring the membrane's electrical capacitance. A side cross-section of the WOP aperture (**Figure 1B**) shows a poly(1,2-butadiene)-b-poly(ethylene oxide) (PBD_n-PEO_m) block-copolymer bilayer that spans the aperture, as well as the inserted α-hemolysin channel, which is the nanopore used in our experiments. We chose this block copolymer membrane material for its chemical compatibility with GdmCl buffers[40]. Also depicted is our use of high GdmCl concentrations to mediate protein unfolding and translocation.

Current vs. voltage curves for single α-hemolysin channels at different buffer conditions used in this study are shown in **Figure 1C**. All curves exhibit significant asymmetry, with higher current amplitudes at positive voltages. The impact of GdmCl on noise in α-hemolysin is moderate: The 10 kHz bandwidth noise values at 300 mV are 7.2 pA for 2.5 M KCl and 10.7 pA for 1 M KCl + 2.0 M GdmCl, respectively. Power spectra at 0 mV and 300 mV for these two buffers (see inset) reveal a small increase in the noise in the low-to-intermediate frequency regime (<5 kHz).

Different types of protein analytes are used to verify the effectiveness of our platform. **Figure 1D** shows the net unfolded protein charge (at pH 7.5)[41] and length (number of aa's) of each protein, as well as pKa-based graphical profiles of the charge[41] and relative volume[42] of each aa residue. We have chosen variants of maltose-binding protein (MBP) in its monomeric form (denoted as either MBP-D10 or D10-MBP depending on the C- or N-terminus attachment of the D_{10} tail) and

its dimeric form (diMBP-D10, with a GGSG linker between two MBP monomers). Another protein we used is green-fluorescent protein (GFP), because its stable β-barrel structure makes this protein notoriously difficult to unfold, and we seek to demonstrate its chemical unfolding and translocation. Anionic D_{10} aspartate tails were added to the proteins in order to direct protein capture, as per an earlier report[26]. The impact of the tail on protein threading is evident in **Figure 1E** and **1F**, where current traces for wild-type MBP (WT-MBP) and MBP-D10 at 175 mV and 1 M KCl, 2 M GdmCl buffer are shown. While in both experiments protein concentration was 350 nM, the capture rates for MBP-D10 (9.3 $s^{-1}\mu M^{-1}$) were ~80% higher than for WT-MBP (5.2 $s^{-1}\mu M^{-1}$), and further, WT-MBP events exhibited a broad range of amplitudes and fast translocations, whereas MBP-D10 events were deep and with most events being uniform in duration, with nearly 79% of the events forming a tight distribution characterized by a ~85% fractional blockade and dwell times between 1 to 10 ms. Such reproducible dwell times and current amplitudes suggests an efficient and deterministic translocation process, mediated by the insertion of the D_{10} tail.

Influence of GdmCl concentration on complete protein unfolding

Bulk measurements suggest that MBP has an unfolding midpoint at 1.0 M GdmCl at room temperature, and further, that MBP is fully denatured in ~1.2 M GdmCl at room temperature[43,44]. We present current recordings of a α-hemolysin channel after the addition of MBP-D10 to the *cis* chamber in 1.0 M, 1.5 M, and 2.0 M GdmCl, respectively. In **Figure 2A**, the current blockade vs. dwell time scatter plots shows two main distributions, highlighted with dashed red and blue dashed circles, in addition to a very “fast” population at ~100 μs which most likely corresponds to protein collisions with the pore entrance. Current traces are shown in **SI Figure S2**, also exhibiting two types of events (long and short), which is consistent to the scatter plots. We attribute the long-spread, long-lived events (red dashed circle) as population P_F, which contains events in which the protein is partly folded, and we ascribe the tight, shorter-lived population P_L (blue dashed circle) to completely linear (or unfolded) proteins, for reasons that are described in the next paragraph. Since these two populations are generally well-resolved, we have quantified the percentage of linear protein events P_L for experiments in various GdmCl concentrations and at different voltages, where P_L is defined as the fraction $P_L = {}^{\#}P_L/({}^{\#}P_F+{}^{\#}P_L)$, ${}^{\#}P_X$ is the number of events in population X (see **SI**, **Figure S15**). As seen in **Figure 2A**, P_L increases with the GdmCl concentration from ~36% at 1.0 M GdmCl to >93% at 2.0 M GdmCl. The existence of one population at 2.0 M GdmCl suggests that the protein is fully unfolded during its translocation through the pore.

Evidence of steady voltage-driven protein translocations

In **Figure 2B**, we present mean dwell times for the P_L populations as a function of voltage for MBP monomer (MBP-D10) and dimer (diMBP-D10), based on the histograms in **SI Figure S9**. Dwell times for population P_L decrease with increasing voltage, showing mild and regular exponential trend, and remarkably, dwell times for the dimeric (di-MBP-D10) are a factor of ~2 longer than for monomeric MBP-D10. Meanwhile, as shown in **SI Figure S10**, the dwell times for P_F population also decrease with increasing voltage from ~1 s to ~10 ms, consistent with translocation. However, dwell times for the partially folded form of protein, P_F, are much longer for both MBP-D10 and diMBP-10, at all voltages, and further, exhibit a steeper voltage dependence than the P_L population. In **Figure 2C,** a related protein “velocity” plot calculated by

dividing the contour lengths (0.34 nm per amino acid in MBP-D10 and diMBP-D10) by the dwell times (from **panel B**) of the proteins, shows that velocity is linearly dependent on voltage for both monomeric and dimeric MBP, but interestingly, independent of the protein length. Electrophoretic mobilities are calculated based on equation

$$\mu = \frac{d}{E \cdot t} \quad , \tag{1}$$

where d is the contour length of the protein molecule and d / t represents field-induced velocity, v_e, of the molecule[45]. To estimate electric field E, we use 5 nm for α-hemolysin as the rough length D of its lumen, therefore we have

$$E = \frac{V}{D} \quad . \tag{2}$$

Combining Equations 1 and 2 we get

$$\mu = \frac{v_e}{V} \times D \quad , \tag{3}$$

where v_e / V represents the slopes of the fitting-lines in **Figure 2C**. Based on that, electrophoretic mobilities for MBP-D10 and diMBP-D10 are calculated as $\mu_m = 8.70 \times 10^{-9}\ cm^2 / V \cdot s$ and $\mu_d = 8.45 \times 10^{-9}\ cm^2 / V \cdot s$, respectively. Datapoints for both MBP-D10 and diMBP-D10 at 175 mV with 2.0 M GdmCl show that higher GdmCl concentrations increase translocating speeds, which suggests that higher GdmCl concentration serves to unfold the protein, as well as enhancing the driving force on the protein chain. The uniform translocation speed, and its direct relationship with voltage, is similar to investigations of ssDNA translocation through α-hemolysin[46-48], although DNA velocities are ~10 times faster than our protein velocities. Finally, a plot of the event rates for these experiments (See **SI**, **Figure S16**) reveals a low-voltage regime characterized by high frequency short-lived collisions and an exponentially increasing capture rate at higher voltages (V > 150 mV), which suggests an entropic barrier for capture[49].

In **Figure 2D** we present dwell-time distributions for GFP-D10 (254 aa), MBP-D10 (389 aa), D10-MBP (389 aa), and diMBP-D10 (764 aa), all at 175 mV applied voltage and with 2 M GdmCl denaturant concentrations. Interestingly, there is no dependence of the protein transport time on its orientation of entry (N-to-C vs. C-to-N terminus), which differs from the orientation dependence of DNA transport through α-hemolysin[50]. The continuous curves in the plots are optimal fits of the dwell-time distributions to the 1D Fokker-Planck equation[51-53], which yields diffusion coefficients D and drift velocities v (shown in the insets for all experiments). Drift velocities for all molecules were in the range of 0.031 – 0.04 nm/μs, which, given the extended backbone distance between amino acids in a protein chain (0.34 nm)[54], translates to a mean residence time of ~10 μs per amino acid in the pore. This mean velocity for a protein chain is roughly an order of magnitude slower than single-stranded DNA transport through α-hemolysin (0.15 nm/μs) [46]. This points to a stark contrast between protein translocation and DNA/RNA translocation, which yields different electromotive forces and pore/polymer interactions. Translocation directions are controlled by tagging the D_{10} tail to N- or C- terminal of MBP protein, which orients threading. In **Figure 2E** we show fractional current blockades vs. dwell times for MBP-D10 (orange dots) and D10-MBP (black dots). To a large extent, there is an overlap in the D10-MBP and MBP-D10 dwell time distributions,

with the exception that the MBP protein with its N-terminus D_{10} tail (D10-MBP) is less effective at being captured by the pore, as indicated by many collisions which exhibit shorter dwell times (~100 μs) and lower current blockades. However, it is noteworthy that protein translocation from either direction proceeds with the same speed, in contrast to transport of DNA[50].

Electroosmotic flow drives protein transport

To determine how Gdm^+ ions enable unidirectional transport of unfolded peptides, we built seven all-atom systems each containing a different 52-residue fragment of the MBP protein (Table S1) threaded through α-hemolysin, a lipid membrane, and 1.5 M GdmCl/1 M KCl electrolyte (see **Figure 3A**). For comparison, two variants of each system were built differing by the composition of the electrolyte solution: 1.5 M GdmCl and 2.5 M KCl. Each system was equilibrated using the all-atom MD method[55] and then simulated under a +200 mV bias for approximately 1,500 ns (see Materials and Methods for details).

In the case of the GdmCl/KCl electrolyte, 94% of the blockade current was carried by Cl^- ions, **Figure 3B**, whereas the current carried by Gdm^+ and K^+ ions was 6 and 0%, respectively. Similarly strong ionic selectivity was observed for pure GdmCl electrolyte, **Figure 3C** and **SI Figure S18**. The ionic selectivity was less pronounced but still substantial (70 and 30% for Cl^- and K^+ currents, respectively) for pure KCl. Consistent with the ion selectivity, we observed strong electro-osmotic effects in all three systems, **Figure 3D, E** and **SI Figure S19**. Further analysis found Gdm^+ ions to accumulate at the inner nanopore surface, in particular, near the termini of the α-hemolysin stem, **Figure 3F** (top) and **SI Figure S20**. In the same regions, individual Gdm^+ ions were observed to remain bound to the nanopore surface for considerable (> 10 ns) intervals of time, **Figure 3F** (bottom). In all system, the local concentrations of ionic species were found to satisfy the local electroneutrality condition, **SI Figure S21**. In the GdmCl/KCl system, however, K^+ ions were almost excluded from the α-hemolysin stem, **SI Figure S21**, which explains their negligible current. Thus, binding of Gdm^+ ions to the inner nanopore surface renders the surface positively charged. That surface charge is compensated by much more mobile chloride ions that carry the majority of the ionic current and produce a strong electro-osmotic effect.

The electro-osmotic effect produced by Gdm^+ binding was found to produce small yet measurable net transport of the unfolded protein through the nanopore. For this analysis, we computed the number of residues translocated through the nanopore constriction as a function of simulation time, **Figure 3G** and **SI Figure S22**. To exclude the effect of peptide chain shrinking or stretching, we next identified the parts of the simulation trajectories where the number of peptide residues within the α-hemolysin stem remained approximately constant, **SI Figure S23**. Averaged over such constant-density trajectory fragments, the peptides were found to move with the average rate of 1.0+/-0.8 and 0.8+/-0.5 residues/μs for pure and mixed GdmCl electrolytes, respectively, and 0.1+/-0.4 residues/microsecond for pure KCl.

Protein-specific current signals

We first analyzed the feasibility of distinguishing N-terminus and C-terminus entry of the MBP protein. In order to extract an “average shape” for C- tagged MBP (MBP-D10) and N- tagged MBP (D10-MBP) events, their barycenters (Fréchet means) were computed using the Soft Dynamic

Time Warping metric[56] (**Supplementary Note 2**). The result of the barycenter computation is a smooth curve, representing the centroid, or the "essence" of the translocation events in the dataset. The barycenters $\vec{x}_{MBPD10}$ and $\vec{x}_{D10MBP}$ are shown in **Figure 4A,** with background traces (black) comprising 100 of their respective resampled events. These curves show a trend of how the events of each protein type tend to progress, on average. Both samples distinctly show opposite locations for local maxima and minima (pink and blue arrows, respectively)**.** Whether due to the proteins secondary structure in the pore, or purely due to sequence variation, the opposing current blockage trends are a strong indication of directional protein translocation.

Finally, to investigate whether the signal properties are reproducible and comprise sufficient information to distinguish different proteins, a support vector machine (SVM) was trained on pure GFP-D10 and MBP-D10 and then used to detect protein type from single-pulse measurements in a mixture experiment (1:1 GFP-D10 to MBP-D10 ratio). **Figure 4B** shows a scatter plot of the fractional blockade versus dwell for the mixture experiments, with the histogram showing partial overlap in the dwell time between the two proteins. To accurately classify the proteins irrespectively of dwell times (which significantly overlap), decision boundaries of the SVM classifier were generated based on 51 signal features from each event, where half of the features are the first 25 Fourier coefficients of the signal and the rest are time-domain features. We used a balanced dataset of 381 GFP-D10 events and 381 MBP-D10 events with dwell times ranging between 500 us and 20 ms for SVM training and testing. The model was trained on 80% of the events and was subsequently tested on the remaining 20% of events. Based on the confusion matrix of the test set (**SI Figure S24**), the model had a classification accuracy of 93.4%, where 70/74 (94.5%) and 73/79 (93.2%) of GFP-D10 and MBP-D10 events were correctly called, respectively (see **SI Section 6** for details). Applying this validated model on 1,142 unlabeled events from the 1:1 mixture experiment, the SVM classified 559 of the events as MBP-D10 (49%) and 583 as GFP-D10 (51%). **Figure 4C** shows the SVM classification results on one of the mixture experiment files, with the confidence score for each call labeled below. Example events from the training sets and mixture classification results are provided in **SI Figure S25**. The obtained ~1:1 event ratio is in excellent agreement with the population sizes obtained from fitting the dwell time histograms of the mixture to the sum of two functions that match the pure GFP-D10 and MBP-D10 dwell time histograms (see **SI**, **Supplementary Note 1** and **Figure S17**), which yields a 53/47 GFP-D10 to MBP-D10 ratio.

Conclusions

We have presented here a method for voltage-driven, single-file, and full-length protein transport through nanopores. Guanidinium ion serves as a dual-purpose agent in this method: first, it unfolds the protein guanidinium linearizes the protein in preparation for transport, and second, they facilitate electroosmosis by binding to the hemolysin pore. Nanopore-based single-molecule protein sequencing, which relies on moving a protein chain through a pore detector, requires keeping an unfolded protein taut at the pore in order to reliably measure a set of amino acids in the chain at any given time. Our findings here present a new paradigm for both unfolding and generating a stretching driving force using a single agent, GdmCl. We envision the use of this effect in future nanopore-based single-molecule protein sequencing applications. For example, high-bandwidth measurements can be combined with GdmCl-mediated protein transport through a higher-resolution pore (such as MspA, which maintains its functionality in 2 M GdmCl[57]) or

asymmetric buffer conditions in which protein unfolding is mediated in one chamber, electro-osmotic forces are used to keep the protein taut at the pore, and enzyme-mediated motion on another chamber of the pore is used to move the protein through the pore in a discrete manner. Moreover, recent work has shown that the use of different anions (such as nitrate) in the buffer has an effect of reducing the noise as compared to chloride[58], which could in principle be used to further improve the signal.

Methods

Polymer bilayer painting and nanopore measurement. The chip with 100 μm SU-8 wedge-on-pillar aperture supported by a 500 μm-thick Si chip[39] was mounted on our custom designed fluidic cell, sealing properly to separate *cis* and *trans* chambers. Both sides of the aperture were pretreated with 4 mg/ml poly(1,2-butadiene)-b-poly(ethylene oxide) (PBD_{11}–PEO_8) block-copolymer (Polymer Source) dissolved in hexane, in order to coat the aperture with a dry and thin polymer layer. The *cis* and *trans* chambers were filled with GdmCl electrolyte (all contain 1 M KCl, 10 mM Tris, pH 7.5), and a pair of Ag/AgCl electrodes were immersed in the electrolyte and connected to an Axon 200B patch-clamp amplifier. Polymer membrane was painted across the aperture using 8 mg/ml polymer dissolved in decane, and at least 60 mins waiting time were required until the polymer membrane thinned to a capacitance value of 60 – 80 pF. After verification of bilayer formation, 0.5 μl of 50 μg/ml α-hemolysin (Sigma-Aldrich) was added to the *cis* chamber and single pore insertion was marked by an ion conductance jump. Denatured protein sample (incubated in GdmCl buffer before use) was added to *cis* chamber and mixed gently by pipetting. Current signals were low-pass filtered at 100 kHz using the Axopatch setting and digitized at 16-bits and 250 kHz sampling rate.

Cloning of the GFP and MBP constructs. The N-terminal 10-aspartate MBP (D10-MBP), C-terminal 10-aspartate MBP (MBP-D10), and C-terminal 10-aspartate GFP (GFP-D10) constructs were obtained by mutagenesis polymerase chain reaction (PCR) using pT7-MBP or pRSETB-GFP as the template plasmid. All primers (Eurofins MWG Operon) used in this study are listed in Table S2. The PCR reaction mixtures were subjected to DpnI digestion for 3 h at 37 °C to degrade the template plasmids. The digested samples were then transformed to chemically competent *E. coli* DH5α cells. The desired mutant plasmids were isolated from colonies and verified by DNA sequencing.

The C-terminal MBP-D10 dimer construct (diMBP-D10) was generated as follows: the first mutagenesis PCR was performed using pT7-hisMBP as the template to remove the stop codon and add a flexible linker GGSG to the C-terminus of the MBP gene. The PCR products were digested with DpnI and transformed into *E. coli* DH5α cells, which resulted in a plasmid pT7-hisMBPggsg containing the hindIII and SfbI restriction sites right after the GGSG linker gene The second PCR was performed with pT7-MBP as the template to introduce HindIII and SfbI cutting sites at the two ends of the MBP gene, and add a D10 at the c-terminal to the MBP fragment. The PCR products and the plasmid pT7-hisMBPgsgg were digested with HindIII and SfbI and ligated by T4 ligase. The ligated products were transformed to chemically competent *E. coli* DH5α cells. The mutant plasmid pT7-diMBP-D10 was verified by enzyme digestion and DNA sequencing.

Expression and purification of GFP and MBP proteins. GFP and MBP protein variants (**Table S1**) were expressed and purified by using similar protocols. Briefly, plasmids were transformed into chemically competent BL21(DE3) *E. coli* cells. The cells were grown in 1 L of LB medium at 37 °C until the OD600 reached 0.6 and induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside. The temperature was then decreased to 16°C for overnight expression. Cells were harvested by centrifugation at 13000 RPM for 25 min. The cell pellets were used for protein purification or frozen at −20 °C for future use. To purify proteins, cells were resuspended in 50 ml of 50 mM Tris-HCl (pH 8.0), 150 mM NaCl buffer and lysed via sonication. The lysate was centrifuged at 13000 RPM for 25 min. The supernatant was filtered through a 0.22 μm syringe filter (CELLTREAT Scientific Products) and then loaded to a Ni-NTA affinity column (ThermoFisher scientific) equilibrated with buffer 50mM Tris-HCl (pH 8.0), 150 mM NaCl. MBP-D10, D10-MBP and diMBP-D10 were eluted in buffer 50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 150 mM imidazole. GFP-D10 was eluted in buffer 50 mM Tris-HCl (pH 8.0), 150mM NaCl, 20 mM imidazole. After Ni-NTA chromatography, MBP-D10, D10-MBP and GFP-D10 exhibited more than 95% purity on SDS-PAGE while the eluted diMBP-D10 fraction contained multiple low-molecular impurity bands. To remove these impurity proteins, the eluted samples were run on a preparative 12 % SDS-PAGE. The band containing the full-length diMBP-D10 was cut out and the protein was extracted from the gel with buffer 50mM Tris-HCl (pH8.0), 8M urea by incubating the gel and the extraction buffer at room temperature for overnight. Supernatant containing the protein was collected by centrifuging the samples at 13000 RPM for 30 min. Protein concentrations of all samples were determined by A280 with Nanodrop and stored at −80 °C for future use.

MD simulation. All MD simulations were performed using the molecular dynamics program NAMD2[59], a 2 femtosecond integration timestep, periodic boundary conditions, CHARMM36[60] force field, and a custom non-bonded fix (NBFIX) corrections for K, Cl and Gdm ions[61]. SETTLE algorithm[62] was used to maintain covalent bonds to hydrogen atoms in water molecules, whereas RATTLE algorithm[63] maintained all other covalent bonds involving hydrogens. The particle-mesh Ewald[64] method was employed to compute long-range electrostatic interactions over a 1.2 Å grid. All van der Waals and short-range electrostatic interactions were evaluated every time step using a cutoff of 12 Å and a switching distance of 10 Å; Full electrostatics were evaluated every second time step.

The all-atom models of α-hemolysin suspended in a lipid bilayer membrane were built using CHARMM-GUI[65]. The initial structural model of α-hemolysin was taken from the Protein Data Bank (PDB ID: 7AHL)[66]. After adding missing atoms and aligning the primary principal axis of the protein with the *z*-axis, the protein structure was merged with a 15×15 nm^2 patch of a pre-equilibrated 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer. The protein-lipid complex was then solvated in a rectangular volume of ~78,500 pre-equilibrated TIP3P water molecules[67]. Gdm^+, K^+, and Cl^- ions were added at random positions corresponding to target ionic concentrations. Additional charges were introduced to neutralize the system. Each final system was 15×15×18 nm^3 in volume and contained approximately 300,000 atoms. Upon assembly, the systems were initially equilibrated using the default CHARMM-GUI's protocol. Specifically, the systems were subjected to energy minimization for 10,000 steps using the conjugate gradient method. Next, lipid tails and protein side chains were relaxed in a 2.5 ns pre-equilibration simulation that was ran while restraining the protein backbones and lipid head groups. This step

was followed by a 25 ns simulation in the NPT (constant number of particles, pressure, and temperature) ensemble using the Nosé-Hoover Langevin piston pressure control[68]. In all simulations, the temperature was maintained at 298.15 K by coupling all non-hydrogen atoms to a Langevin thermostat with a damping constant of 1 ps^{-1}.

The atomic coordinates of the maltose-binding protein (MBP) were obtained from the Protein Data Bank (entry 1JW4)[69]. The missing hydrogen atoms were added using the psfgen plugin of VMD[70]. The protonation state of each titratable residue was determined using PROPKA[71] according to the experimental pH conditions (7.5 pH). Next, the protein was split into seven peptide fragments producing six 53-residue and one 52-residue peptides. The N-terminal of each peptide was terminated with a neutral acetyl group (ACE patch) whereas the C-terminal was terminated with an N-methyl group (CT3 patch). Each peptide was stretched using constant velocity SMD in vacuum, followed by 5 ns equilibration in a 1.5 M GdmCl solution. During the 150 ps SMD run, the C-terminal of the peptide was kept fixed while the N-terminal was coupled to a dummy particle by means of a harmonic potential (k_{spring} = 7 kcal/(mol $Å^2$)) and the dummy particle was pulled with a constant velocity of 1 Å/ps. At the end of the equilibration step, each peptide fragment had a contour length of approximately 167 Å, ~3.16 Å per residue. Next, we used the phantom-pore method[50] to convert the geometrical shape of the α-hemolysin nanopore to a mathematical surface. To fit the stretched peptide into the α-hemolysin pore, the phantom pore surface was initially made to represent a nanopore that was 1.4 times wider than the pore of α-hemolysin. During a 2 ns simulation, the phantom pore was gradually shrunk to match the shape of the α-hemolysin nanopore while all atoms of the peptide and all ions laying outside of the potential were pushed toward the center of the nanopore using a constant 50 pN force. At the end of the simulation, the each peptide fragment and all guanidinium ions residing within 3 Å of any peptide atom were placed inside the pre-equilibrated α-hemolysin system having the peptide's backbone approximately aligned with the nanopore axis. Prior to the production runs, each system was equilibrated for 10 ns in the NPT ensemble at 1.0 bar and 298.15 K with all C_α atoms of the α-hemolysin protein restrained to the crystallographic coordinates.

All production simulations were carried out in the constant number of particles, volume, and temperature ensemble (NVT) under a constant external electric field applied normal to the membrane, producing a ±200 mV transmembrane bias. To maintain the structural integrity of the nanopore, all C_α atoms of the protein were restrained to same coordinates as in the last frame of the equilibration trajectory using harmonic potentials with spring constants of 1 kcal/(mol $Å^2$). The ionic currents were calculated as described previously[55]. The quantify protein translocation, we defined the number of residues translocated as the number of non-hydrogen backbone atoms passing below the α-hemolysin constriction divided by the total number of non-hydrogen backbone atoms in one residue. The constriction's z-coordinate was defined by the center of mass of the backbone atoms of residues 111, 113, and 147. The concentration profile and guanidinium binding analyses were carried out using in-house VMD scripts. All MD trajectories were visualized using VMD[70].

Data analysis. All data parsing (excluding DTW and SVM) were performed using the Pyth-ion package (https://github.com/wanunulab/Pyth-Ion) and figures were generated using Igor. DTW and SVM analyses were conducted via a Jupyter notebook python script, tslearn[72], SciKit-Learn[73], and a modified version of the PyPore[74] nanopore data analysis library. The Jupyter notebook and

associated files are available on GitHub (https://github.com/wanunulab/protein-gd). A detailed description of the DTW and SVM analyses is provided in **SI Section 6**.

Supporting Information Available

Data analysis details, example datasets, and methods description.

Acknowledgment

We thank Nikolai Slavov for helpful discussions regarding protein sequencing. We acknowledge funding from the National Institutes of Health grant # HG0011087 (MW); GM115442 (MC); the National Science Foundation grant PHY-1430124 (AA). The supercomputer time was provided through the XSEDE allocation grant (MCA05S028) and the Leadership Resource Allocation MCB20012 on Frontera of the Texas Advanced Computing Center.

Figures

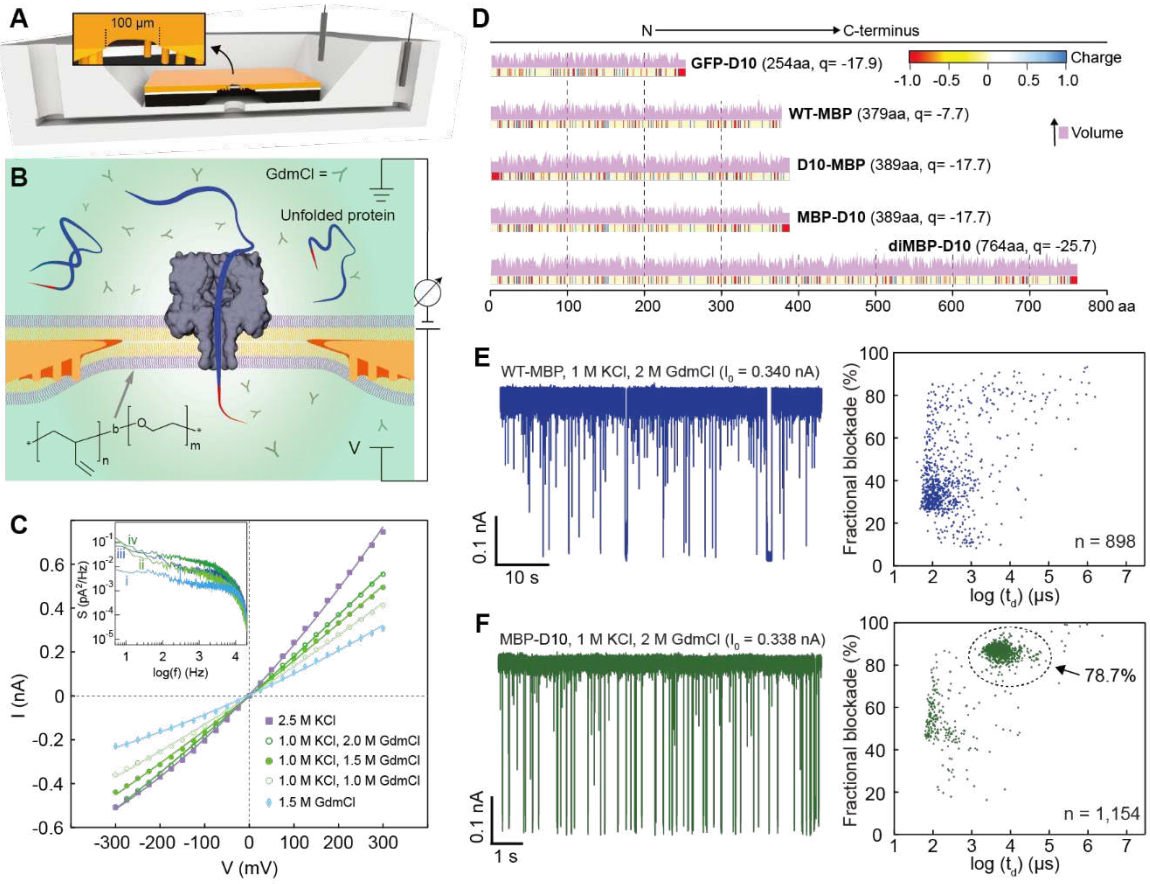

Figure 1. Enzyme-free full-length protein translocation through nanopores. **A)** Cartoon representation of our fluidic cell that houses a Si-chip (black) upon which an SU-8 wedge-on-pillar (WOP) aperture is suspended[39]. **B)** Schematic cut-away view of a PBD_n-PEO_m block-copolymer bilayer (inset shows polymer structure) suspended on the WOP aperture (orange depicts lipid solvent), with an α-hemolysin nanopore inserted into the bilayer. The guanidinium chloride (GdmCl) buffer unfolds the analyte proteins present in the buffer while leaving the α-hemolysin nanopore intact. A pair of Ag/AgCl electrodes generates transmembrane voltage and measures ionic current through the nanopore. **C)** Current-voltage dependence of a single α-hemolysin channel at several buffer conditions (*V* is applied on *trans* chamber). **D)** Graphical representations of the formal charges at pH 7.5 for the protein constructs used in this study in their unfolded (solvent accessible) state. A plot above each charge graph shows the relative amino acid volumes (pink). **E)** Current vs. time trace (left) and the fractional blockade vs. dwell time scatter plot (right) recorded from wild-type maltose binding protein (WT-MBP) at *V* = 175 mV, [WT-MBP] = 0.35 µM. Open pore current (I_o) is indicated above each trace. **F)** Same as in panel E but for MBP containing a C-terminus aspartate tail (MBP-D10), [MBP-D10] = 0.35 µM. Fraction of detectable events in indicated near the dashed circle in the scatter plot.

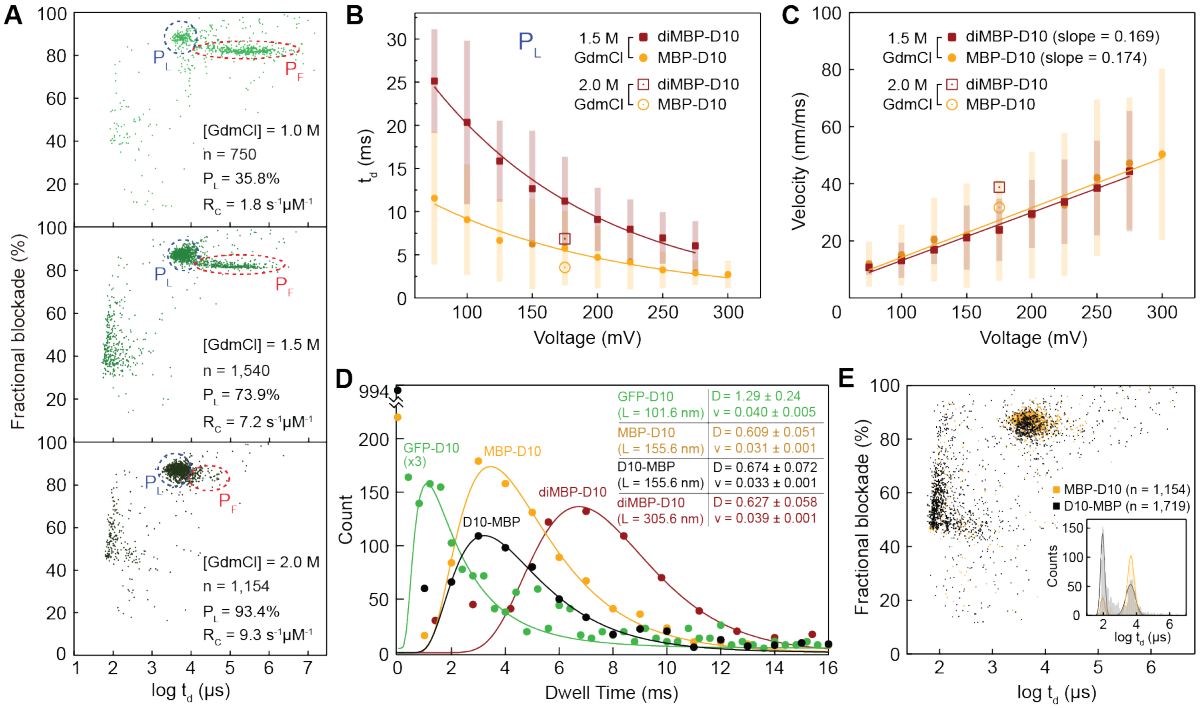

Figure 2. Transport properties of unfolded protein analytes. **A)** Fractional current blockade vs. dwell time scatter plots for MBP-D10 in 1.0 M, 1.5 M and 2.0 M GdmCl buffer (+1 M KCl). Red and blue ovals show populations that correspond to P_F (partially folded) and P_L (linear, or unfolded) states of MBP-D10, respectively. [MBP-D10] = 0.7 μM for 1.0 M GdmCl and 0.35 μM for 1.5, 2.0 M GdmCl experiments. **B)** Mean dwell time vs. voltage with exponential fitting for the P_L populations of MBP-D10 and diMBP-D10, respectively (error bars represent the FWHM of the distribution fits). **C)** Protein transport velocities calculated from estimated protein contour length and observed dwell times as a function of applied voltage (error bars are based on the dwell time distribution widths shown in **panel B**). Buffer conditions for data shown in **panels B and C** are 10 mM Tris, pH 7.5, 1 M KCl and either 1.5 or 2.0 M GdmCl. For the latter, data at only one voltage (175 mV) are shown. **D)** Dwell time histograms for GFP-D10, MBP-D10, D10-MBP and diMBP-D10 along with the mean diffusion coefficients ($nm^2/\mu s$) and velocities (nm/μs) determined from fits to the 1D Fokker-Planck equation[51-53]. **E)** Fractional current blockade vs. dwell time scatter plots and dwell time histograms for C-terminus (MBP-D10) vs. N-terminus (D10-MBP) threading and transport of full-length MBP. Experiments in **panel D and E** were performed in 1 M KCl, 2.0 M GdmCl, 10 mM Tris, pH 7.5, under a 175 mV bias applied to the *trans* chamber.

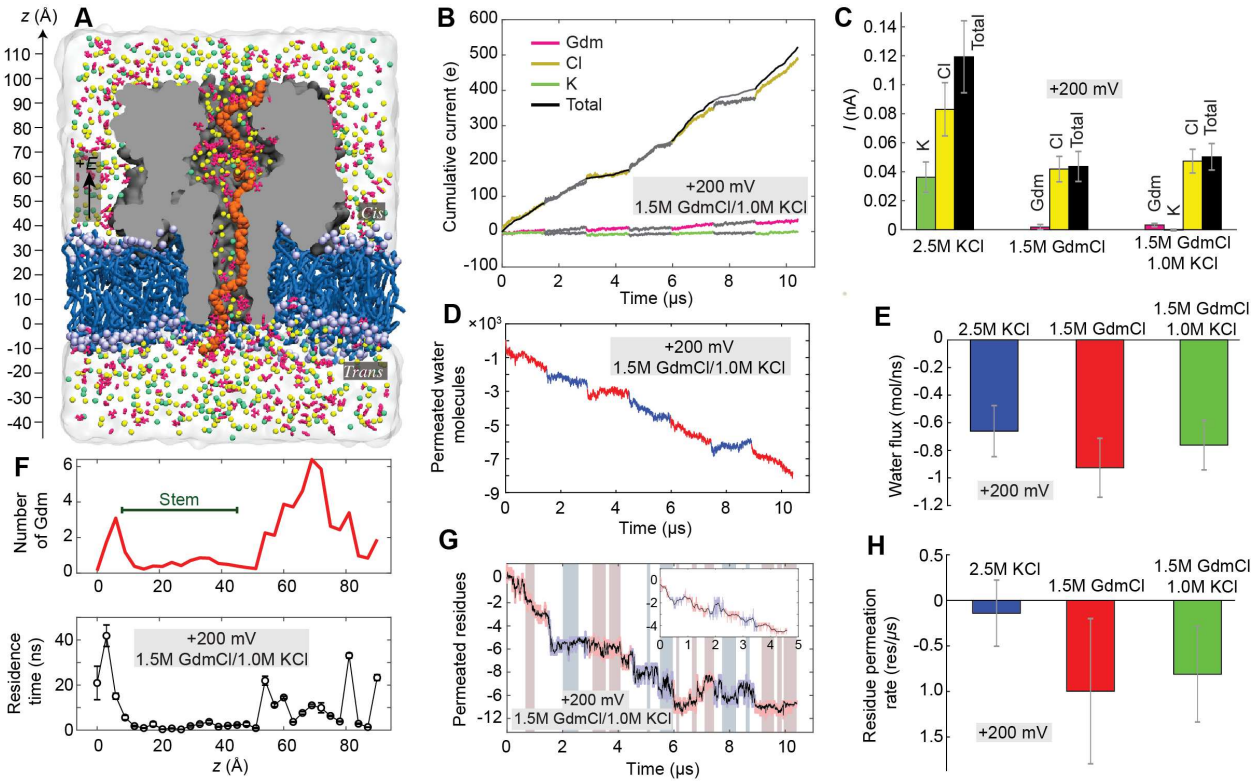

Figure 3. MD simulation of ion, water, and peptide transport through α-hemolysin. **A)** All-atom model of α-hemolysin (gray) containing a fragment of the MBP protein (orange), embedded in a lipid membrane (blue) and submerged in the 1.5 M GdmCl, 1 M KCl electrolyte mixture. **B)** Total charge carried by ion species in seven independent MD simulations differing by the sequence and initial conformation of the MBP fragment. Hereafter each trace is shown using two alternating colors to indicate data from independent trajectories. The traces are added consecutively to appear as a continuous permeation trace. The slope indicates the average current. **C)** Average ionic current for the three electrolyte conditions. Hereafter the average and the standard error are calculated considering each trajectory-averaged value as a result of an independent measurement. **D)** Number of water molecules permeated through the α-hemolysin constriction (residues 111, 113, and 147). Negative values indicate transport in the negative *z*-axis direction (defined in panel A). **E)** Average water flux for each electrolyte condition. **F)** The number of Gdm^+ ions within 3 Å of the nanopore inner surface (top) and their average residence time (bottom) along the transmembrane pore of α-hemolysin. **G)** Number of amino acids permeated through α-hemolysin constriction under a +200 mV bias in the 1.5 M GdmCl/1.0 M KCl electrolyte simulations. Highlights indicate the parts of the trajectories where the peptide density within the stem (8<z<45 Å) of α-hemolysin is constant; the inset shows consecutive addition of the highlighted regions. **H)** Average number of translocated amino acids for each electrolyte condition.

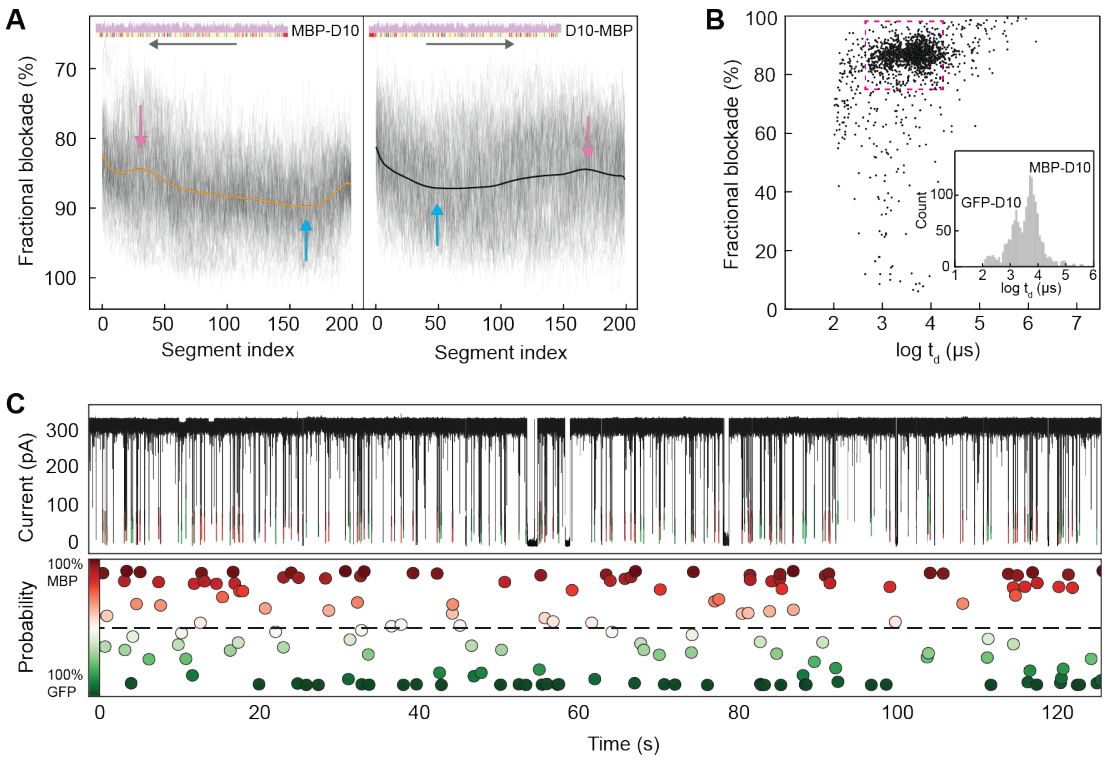

Figure 4. Single-molecule fingerprinting of full-length proteins. **A)** Barycenter of 623 MBP (orange) and 478 D10-MBP (black) translocation events. The background shows 100 time-warped events, generated by segmenting each event to 200 segments and superimposing all event segments in the plot, along with a barycenter line (solid curve). **B)** Fractional current blockade vs. dwell time recorder for an equimolar (0.35 μM each) mixture of MBP-D10 and GFP-D10. The inset shows the distribution of the dwell time. **C)** Post-training SVM classification of the MBP-D10 (red) and GFP-D10 (green) mixture experiment with a probability classification estimate associated with each translocation event (see **SI, Section 6**). All experiments were performed in 1.0 M KCl, 2.0 M GdmCl, 10 mM Tris, pH 7.5 and under a 175 mV bias applied to the *trans* chamber.

References

1 Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345-353, doi:10.1038/nature24286 (2017).

2 Ameur, A., Kloosterman, W. P. & Hestand, M. S. Single-Molecule Sequencing: Towards Clinical Applications. *Trends in Biotechnology* **37**, 72-85, doi:https://doi.org/10.1016/j.tibtech.2018.07.013 (2019).

3 Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133, doi:10.1126/science.1162986 (2009).

4 Venkatesan, B. M. & Bashir, R. Nanopore sensors for nucleic acid analysis. *Nature Nanotechnology* **6**, 615-624, doi:10.1038/nnano.2011.129 (2011).

5 Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* **34**, 518-524, doi:10.1038/nbt.3423 (2016).

6 Jass, J. R. Serrated adenoma of the colorectum and the DNA-methylator phenotype. *Nature Clinical Practice Oncology* **2**, 398-405, doi:10.1038/ncponc0248 (2005).

7 Mohrenweiser, H. W., Wilson, D. M. & Jones, I. M. Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **526**, 93-125, doi:https://doi.org/10.1016/S0027-5107(03)00049-6 (2003).

8 Smith, L. M. *et al.* Proteoform: a single term describing protein complexity. *Nature Methods* **10**, 186-187, doi:10.1038/nmeth.2369 (2013).

9 Bogaert, A., Fernandez, E. & Gevaert, K. N-Terminal Proteoforms in Human Disease. *Trends in Biochemical Sciences* **45**, 308-320, doi:10.1016/j.tibs.2019.12.009 (2020).

10 Tolsma, Thomas O. & Hansen, Jeffrey C. Post-translational modifications and chromatin dynamics. *Essays in Biochemistry* **63**, 89-96, doi:10.1042/ebc20180067 (2019).

11 Conibear, A. C. Deciphering protein post-translational modifications using chemical biology tools. *Nature Reviews Chemistry* **4**, 674-695, doi:10.1038/s41570-020-00223-8 (2020).

12 Slavov, N. Single-cell protein analysis by mass spectrometry. *Current Opinion in Chemical Biology* **60**, 1-9, doi:https://doi.org/10.1016/j.cbpa.2020.04.018 (2021).

13 Specht, H. & Slavov, N. Transformative Opportunities for Single-Cell Proteomics. *J Proteome Res* **17**, 2565-2571, doi:10.1021/acs.jproteome.8b00257 (2018).

14 Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology* **22**, 50, doi:10.1186/s13059-021-02267-5 (2021).

15 Goenka, S. D. *et al.* Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nature Biotechnology*, doi:10.1038/s41587-022-01221-5 (2022).

16 Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods* **18**, 165-169, doi:10.1038/s41592-020-01041-y (2021).

17 Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nature Nanotechnology* **9**, 466-473, doi:10.1038/nnano.2014.54 (2014).

18 Kennedy, E., Dong, Z., Tennant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm3 resolution using a subnanometre-diameter pore. *Nature Nanotechnology* **11**, 968-976, doi:10.1038/nnano.2016.120 (2016).

19 Swaminathan, J. *et al.* Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nature Biotechnology* **36**, 1076-1082, doi:10.1038/nbt.4278 (2018).

20 van Ginkel, J. *et al.* Single-molecule peptide fingerprinting. *Proceedings of the National Academy of Sciences* **115**, 3338-3343, doi:10.1073/pnas.1707207115 (2018).

21 Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. *Nature Nanotechnology* **13**, 786-796, doi:10.1038/s41565-018-0236-6 (2018).

22 Alfaro, J. A. *et al.* The emerging landscape of single-molecule protein sequencing technologies. *Nature Methods* **18**, 604-617, doi:10.1038/s41592-021-01143-1 (2021).

23 Keyser, U. F. *et al.* Direct force measurements on DNA in a solid-state nanopore. *Nature Physics* **2**, 473-477, doi:10.1038/nphys344 (2006).

24 Stefureac, R., Long, Y.-t., Kraatz, H.-B., Howard, P. & Lee, J. S. Transport of α-Helical Peptides through α-Hemolysin and Aerolysin Pores. *Biochemistry* **45**, 9172-9179, doi:10.1021/bi0604835 (2006).

25 Movileanu, L. Squeezing a single polypeptide through a nanopore. *Soft Matter* **4**, 925-931, doi:10.1039/B719850G (2008).

26 Rodriguez-Larrea, D. & Bayley, H. Multistep protein unfolding during nanopore translocation. *Nat Nanotechnol* **8**, 288-295, doi:10.1038/nnano.2013.22 (2013).

27 Rosen, C. B., Bayley, H. & Rodriguez-Larrea, D. Free-energy landscapes of membrane co-translocational protein unfolding. *Commun Biol* **3**, 160, doi:10.1038/s42003-020-0841-4 (2020).

28 Payet, L. *et al.* Thermal unfolding of proteins probed at the single molecule level using nanopores. *Anal Chem* **84**, 4071-4076, doi:10.1021/ac300129e (2012).

29 Oukhaled, G. *et al.* Unfolding of Proteins and Long Transient Conformations Detected by Single Nanopore Recording. *Physical Review Letters* **98**, 158101, doi:10.1103/PhysRevLett.98.158101 (2007).

30 Pastoriza-Gallego, M. *et al.* Dynamics of unfolded protein transport through an aerolysin pore. *J Am Chem Soc* **133**, 2923-2931, doi:10.1021/ja1073245 (2011).

31 Merstorf, C. *et al.* Wild Type, Mutant Protein Unfolding and Phase Transition Detected by Single-Nanopore Recording. *ACS Chemical Biology* **7**, 652-658, doi:10.1021/cb2004737 (2012).

32 Pastoriza-Gallego, M. *et al.* Evidence of Unfolded Protein Translocation through a Protein Nanopore. *ACS Nano* **8**, 11350-11360, doi:10.1021/nn5042398 (2014).

33 Cressiot, B. *et al.* Protein Transport through a Narrow Solid-State Nanopore at High Voltage: Experiments and Theory. *ACS Nano* **6**, 6236-6243, doi:10.1021/nn301672g (2012).

34 Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an alpha-hemolysin nanopore. *Nat Biotechnol* **31**, 247-250, doi:10.1038/nbt.2503 (2013).

35 Nivala, J., Mulroney, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among Protein Variants Using an Unfoldase-Coupled Nanopore. *ACS Nano* **8**, 12365-12375, doi:10.1021/nn5049987 (2014).

36 Yan, S. *et al.* Single Molecule Ratcheting Motion of Peptides in a Mycobacterium smegmatis Porin A (MspA) Nanopore. *Nano Letters*, doi:10.1021/acs.nanolett.1c02371 (2021).

37 Brinkerhoff, H., Kang Albert, S. W., Liu, J., Aksimentiev, A. & Dekker, C. Multiple rereads of single proteins at single–amino acid resolution using nanopores. *Science* **0**, eabl4381, doi:10.1126/science.abl4381.

38 Meller, A. & Branton, D. Single molecule measurements of DNA transport through a nanopore. *Electrophoresis* **23**, 2583-2591, doi:10.1002/1522-2683(200208)23:16<2583::Aid-elps2583>3.0.Co;2-h (2002).

39 Kang, X., Alibakhshi, M. A. & Wanunu, M. One-Pot Species Release and Nanopore Detection in a Voltage-Stable Lipid Bilayer Platform. *Nano Letters* **19**, 9145-9153, doi:10.1021/acs.nanolett.9b04446 (2019).

40 Yu, L. *et al.* Stable polymer bilayers for protein channel recordings at high guanidinium chloride concentrations. *Biophysical Journal* **120**, 1537-1541, doi:https://doi.org/10.1016/j.bpj.2021.02.019 (2021).

41 Haynes, W. M. *CRC Handbook of Chemistry and Physics*. (CRC Press, 2016).

42 Perkins, S. J. Protein volumes and hydration effects. *European Journal of Biochemistry* **157**, 169-180, doi:https://doi.org/10.1111/j.1432-1033.1986.tb09653.x (1986).
43 Liu, G. P., Topping, T. B., Cover, W. H. & Randall, L. L. Retardation of folding as a possible means of suppression of a mutation in the leader sequence of an exported protein. *Journal of Biological Chemistry* **263**, 14790-14793, doi:https://doi.org/10.1016/S0021-9258(18)68107-4 (1988).
44 Sheshadri, S., Lingaraju, G. M. & Varadarajan, R. Denaturant mediated unfolding of both native and molten globule states of maltose binding protein are accompanied by large deltaCp's. *Protein Sci* **8**, 1689-1695, doi:10.1110/ps.8.8.1689 (1999).
45 Nakane, J., Akeson, M. & Marziali, A. Evaluation of nanopores as candidates for electronic analyte detection. *ELECTROPHORESIS* **23**, 2592-2601, doi:https://doi.org/10.1002/1522-2683(200208)23:16<2592::AID-ELPS2592>3.0.CO;2-L (2002).
46 Meller, A., Nivon, L. & Branton, D. Voltage-Driven DNA Translocations through a Nanopore. *Physical Review Letters* **86**, 3435-3438, doi:10.1103/PhysRevLett.86.3435 (2001).
47 Meller, A. & Branton, D. Single molecule measurements of DNA transport through a nanopore. *ELECTROPHORESIS* **23**, 2583-2591, doi:https://doi.org/10.1002/1522-2683(200208)23:16<2583::AID-ELPS2583>3.0.CO;2-H (2002).
48 Hornblower, B. *et al.* Single-molecule analysis of DNA-protein complexes using nanopores. *Nature Methods* **4**, 315-317, doi:10.1038/nmeth1021 (2007).
49 Henrickson, S. E., Misakian, M., Robertson, B. & Kasianowicz, J. J. Driven DNA Transport into an Asymmetric Nanometer-Scale Pore. *Physical Review Letters* **85**, 3057-3060, doi:10.1103/PhysRevLett.85.3057 (2000).
50 Mathé, J., Aksimentiev, A., Nelson, D. R., Schulten, K. & Meller, A. Orientation discrimination of single-stranded DNA inside the α-hemolysin membrane channel. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12377-12382, doi:10.1073/pnas.0502947102 (2005).
51 Muthukumar, M. Polymer translocation through a hole. *The Journal of Chemical Physics* **111**, 10371-10374, doi:10.1063/1.480386 (1999).
52 Ammenti, A., Cecconi, F., Marini Bettolo Marconi, U. & Vulpiani, A. A Statistical Model for Translocation of Structured Polypeptide Chains through Nanopores. *The Journal of Physical Chemistry B* **113**, 10348-10356, doi:10.1021/jp900947f (2009).
53 Ling, D. Y. & Ling, X. S. On the distribution of DNA translocation times in solid-state nanopores: an analysis using Schrödinger's first-passage-time theory. *Journal of Physics: Condensed Matter* **25**, 375102 (2013).
54 Yang, G. *et al.* Solid-state synthesis and mechanical unfolding of polymers of T4 lysozyme. *Proceedings of the National Academy of Sciences* **97**, 139, doi:10.1073/pnas.97.1.139 (2000).
55 Aksimentiev, A. & Schulten, K. Imaging α-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map. *Biophysical Journal* **88**, 3745-3761, doi:https://doi.org/10.1529/biophysj.104.058727 (2005).
56 Cuturi, M. & Blondel, M. in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70 (eds Precup Doina & Teh Yee Whye) 894--903 (PMLR, Proceedings of Machine Learning Research, 2017).
57 Pavlenok, M., Yu, L., Herrmann, D., Wanunu, M. & Niederweis, M. Control of subunit stoichiometry in single-chain MspA nanopores. *Biophysical Journal*, doi:https://doi.org/10.1016/j.bpj.2022.01.022 (2022).
58 Ouldali, H. *et al.* Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat Biotechnol* **38**, 176-181, doi:10.1038/s41587-019-0345-2 (2020).
59 Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics* **153**, 044130, doi:10.1063/5.0014475 (2020).

60 Klauda, J. B. *et al.* Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *The Journal of Physical Chemistry B* **114**, 7830-7843, doi:10.1021/jp101759q (2010).
61 Yoo, J. & Aksimentiev, A. New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Physical Chemistry Chemical Physics* **20**, 8432-8449, doi:10.1039/C7CP08185E (2018).
62 Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **13**, 952-962, doi:https://doi.org/10.1002/jcc.540130805 (1992).
63 Andersen, H. C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics* **52**, 24-34, doi:https://doi.org/10.1016/0021-9991(83)90014-1 (1983).
64 Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089-10092, doi:10.1063/1.464397 (1993).
65 Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **29**, 1859-1865, doi:https://doi.org/10.1002/jcc.20945 (2008).
66 Song, L. *et al.* Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science* **274**, 1859-1865 (1996).
67 Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **79**, 926-935 (1983).
68 Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **101**, 4177-4189, doi:10.1063/1.467468 (1994).
69 Duan, X. & Quiocho, F. A. Structural Evidence for a Dominant Role of Nonpolar Interactions in the Binding of a Transport/Chemosensory Receptor to Its Highly Polar Ligands. *Biochemistry* **41**, 706-712, doi:10.1021/bi015784n (2002).
70 Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33-38, doi:https://doi.org/10.1016/0263-7855(96)00018-5 (1996).
71 Li, H., Robertson, A. D. & Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Structure, Function, and Bioinformatics* **61**, 704-721, doi:https://doi.org/10.1002/prot.20660 (2005).
72 Tavenard, R. *et al.* Tslearn, A Machine Learning Toolkit for Time Series Data. *J. Mach. Learn. Res.* **21**, 1-6 (2020).
73 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).
74 Schreiber, J. & Karplus, K. Analysis of nanopore data using hidden Markov models. *Bioinformatics* **31**, 1897-1903, doi:10.1093/bioinformatics/btv046 (2015).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- YuetalSI2022Final.pdf