UNIVERSITY OF CALIFORNIA
SANTA CRUZ

UNFOLDASE-MEDIATED PROTEIN TRANSLOCATION THROUGH A NANOPORE

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

DOCTOR OF PHILOSOPHY

in

CHEMISTRY

by

Jeffrey M. Nivala

December 2014

The Dissertation of Jeffrey M. Nivala
is approved:

_____
Professor David Deamer, chair

_____
Professor Michael Stone

_____
Professor Mark Akeson

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABSTRACT

Unfoldase-mediated protein translocation through a nanopore

by

Jeffrey M. Nivala

Understanding the operating principles of life requires complete characterization of cellular biology at the molecular level. While genomic analysis illuminates the blueprints used by organisms to store and propagate information, proteins are the principal active ingredients in the recipe of life. Thus, our ability to perceive biological processes hinges on describing the structure and function of proteomes—robust methods to identify and characterize proteins are vital to this effort. The primary focus of this dissertation is to develop a new method of protein analysis by coupling a protein unfoldase to a nanopore sensor. In this system, intact protein strands are interrogated as they are enzymatically translocated through the sensitive nanopore lumen. This process results in a series of ionic current blockades that are diagnostic of protein structure at the single-molecule level. This work represents the first steps towards developing the principles of this technology as a general platform for protein identification. This analytical approach is aimed at achieving the resolution required to fully grasp the complexities of the proteome.

ACKNOWLEDGMENTS

I will begin by thanking my advisor, Prof. Mark Akeson. Mark gave me the freedom to pursue this project, and never hesitated to provide me with the support I needed to succeed. He taught me many things about being a scientist that I never could have learned in a classroom or at the lab bench.

I would like to thank Prof. David Deamer for his encouraging vision and serving as chair of my committee. Dave was always available for advice and consistently inspired me to see the bigger picture. I would also like to thank Prof. Michael Stone for taking an interest in my work and being there for me whenever I needed guidance. Michael always knows the critical questions to ask.

I am grateful to the entire Nanopore group. I would especially like to thank David Bernick and Hugh Olsen. Both have been true mentors and friends to me. Doug Marks for breaking ground with me on this project. Robin Abu-Shumays for being the perpetual lab sunshine—emanating positive vibes and optimistic thinking. Kate Lieberman for sharing her wisdom and always having an open door for me. And all my lab comrades: Andrew Smith (thanks for joining me in the Wild West), Miten Jain, Logan Mulroney, Joey DαHL, Art Rand, and numerous (or notorious) undergrads.

Additional co-authors also contributed to work in Chapter 2 (Mark Akeson & Doug Marks), which was previously published (*Nat. Biotechnol.* 31, 247–250 (2013)), and Chapter 3 (Mark Akeson, Logan Mulroney, Gabriel Li, and Jacob Schrieber), which is currently in review.

Finally, I would not be where I am today without the unconditional support and love of my parents. My crazy brothers. And my best friend and wife, Kristen.

Chapter 1

INTRODUCTION

The ~20,000 genes encoded within the human genome are differentially expressed, translated, and modified to produce a bewildering multiplicity of protein variations approaching the millions. This proteomic complexity is highly variable between cell types, time, and disease states[1]. Genetic mutations implicated in disease ultimately manifest themselves in protein aberrations, while drugs rely on targeting specific proteins for efficacy. Thus, understanding the assortment of proteins integral to complex biological systems is dependent on our ability to accurately identify and characterize them.

Compared to genome and transcriptome analysis, the proteome represents a much more formidable challenge[2]. Proteins are the terminal products of gene expression, and are not amenable to the next-generation sequencing technologies that have revolutionized genomic and transcriptomic analysis. Protein mass spectrometry (MS),

immuno-staining, chromatography, and gel-based separation methods are the most widely used techniques to identify and quantitate proteins from complex samples[3-7]. However, even with recent advances that enable identification of some posttranslational modifications (PTMs), splicing variants, and partial *de novo* sequencing, each of these technologies has associated limitations[8]. For instance, these techniques are inherently ensemble detection methods. Hundreds to billions of molecules are required for accurate identification of any particular protein, while inter- and intra-protein species heterogeneity and a high dynamic concentration range can obscure or provide erroneous results[4,5,8]. Compounding these problems, most MS applications necessitate the fractionation of proteins into small fragments. This leads to a complex bioinformatics problem that is fraught with potential for inaccuracies when the fragments from heterogeneous samples are computationally stitched back together for protein identification. PTMs (which occur on most human proteins) and isoforms exacerbate this fractionation problem by losing the information encoded within combinatorial modification forms. One possible way to overcome these complications is to develop new technologies capable of analyzing individual intact protein molecules.

A recent White Paper based on a meeting hosted by the National Institute of Standards and Technology (NIST) defined a Life Sciences Grand Challenge for Proteomics Technologies[9]. This report called for the development of technologies that enable "ultra sensitive single-cell and single-molecule analyses of proteins." In

accord with these goals, my dissertation has focused on developing a single-molecule protein analysis method using nanopore sensing technology.

1.1 Nanopore sensors

Wallace H. Coulter first conceived of a method to count and size small particles suspended in solution in the 1940s[10]. This invention (now termed a "Coulter counter") may be considered one of the first predecessors to nanopore-based sensors. Similar in concept to a Coulter counter, nanopore instruments (Figure 1.1) are composed of a small aperture or pore imbedded within an insulating membrane that separates two volumes of conductive solution. An amplifier is used to apply a constant voltage across the membrane and measure ionic current flow through the pore over time. Particles that flow through or block the aperture result in an attenuation of current, enabling electronic particle detection.

Compared to other single-molecule sensing methods (e.g. atomic force microscopy or optical tweezers), nanopores do not require analyte molecules to be attached to the sensor. This simplifies experimental setup and enables rapid probing of many analyte molecules in succession throughout a single experiment.

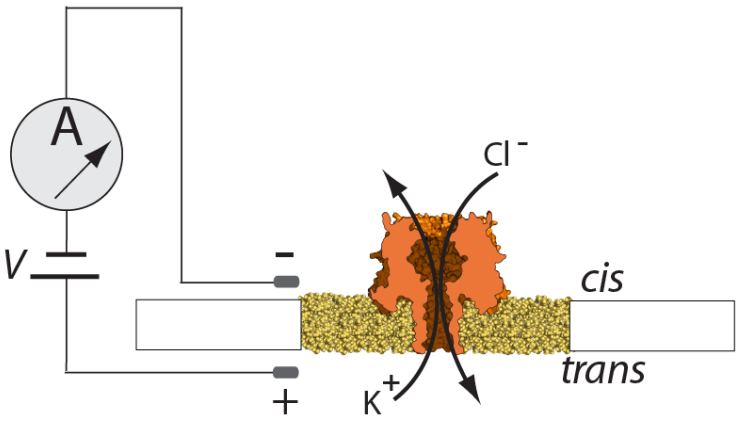


Figure 1.1: Nanopore sensor.
A single nanopore is embedded within an insulating membrane separating two volumes of ionic solution (termed *cis* and *trans*). A constant voltage is applied across the membrane and the resulting ionic current flow (e.g. potassium (K^+) and chloride (Cl^-) ions) through the pore is measured over time. Analytes that pass into the pore vestibule cause a detectable change in the ionic current.

1.1.1 DNA sequencing

Since their inception in the 1990s, nanopores have proven to be powerfully sensitive single-molecule sensors[11]. Single-molecule nanopore analysis has been extensively applied to the study of nucleic acids, particularly in the field of DNA sequencing, where it stands poised as a leading third-generation sequencing technology[12-18].

The promise of personalized genomic medicine has driven the advance of DNA sequencing technologies. 2nd-generation sequencing methods have precipitously dropped the cost of sequencing over the last decade, recently reaching the $1000 human genome milestone. However, some or all of these technologies still have

significant drawbacks including high instrument cost, short read length, and the need for PCR-based sample amplification. Nanopore-based sequencing was originally conceived over 20 years ago by Deamer, Branton, and Church, and has the potential to eliminate these weaknesses[11]. Nanopore sensing relies simply on electrical detection (lowering cost), has theoretically unlimited read length potential, and is inherently single-molecule (not requiring PCR amplification).

With recent technological breakthroughs, the first proof-of-principle studies demonstrating nanopore DNA sequencing have come to light. These advances included the use of a motor (phi29 DNA polymerase) to control DNA movement through the pore[15], and a protein pore (MspA) sensitive enough to achieve single-nucleotide resolution[17]. The UCSC Nanopore group pioneered the use of DNA polymerases in controlling the translocation of DNA through a nanopore[15,19]. This work demonstrated that processive enzymes can be used as motors to finely control polymer translocation, facilitating sequence analysis.

1.1.2 Protein analysis

As nanopore DNA sequencing approaches commercial implementation, nanopore analysis of protein composition and function has also gained momentum in research laboratories. These nanopore protein experiments fall into three main categories: i) experiments that examine protein domains that are captured in the pore lumen but do

not translocate through the pore[20,21]; ii) experiments that report protein activity in bulk phase based on capture and counting of modified substrates[22]; and iii) experiments that examine protein activity at the nanopore using ligands associated with the pore[23,24].

A fourth, relatively new category of nanopore experiments examines protein structure and composition as single polymers are captured and translocated through the pore lumen. Because charge distribution along protein strands is not uniform, translocation cannot be systematically driven by an applied electric field across the nanopore as is the case for polynucleotides. Initiation of protein translocation can be potentiated by attaching a polyanion to the end of the protein strand which is captured in the nanopore electric field[25,26]. Alternatively, solid state nanopores larger than a protein's folded structure can be used to study translocation without unfolding[27,28].

The aforementioned classes of protein experiments, which rely on voltage-mediated forces, have limited control over the protein translocation topology and rate. As such, these methods would not be efficient at unfolding and translocating larger multidomain proteins linearly though a pore, a feature required for sequence analysis.

1.1.3 Types of nanopores

Nanopores can be divided into two categories: biological and solid-state. The most common types of biological nanopores are naturally occurring proteins that serve as

transport channels through lipid bilayers (e.g. pore-forming toxins or small molecule conductors)[12]. More recently, artificial biological nanopores have also been engineered out of DNA (i.e. DNA origami pores)[29]. Solid-state nanopores, on the other hand, can be fabricated out of inorganic materials like silicon-nitride or graphene[30]. In comparing biological and solid-state nanopores, each has associated advantages and weaknesses. Briefly, biological nanopores are atomically precise structures that can be modified at precise locations[31]. That is, individual protein pores are exact replicas of each other and are composed of amino acids that can be mutated to include diverse chemical groups at specific positions within their structure. Solid-state nanopores, however, suffer from structural irreproducibility at the atomic-scale. This causes pore-to-pore variation that can make comparison of results from one pore to another difficult. Benefits of using solid-state pores include enhanced stability (e.g. there is no need for a lipid bilayer) and the ability to easily fabricate nanopore arrays for high-throughput analysis.

1.1.4 Alpha-hemolysin

The protein alpha-hemolysin (αHL) is a pore-forming toxin secreted by the human pathogen *Staphylococcus aureus*[32]. It forms a heptameric pore that is ~10 nm long, and has roughly three limiting constrictions that are ~1.5-2.5 nm in diameter (Figure 1.2). αHL was the first nanopore used to detect DNA translocation[33]. This protein is still the most widely used in contemporary protein nanopore research because of its well-defined three-dimensional structure, minimal tendency to gate, stability, robust

tolerance to mutagenesis, and innate potency to form stable channels in synthetic lipid bilayers.

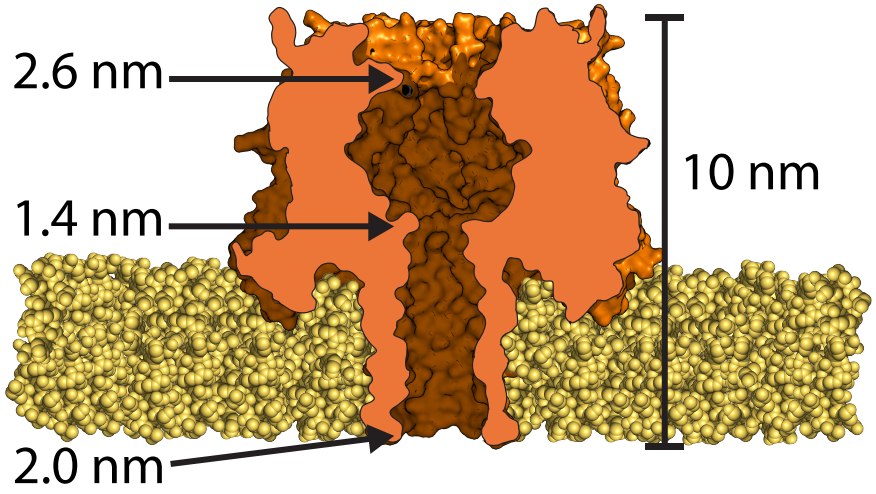


Figure 1.2: Structural dimensions of αHL nanopore.
The αHL pore structure contains several limiting constrictions. This feature makes ionic current flow through the pore sensitive to multiple sites within a polymer's sequence as it translocates.

Although αHL can discriminate between all four canonical DNA bases[34], its long cylinder-like interior lumen make it sensitive to as many as 10 to 15 DNA bases at a time. This relatively large "reading frame" makes strand-based DNA sequencing with αHL difficult. In contrast, the mycobacterial porin MspA has since been developed for DNA sequencing applications because of its more favorable lumen dimensions[35,36]. The MspA structure contains a single limiting constriction that is only sensitive to 3-5 DNA bases at a time. This feature simplifies conversion of ionic current data to nucleotide sequence in the strand-sequencing approach[18].

Despite the MspA nanopore being the current state-of-art in nanopore DNA sequencing applications, I chose to perform all of the experiments contained in this dissertation with αHL. This choice was largely a practical one, as the UCSC Nanopore group had only just begun to work with MspA when I started my experiments. However, it will be interesting to see what advantages MspA provides over αHL for protein sequence analysis as the technology advances.

1.2 Overview of the protein unfoldase ClpX

ClpX is a component of the ClpXP proteasome-like complex that is responsible for the targeted degradation of numerous protein substrates in *Escherichia coli* and other organisms[37]. Within this complex, ClpX forms a homohexameric ring that uses ATP hydrolysis to unfold and translocate proteins through its central pore and into a proteolytic chamber (ClpP) for degradation. The canonical ClpX recognition motif is the 11 amino acid ssrA tag (AANDENYALAA), though other motifs exist[37]. The ssrA tag is added to the C-terminal of nascent proteins on stalled ribosomes via transfer messenger RNA (tmRNA) and an associated protein complex[38]. This releases the bound ribosome and targets the incomplete protein for degradation. Importantly, this prevents aggregation of potentially toxic misfolded protein species. ClpX also has critical roles as a chaperone involved in protein complex disassembly[39].

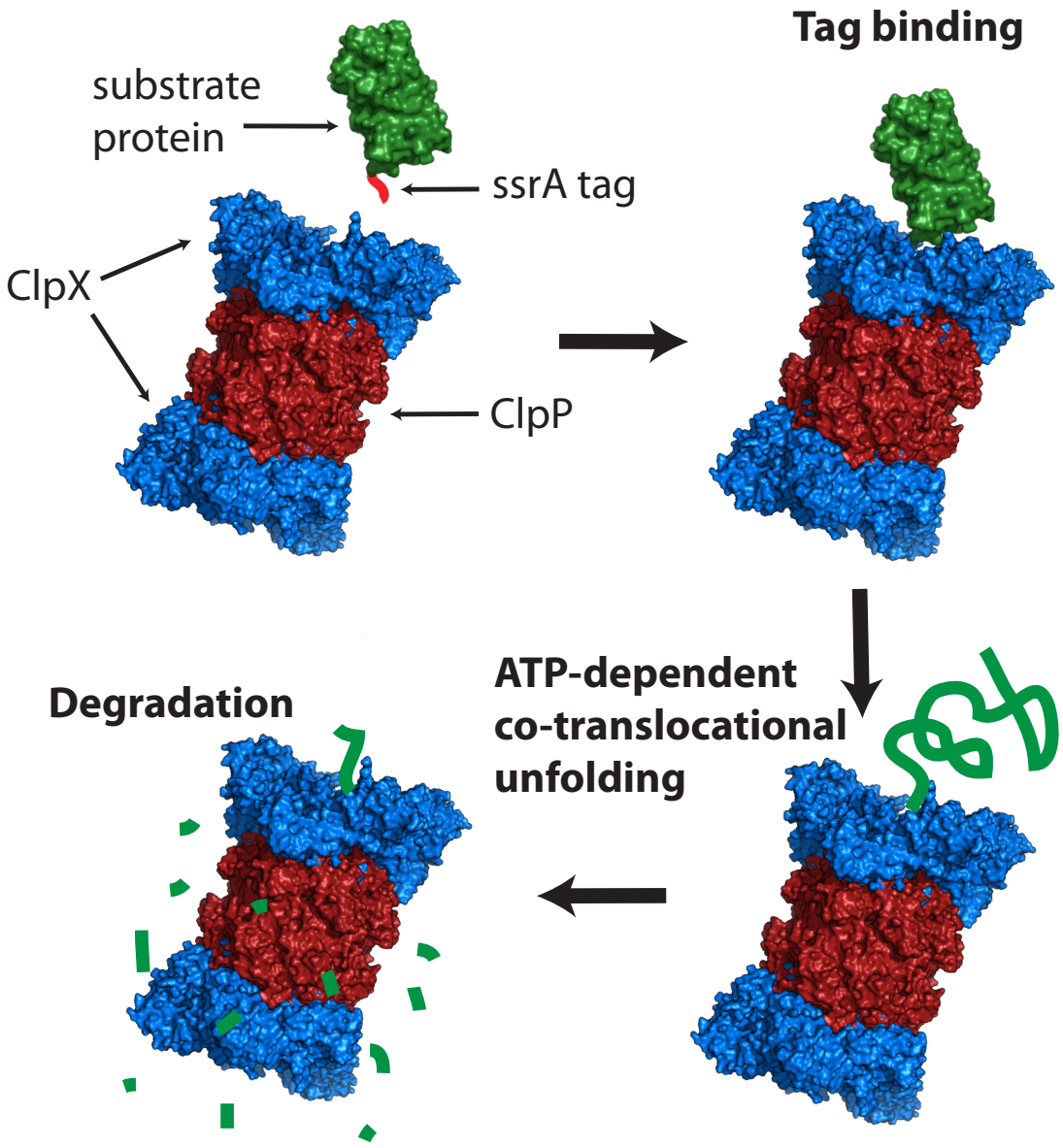


Figure 1.3: ClpXP-mediated protein unfolding and degradation.
ClpX binds specific motifs typically displayed on substrate protein termini (e.g. the C-terminal ssrA tag). After binding the tag, ATP-hydrolysis drives repeated rigid-body movements between the ClpX subunits. These movements pull the substrate protein against the ClpX ring and through its narrow lumen, denaturing the substrate's tertiary structure. As the protein is unfolding, it is translocated in the ClpP lumen where it is digested.

Currently, ClpX is the most well characterized AAA+ protein unfoldase. Though initially discovered nearly 20 years ago, single-molecule studies have only recently shed light on its enzymatic mechanism[40,41]. As shown in Figure 1.3, ClpX unfolds substrate proteins by repeated ATP-fueled mechanical pulling attempts which, when coincident with transient stochastic reductions in substrate structural stability, result in denaturation and translocation of the protein through the enzyme's narrow hexameric ring.

I reasoned that ClpX could be used as a molecular motor to control protein translocation through a nanopore because it generates sufficient mechanical force (~20 pN) to denature stable protein folds, and because it translocates along proteins at a rate suitable for primary sequence analysis by nanopore sensors (up to 80 amino acids per second)[40,41].

1.3 Overview of thesis

The work contained in this dissertation demonstrates that ClpX can be coupled to an αHL nanopore sensor as a motor that can both unfold stable protein domains and translocate them linearly through the pore in an ATP-dependent manner (Chapter 2). Further, I prove this technology can detect subtle modifications in protein sequence that result in structural modifications and altered unfolding pathways, and can ultimately be used to discriminate among such protein variants at the single-molecule

level (Chapter 3). Finally, I discuss tagging of endogenous proteins, barcoding of the tags for sample multiplexing, and high-throughput nanopore analysis using a MinION nanopore array device (Chapter 4). This represents the foundation for a single-molecule protein analysis technique that could complement ensemble methods such as protein mass spectrometry.

Chapter 2

UNFOLDASE-MEDIATED PROTEIN TRANSLOCATION THROUGH AN αHL NANOPORE

2.1 Nanopore experiments

2.1.1 Apparatus

The apparatus used for single-channel nanopore experiments is diagrammed in Figure 2.1. This device (manufactured in-house) consists of two ~100 uL volume wells milled into a small Teflon disc. Each well has two holes: one on the side to allow for Ag/AgCl electrode placement, and one on the bottom for insertion of a U-tube that connects the two wells. The end of the U-tube in the *cis* well contains an ~20 uM aperture fabricated out of Teflon heat-shrink tubing.

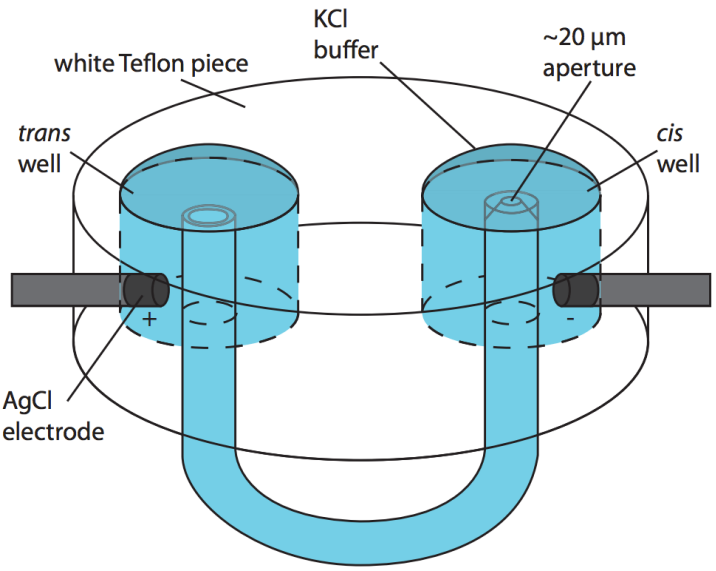


Figure 2.1: Single-channel nanopore apparatus.
A Teflon U-tube connects two ~100 uL volume wells etched into a Teflon disc. In the *cis* well, the end of the U-tube is sealed with an ~20 uM aperture made of Teflon heat-shrink tubing. Ag/AgCl electrodes connected to a patch-clamp amplifier are placed in the wells. A conductive buffer solution in the device completes the circuit. Figure reproduced with permission from Noah Wilson.

A single αHL channel in a lipid bilayer was formed in five steps: 1) The aperture was pre-treated with a solution of diphytanoylphosphatidylcholine (DPhPC) lipid dissolved in hexane. This step serves to clean the aperture and form a thin layer of lipid coating. 2) After pretreatment, the entire apparatus was filled with the buffered ionic solution that was used for the experiment. 3) A small "lipid ball" (~1-2 mm diameter) was formed by mixing ~0.5 uL of hexadecane with several micrograms of dried lipid. The lipid ball was then gently rolled over the aperture and surrounding surface until a visible amount was deposited. This lipid served as the reservoir from which the lipid bilayer was formed. 4) A micropipette with an empty tip was used to

blow an air bubble over the aperture. Once the air bubble was withdrawn back into the tip, it spread the lipid and typically formed a bilayer over the aperture. Stable bilayer formation was evident when no ionic conductance through the aperture occurred under an applied voltage. 5) After establishing a stable bilayer, a small amount (~10-100 ng) of αHL protein was added to the *cis* solution. Spontaneous αHL insertions typically occurred within 10 minutes. Alternatively, insertions were catalyzed by reforming the bilayer.

2.1.2 Electronics and data collection

After setup of the nanopore device and insertion of a single αHL nanopore into the lipid bilayer, ionic current through the nanopore was measured between Ag/AgCl electrodes in series with an integrating patch clamp amplifier (Axopatch 200B, Molecular Devices) in voltage clamp mode with a constant 180 mV potential across the bilayer. Data were recorded at 100 kHz bandwidth in whole cell configuration using an analog-to-digital converter (Molecular Devices), then filtered at 2 kHz using an analog lowpass Bessel filter.

2.1.3 Experimental conditions

Initial experiments, those contained in this chapter, were performed in PD buffer (200 mM KCl, 5 mM $MgCl_2$, 10% glycerol, and 25 mM HEPE-KOH, pH 7.65). However, we found that the longevity of nanopore experiments increased with a buffer containing higher salt. The improved buffer (called protein translocation buffer or PT

buffer) contained 300 mM KCl, 10 mM $MgCl_2$, 10% glycerol, 1 mM DTT, 1 mM EDTA, 5 mM ATP, and 10 mM HEPES-KOH pH 7.6. In addition to increasing experiment longevity, the higher salt concentration also increased the signal to noise ratio of the data. This was the buffer used for experiments contained in Chapter 3, and all subsequent work.

$ClpX_6$ was diluted in buffer for a final concentration of 100 nM. For experiments contained in Chapter 3, two additions were made which also improved the efficiency of data collection: 1) $ClpP_{14}$ was added to a final concentration of 300 nM, and 2) the solution was supplemented by an ATP regeneration mix (8 mM creatine phosphate and 0.08 mg/mL creatine phosphokinase). The ClpX(P/ATP-regeneration) mix solution was used to fill the entire system before isolation of a single αHL nanopore. Upon insertion, the *cis* well was perfused with ~6 mL buffer. Experiments were conducted at 30 °C with ~1 μM substrate protein added to the *cis* well.

2.2. ClpX-mediated protein translocation through αHL

Protein sequencing using nanopores is technically challenging for two reasons: (i) both tertiary and secondary structures must be unfolded to allow the denatured protein to thread through the nanopore sensor with amino acid residues in single-file order; and (ii) processive unidirectional translocation of the denatured polypeptide through the nanopore electric field must be achieved despite a nonuniform charge along the polypeptide chain. To address these challenges, we devised a general method for

enzyme-controlled unfolding and translocation of native proteins through a nanopore sensor using the protein unfoldase ClpX.

2.2.1 S1: a model single domain protein

For our initial experiments, we used a modified version of the ubiquitin-like protein Smt3 as the substrate[42]. Smt3 comprises 98 amino acids arranged into four β-strands and a single α-helix. To facilitate nanopore analysis, we modified the engineered Smt3 protein, designated 'S1', in two ways. First, it was appended with a 65-amino-acid-long glycine/serine tail including 13 interspersed negatively charged aspartate residues (PolyGSD, Appendix A).

This unstructured polyanion was designed to promote capture and retention of S1 in the electric field across the nanopore. Based on its crystal structure[43], the Smt3 folded domain is predicted to sit on top of the αHL vestibule. Second, the appended polyanion was capped at its C terminus with an ssrA tag, the 11-amino-acid ClpX-targeting motif[44]. Experiments conducted in bulk phase confirmed that ClpX unfolds and translocates proteins appended with this unique polyanion tag in an ATP-dependent manner (Figure 2.2). This long polyGSD-ssrA tag also allowed ClpX to specifically bind to the C terminus of the protein when it threaded through the pore into the *trans* compartment (Figure 2.3b and c,i).

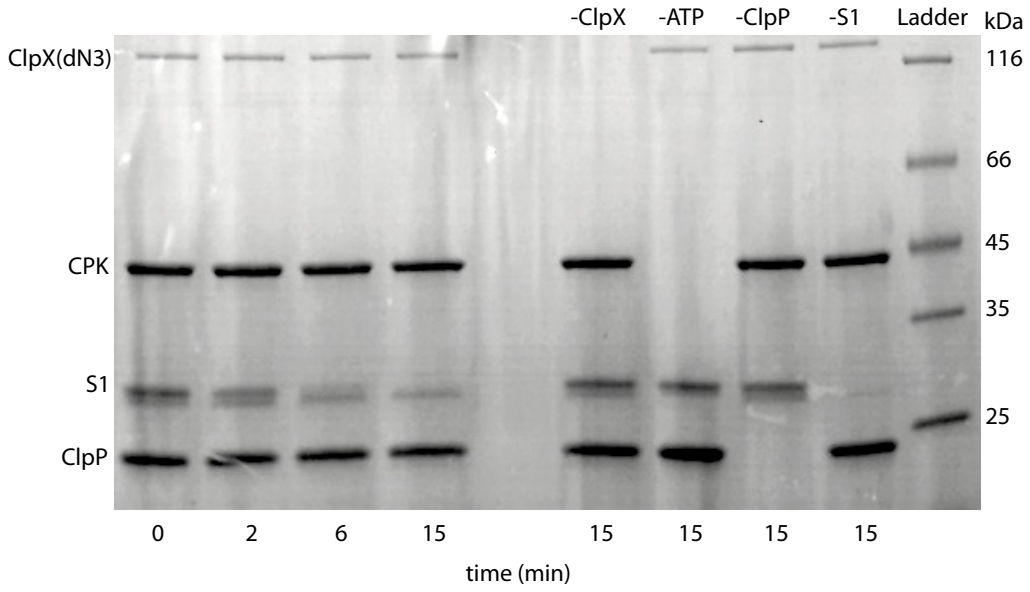


Figure 2.2: Bulk phase assays of ClpX/ATP-dependent unfolding and translocation of substrate proteins bearing the long, charged ssrA-tagged C-terminal tail. SDS/PAGE gel showing substrate protein S1 degradation by ClpXP in the presence of ATP. Lanes 1-4 are a time course of S1 digestion in the presence of ClpX, ClpP, and ATP. Reactions minus ClpX (lane 6), minus ClpP (lane 7), or minus ATP (lane 8) showed no comparable degradation. Lane 9 is absent S1.

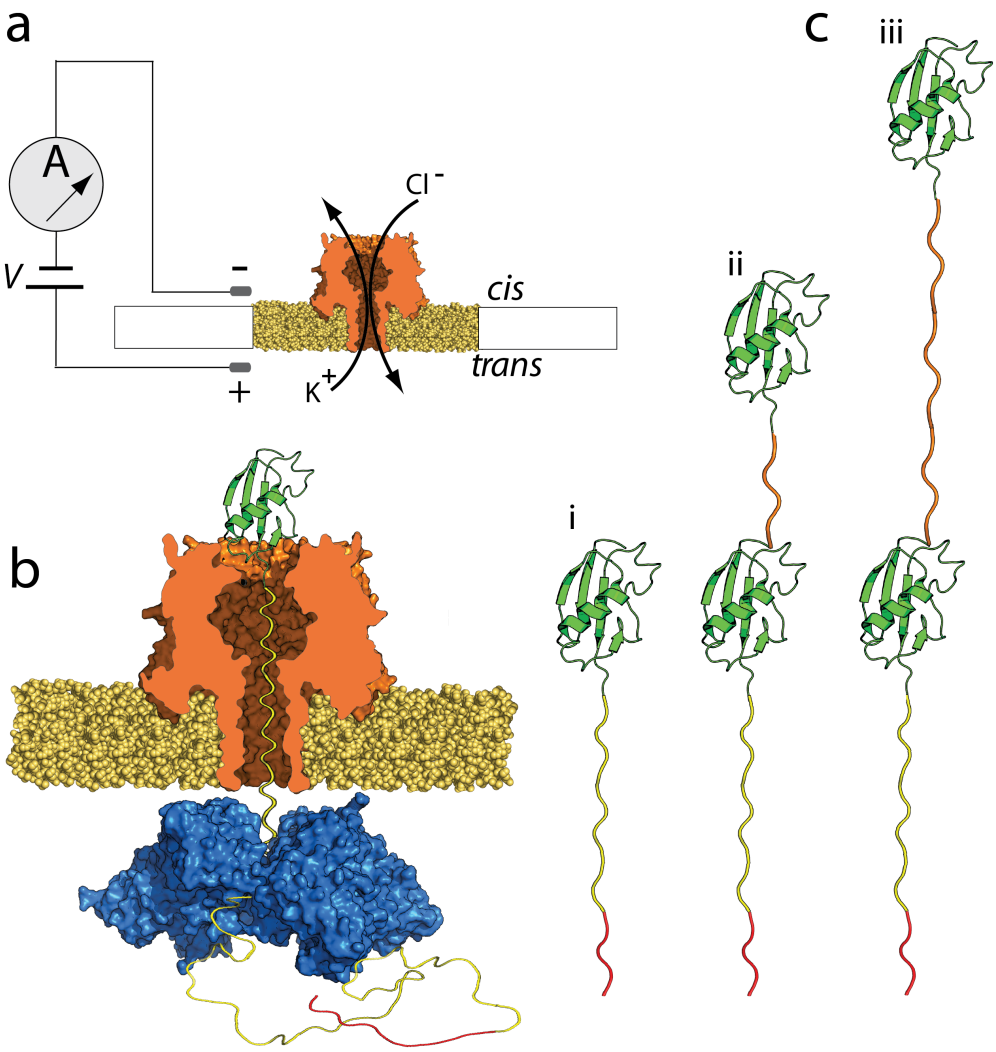


Figure 2.3: Experimental set-up. (**a**) Nanopore sensor. A single αHL pore is embedded in a lipid bilayer separating two polytetrafluoroethylene wells each containing 100 µl of 0.2 M KCl solution at 30 °C. Voltage is applied between the wells (*trans* side +180 mV), causing ionic current flow through the channel. Current diminishes in the presence of a captured protein molecule. (**b**) Protein capture in the nanopore. A model protein bearing an Smt3 domain (green) at its N terminus is coupled to a charged flexible linker (yellow) with an ssrA tag (red) at its C terminus. As a result of the applied voltage, the charged, flexible tag is threaded through the pore into the *trans*-side solution until the folded Smt3 domain prevents complete translocation of the captured protein. ClpX present in the *trans* solution binds the C-terminal ssrA sequence. Fueled by ATP hydrolysis, ClpX translocates along the protein tail toward the channel, and subsequently catalyzes unfolding and

translocation of the Smt3 domain through the pore. (**c**) Engineered proteins used in this study. S1, a protein bearing a single N-terminal Smt3-domain coupled to a 65-amino-acid-long charged flexible segment capped at its carboxy terminus with the 11 amino acid ClpX-targeting domain (PolyGSD, ssrA tag) (**i**); S2-35, similar to S1 but appended at its N terminus by a 35-amino-acid linker and a second Smt3 domain (**ii**); S2-148, identical to S2-35 except for an extended 148-amino-acid linker between the Smt3 domains (**iii**). The linker lengths in this panel are not to scale.

Representative ionic current traces for capture and translocation of protein S1 in the presence of ClpX and ATP are shown in Figure 2.4a and Appendix B. From the open channel current of ~34 ± 2 pA (Figure 2.4a,i), S1 capture resulted in a current drop to ~14 pA (Figure 2.4a,ii). This stable current lasted for tens of seconds and was observed in the presence or absence of ClpX and ATP added to the *trans* compartment. This is consistent with the Smt3 structure held stationary atop the pore vestibule by the electrical force acting on the charged polypeptide tail in the pore electric field. In the presence of ClpX and ATP, this initial ionic current state was often followed by a progressive downward current ramp reaching an average of ~11 pA with a median duration of 4.2 s (Figure 2.4a,iii and Figure 2.5). This current ramp was observed with protein S1 a total of 45 times over ~5.5 h of experimentation when ClpX and ATP were present; in contrast, the ramp was never observed after ionic current state ii (Figure 2.4a,ii) when ClpX or ATP was absent from the *trans* solution over ~2.3 and ~1.7 h of experimentation, respectively. In a majority of events, the ClpX-dependent ramping state terminated with an abrupt ionic current decrease to ~3.9 pA (Figure 2.4a,iv). The median duration for state iv was ~700 ms before it ended in a rapid increase to open the channel current (Figure 2.4a,i′). As an additional

control, we constructed a variant of S1 appended with three additional amino acid residues at the C terminus (protein S1-RQA, Appendix A). Because ClpX recognition of the ssrA tag is dependent upon the C terminus α-carboxyl group, the additional residues placed between the tag sequence and the C terminus thereby inhibit ClpX binding[45]. In the presence of ATP and ClpX, we never observed the ramping state with protein S1-RQA over ~1.7 h of experimentation (data not shown).

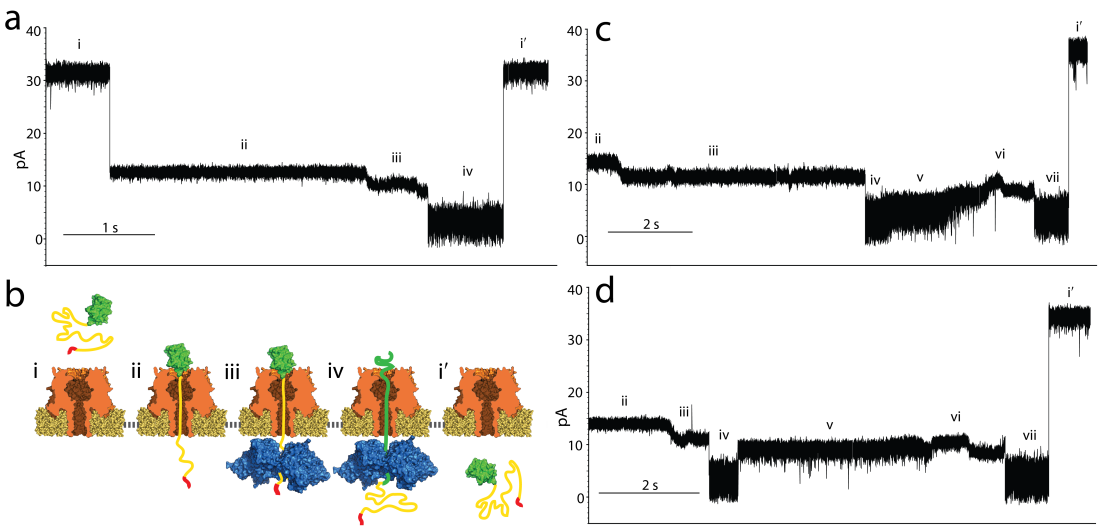


Figure 2.4: Ionic current traces during ClpX-mediated protein translocation. (**a**) S1 translocation. Open channel current through the αHL nanopore under standard conditions (~34 ±2 pA, RMS noise 1.2 ± 0.1 pA) (**i**). Capture of the S1 substrate. Upon protein capture, the ionic current drops to ~14 pA (~0.7 pA RMS noise) (**ii**). ClpX-mediated ramping state. The ionic current decreases to ~10 pA and is characterized by one or more gradual amplitude transitions. This pattern is only observed in the presence of ClpX and ATP (*trans* compartment) (**iii**). Smt3 domain unfolding and translocation through the nanopore (~3.8 pA, 1.7 pA RMS noise) (**iv**). Return to open channel current upon completion of substrate translocation to the *trans* compartment (**i′**). (**b**) Working model of ClpX-mediated translocation of S1. Roman numerals used to label panels correspond to ionic current states in **a**. (**c**) S2-35 translocation. Open channel current (**i**) is not shown. States ii–iv are identical to states

ii-iv in **a**. Gradual increase in ionic current to ~10 pA. In our working model this corresponds to a transition from Smt3 domain translocation to linker region translocation (**v**). A second putative ramping state that closes resembles ramping state iii (**vi**). A second putative Smt3 translocation state with ionic current properties that closely resemble state iv (**vii**). Return to open channel current (**i′**). (**d**) S2-148 translocation. Ionic current states i-iv and vi-i′ were nearly identical to those states for S2-35 translocation in **c**. (**v**) In our working model, this ionic current state corresponds to translocation of the 148-amino-acid linker. Its amplitude is ~3 pA higher than the S2-35 linker amplitude (~9 pA), and it has a median duration ~2.5 fold longer than the comparable S2-35 state v. Translocation events that included ramping state iii were observed 62 times for protein S2-35 (7.3 h of experimentation), and 66 times for protein S2-148 (4.3 h of experimentation), when ClpX and ATP were present. In the absence of ClpX, these ramping states were never observed for S2-35 (1.7 h of experimentation) and S2-148 (1.2 h of experimentation).

Based on these data we hypothesized that ClpX served as a molecular machine that used chemical energy derived from ATP hydrolysis to pull the S1 protein through the nanopore. In the proposed process, an open channel (Figure 2.4b,i) captures protein S1 with the Smt3 segment perched above the pore vestibule with the slender, charged polypeptide tail segment extended into the pore lumen, and the ssrA tag in the *trans* compartment (Figure 2.4b,ii). In this ionic current state, ClpX is not bound to S1 or, alternatively, is bound but is still distant from the pore; ClpX advances along the S1 strand toward the *trans*-side orifice of the αHL pore until it makes contact (Figure 2.4b,iii). At this time, ionic current decreases owing to proximity of ClpX to the pore; under the combined force exerted by ClpX and the pore electric field, the Smt3 structure atop the pore is sequentially denatured, thus allowing the polypeptide to advance through the nanopore (Figure 2.4b,iv). In this state, the ionic current has decreased because larger amino acids (or Smt3 secondary structures) have entered the

pore lumen. This ionic current state persists until the S1 protein is completely pulled into the *trans* compartment resulting in a return to the open channel current (Figure 2.4b,i′).

2.2.2 S2-35 and S2-148: model two domain proteins

This model makes a testable prediction. If the observed states are due to processive movement of polypeptide segments into the pore lumen driven in part by ClpX, then changing the protein primary structure should result in sequential ClpX- and ATP-dependent changes in the ionic current pattern. In particular, addition of a second Smt3 domain should result in a second ramping state (Figure 2.4a,iii) followed by a second Smt3 translocation state centered at ~4 pA (Figure 2.4a,iv). As a test, we fused a flexible glycine/serine-rich 35 amino acid linker to the N terminus of the S1 protein and capped this with a second Smt3 domain (protein S2-35, Figure 2.3c,ii and Appendix A). Thus, the single folded-component sequence of S1 (that is, Smt3) is repeated twice in S2-35.

When protein S2-35 was captured in the nanopore with ClpX and ATP present in the *trans* compartment, an ionic current pattern with eight reproducible states was observed (Figure 2.4c and Appendix B). The first four states (Figure 2.4c,i-iv) were identical to states i–iv caused by S1 translocation (compare Figure 2.4a and c). This similarity included ramping state iii that is diagnostic for ClpX engagement, and the Smt3-dependent state iv. However, beginning at state v, the S2-35 pattern diverged

from the S1 pattern (compare Figure 2.4a and c). That is, following Smt3 translocation state iv, a typical S2-35 ionic current trace did not proceed to the open channel current but instead transitioned to a ~6.3-pA state with a median duration of 1.5 s (Figure 2.4c,v). This was followed by a ~8.5-pA state (Figure 2.4c,vi) that closely resembled ramping state iii, and a subsequent ionic current state that closely resembled the putative Smt3 translocation state iv (Figure 2.4c,vii). In other words, consistent with our model, the putative ClpX-bound and Smt3-dependent states that were observed once during S1 events (Figure 2.4a) were observed twice during S2-35 events (Figure 2.4c). These analogous states for the two constructs shared nearly identical amplitudes, root mean square (RMS) noise values and durations.

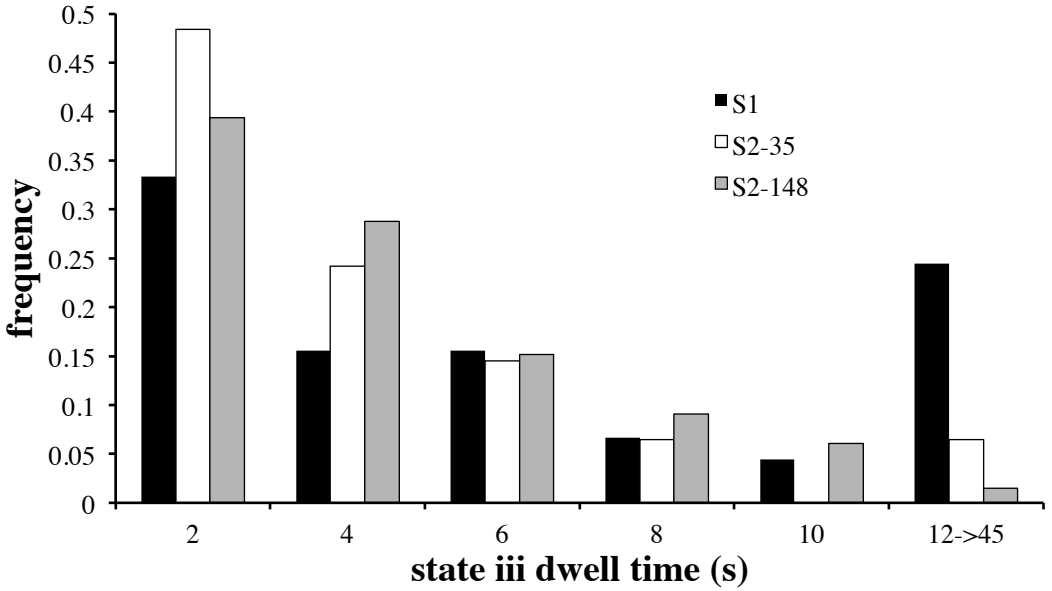


Figure 2.5: Comparison of ionic current state iii dwell times for ClpX/ATP-dependent translocation events. Black bars: median = 4.15 s, IQR = 7.52, n = 45. White bars: median = 2.12 s, IQR = 3.23, n = 62. Gray bars: median = 2.76 s, IQR = 3.98, n = 66.

This dependence of ionic current on protein structure is consistent with ClpX-driven protein translocation through the nanopore. As an additional test, we re-examined ionic current state v observed during S2-35 translocation. This state is consistent with movement of the 35-amino-acid linker through the nanopore based on two observations: (i) its average ionic current is measurably higher than that of surrounding states (Figure 2.4c) as expected for an amino acid sequence with few bulky side chains; and (ii) in the time domain, state v occurs between Smt3-dependent states iv and vi as expected, given its position along the S2-35 primary sequence (Figure 2.3c,ii and Appendix A).

If state v corresponds to translocation of the polypeptide linker under ClpX control, then changes in the length and composition of this linker should result in duration and current amplitude changes. To test this, we designed a third protein in which the S2-35 linker region was appended with an additional 113 amino acids, yielding a final construct consisting of two Smt3 domains separated by an extended 148-amino-acid flexible linker (protein S2-148, Figure 2.3c,iii and Appendix A). As predicted, when this protein was captured in the nanopore under standard conditions in the presence of ClpX and ATP, eight reproducible states similar to S2-35 events were observed (Figures 2.4d, 2.5, 2.6a, 2.7, 2.8 and Appendix B). Importantly, however, the S2-35 and S2-148 events differed substantially at state v (compare (Figure 2.4c and d). That is, the S2-148 state v had a higher mean residual current than did S2-35 (~9 versus ~6 pA, respectively), and a median duration ~2.5 fold longer than that of S2-35 state v

(Figure 2.6b). The increased duration of S2-148 state v relative to S2-35 state v was anticipated and consistent with the model described in Figure 2.4. The increased current level was likely due to differences in linker amino acid composition between the two proteins (S2-35 linker: 51% Gly, 34% Ser, 15% other; S2-148 linker: 34% Gly, 32% Ser, 19% Ala, 15% other). However, confirmation of this hypothesis will require systematic testing of the relationship between amino acid identity and resistance to ionic current through the pore lumen.

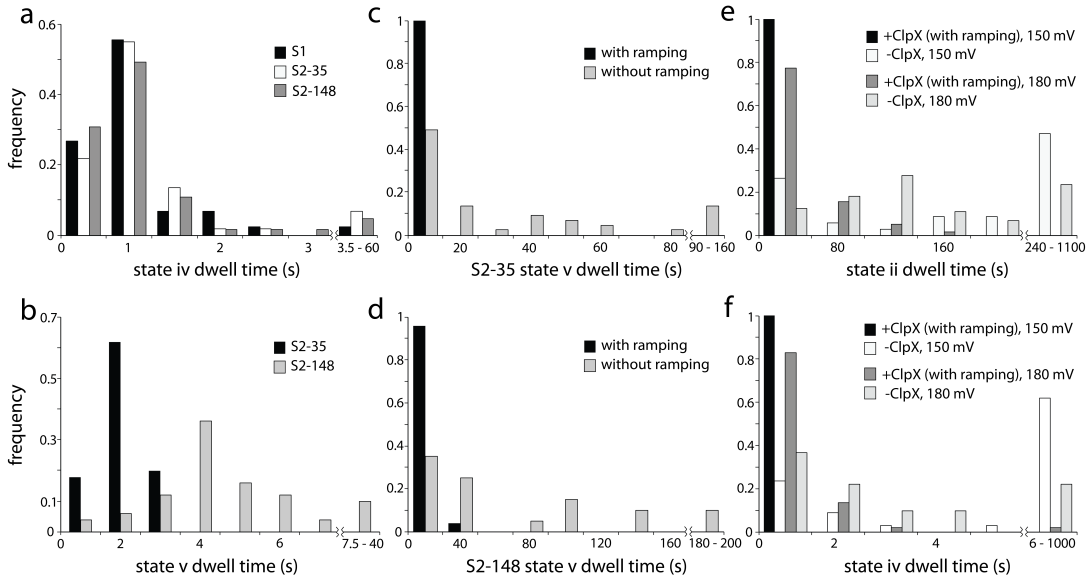


Figure 2.6: Ionic current state dwell times during translocation of model proteins through the nanopore. (**a**) Comparison of putative Smt3 translocation (state iv) dwell times for three model proteins. Values are from events that included the ClpX-dependent ramping state (Figure 2.4,iii). Black bars: median = 0.71 s, interquartile range (IQR) = 0.41, *n* = 45. Gray bars: median = 0.64 s, IQR = 0.47, *n* = 60. White bars: median = 0.63 s, IQR = 0.40, *n* = 65. (**b**) Comparison of putative linker region (state v) dwell times for S2-35 and S2-148 proteins. Values are from events that included the ClpX-dependent ramping state (Figure 2.4,iii). Black bars: median = 1.52 s, IQR = 0.68, *n* = 50. Gray bars: median = 3.62 s, IQR = 2.03, *n* = 50. (**c**) State v translocation dwell times for S2-35 events. Black bars: median = 1.52 s, IQR = 0.68, *n* = 50. Gray bars: median = 11.45 s, IQR = 39.53, *n* = 45. (**d**) State v

translocation dwell times for S2-148 translocation events. Black bars: median = 3.62 s, IQR = 2.03, *n* = 50. Gray bars: median = 37.07 s, IQR = 89.80, *n* = 20. (**e**) State ii dwell times for protein substrates at two voltages. Black bars: median = 4.89 s, IQR = 6.72, *n* = 104. White bars: median = 165.50 s, IQR = 351.75, *n* = 34. Gray bars: median = 17.26 s, IQR = 31.08, *n* = 173. Light gray bars: median = 90.0 s, IQR = 112.0, *n* = 72. (**f**) State iv dwell times for the S1 protein substrate at two voltages. Black bars: median = 0.33 s, IQR = 0.40, *n* = 104. White bars: median = 8.25 s, IQR = 26.82, *n* = 34. Gray bars: median = 0.65 s, IQR = 0.45, *n* = 52. Light gray bars: median = 1.74 s, IQR = 3.12, *n* = 41.

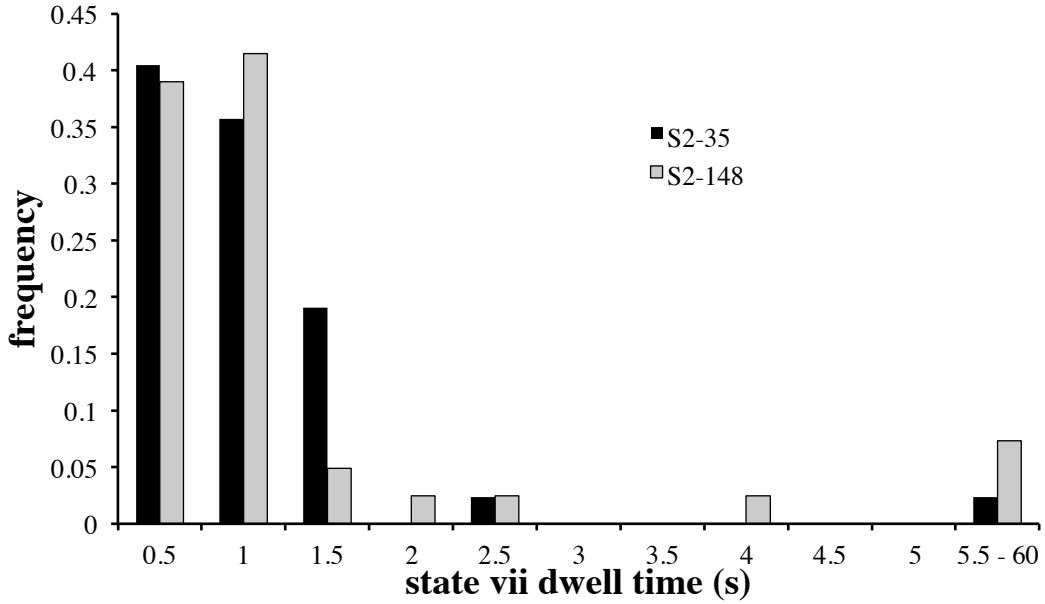


Figure 2.7: Comparison of ionic current state vi dwell times for ClpX/ATP-dependent translocation events. Black bars: median = 1.80 s, IQR = 0.97, n = 44. Gray bars: median = 1.61 s, IQR= 0.97, n = 44.

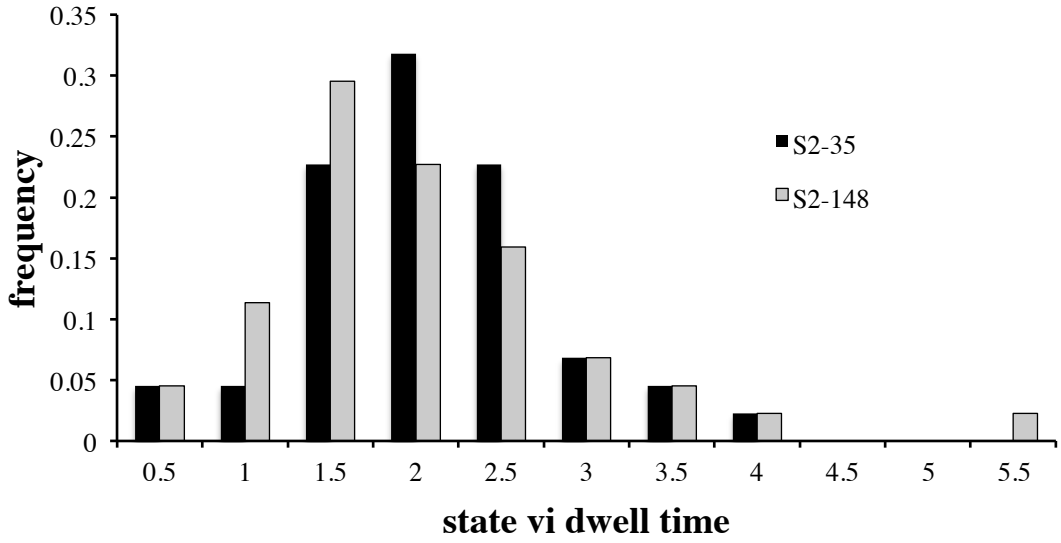


Figure 2.8: Comparison of ionic current state vii dwell times for ClpX/ATP-dependent translocation events. Analyzed events included ClpX/ATP-dependent ramping states iii and vi. Black bars: median = 0.63 s, IQR = 0.93, n = 42. Gray bars: median = 0.66 s, IQR = 0.90, n = 41.

2.3 Voltage-mediated translocation

Prior studies have shown that proteins can translocate through nanopores under an applied voltage without the assistance of processive enzymes[46-51]. This was also the case in our experiments, that is, translocation of the three model proteins was observed in the absence of ClpX- or ATP-dependent mechanical work performed on captured strands (Figure 2.9). However, these ClpX- and ATP-independent translocation events lacked the diagnostic ramping states (Figure 2.4), and they were measurably longer and more variable in duration than were ClpX-mediated translocation events (Figure 2.6c-f). This is consistent with an unregulated translocation process that depends upon random structural fluctuations of the captured protein molecule and intermittent electrical force acting on amino acid segments with

variable charge density in the pore electric field. This model predicts that ClpX-dependent translocation will be relatively unaffected by changes in applied voltage. This proved to be true. State ii and iv dwell times for ClpX- and ATP-dependent S1 translocation events acquired at 150 mV were comparable to those acquired at 180 mV (Figure 2.6e,f). At both voltages, these events were consistently faster and more narrowly distributed than ClpX- or ATP-independent events. Thus, ClpX activity (not voltage) is dominating the unfolding and translocation process.

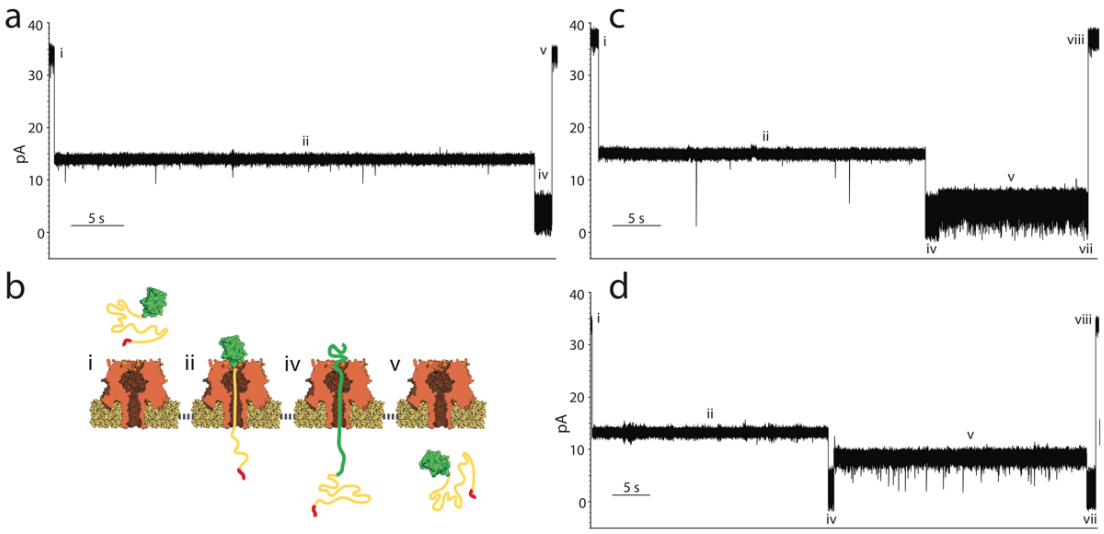


Figure 2.9: Ionic current traces showing translocation of the three model proteins absent ClpX/ATP-dependent mechanical work (no ramping states iii/vi). Following state ii, all protein substrates we examined eventually unfolded and translocated due to the 180 mV applied potential. All events exhibited more widely distributed state dwell times compared to ClpX-mediated events (Fig. 2.6). (**a**) S1 translocation. Note the absence of state iii compared to Figure 2.4a. (**b**) Model of ClpX/ATP-independent protein S1 trannslocation. Cartoons i-i′ correspond to ionic current states i-i′ in a. (**c**) S2-53 translocation. Note the absence of ramping states iii and vi compared to Figure 2.4c. (**d**) S2-148 translocation. Note the asence of states iii and vi compared to Figure 2.4d.

Chapter 3

DISCRIMINATION AMONG PROTEIN VARIANTS

The results presented in Chapter 2 suggest that nanopore devices could be used for sequential protein analysis and identification. To examine this further, we designed experiments to answer two questions. First, can the nanopore device discriminate among distinct protein domains in series along individual protein strands as they are driven through the pore sensor?; and 2) Can the nanopore device discriminate among variants of these protein domains, e.g. structural modifications arising from point mutations, truncations, and rearrangements? Such changes are common to pathogenic protein variants[52-55], and they should be detectable if protein sequence, stability, and unfolding pathways do in fact account for the ionic current patterns we observe.

We addressed these questions using ~700 amino-acid-long engineered proteins bearing well-characterized folded domains. The motor used in these experiments was

ClpXP, the complete proteasome-like complex of *E. coli* and other prokaryotes. We found that specific point mutations, proteolytic cleavage, and sequence rearrangements in these domains resulted in detectable ionic current pattern changes. Naive Bayes-derived decision boundaries applied to our data resulted in single protein identification at 86.4% to 98.7% accuracy.

3.1 S2-GT: a model multidomain protein

3.1.1 Experimental Optimization

In the present study, our set-up included two modifications that increased the number of protein strands that could be analyzed per experiment and the efficiency of translocation for each captured protein relative to experiments in Chapter 2. These were: 1) addition of a conventional ATP-regeneration mixture to the *trans* compartment to maintain a constant ATP concentration over time; and 2) supplementation of the ClpX motor with ClpP to form the ClpXP complex. In the bacterial cell, when an ssrA-tagged protein is unfolded by ClpX, it is threaded into the lumen of an associated compartmentalized peptidase, ClpP, where it is degraded[37]. We found that protein translocations driven by the ClpXP protease were less prone to long off-pathway stalls and slips than were translocations driven by ClpX alone. This observation is consistent with previous studies showing that ClpXP is a more robust unfoldase than is ClpX[40,41]. In addition, trimming of the substrate protein by ClpP in the *trans* compartment reduced the frequency of irreversible protein captures in the αHL pore.

3.1.2 Engineering and design of S2-GT

The reference protein used for our experiments, 'S2-GT', was an ~700 amino acid (aa) strand composed of four folded domains connected by short aa linkers. Based on crystal structures, the individual protein domains (ubiquitin-like protein (Smt3), titin fragment (titin I27), and green fluorescent protein (GFP)) were too large to pass through the αHL pore without unfolding of their native tertiary structures[43,56,57]. Each strand was capped at its carboxy terminus by an aa polyanion (polyGSD), and the ClpX-recognition ssrA motif (Figure 3.1b, and Appendix A). At 180 mV applied potential, the polyGSD tail threaded into the pore and the ssrA tag became accessible to ClpXP in the *trans* solution (Figure 3.1b).

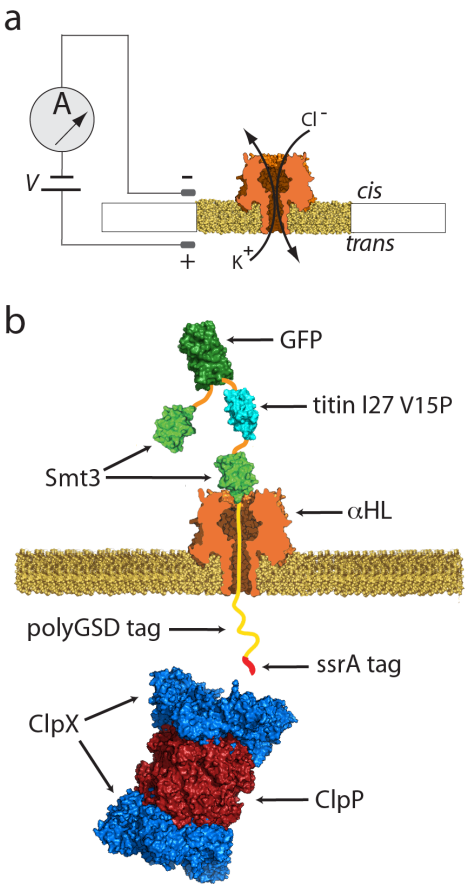


Figure 3.1: Experimental set-up for S2-GT studies. (a) Nanopore sensor. A single αHL pore is embedded in a lipid bilayer separating two polytetra-fluoroethylene wells each containing 100 µl of 0.3 M KCl solution at 30 °C. Voltage is applied between the wells (*trans* side +180 mV), causing ionic current flow through the channel. (b) Protein S2-GT capture in the nanopore. S2-GT is a model protein bearing four folded domains: Smt3 (light green), GFP (dark green), and titin I27 V15P (cyan), coupled to a negatively charged flexible polyGSD region (yellow) and an ssrA tag (red) at its C-terminus. As a result of the applied voltage, the negatively charged polyGSD tag is threaded through the pore into the *trans*-side solution until the first folded Smt3 domain prevents further translocation of the captured protein. ClpXP present in the *trans* solution binds to the C-terminal ssrA sequence. Fueled by ATP, ClpXP translocates along the protein tail toward the channel, and catalyzes sequential unfolding and translocation of the entire multidomain protein through the pore.

3.1.3 ClpXP-mediated unfolding and translocation of S2-GT

Prior to nanopore-based experiments, we tested if ClpXP could unfold and degrade the S2-GT protein in bulk phase. Fluorescence-based experiments confirmed that ClpXP was capable of degrading the GFP domain within S2-GT (Figure 3.2).

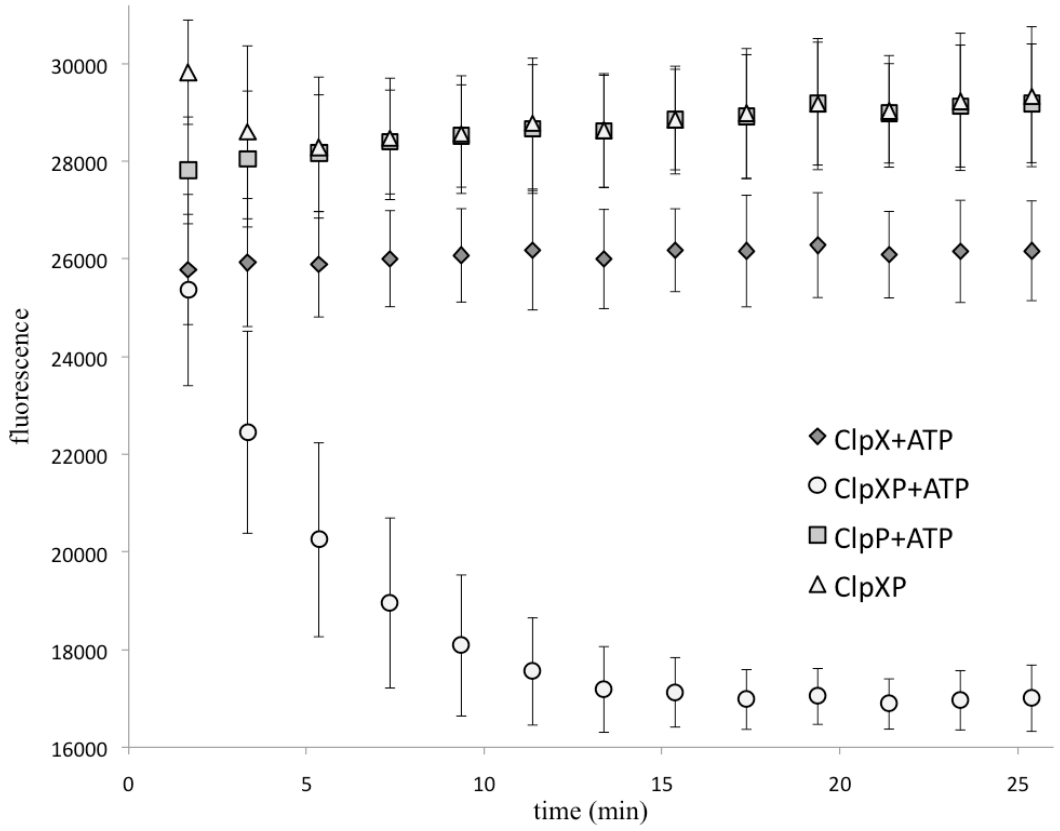


Figure 3.2: Fluorescence-based bulk phase ClpXP activity assay using the S2-GT protein. ClpX(P)/ATP-dependent quenching of the GFP domain within S2-GT. ClpX+ATP n=4, ClpXP+ATP n=4, ClpP+ATP n=4, ClpXP n=3.

Ionic current traces that arise from ClpXP/ATP-dependent translocation of the S2-GT protein are illustrated in Figure 3.3a and b (also see Appendix B). Each event began at ~53 pA (open channel current Figure 3.3b,i), followed by a drop to ~22 pA upon

S2-GT capture in the pore (Figure 3.3b,ii). This ionic current state persisted until ClpXP bound to the ssrA tag and began pulling on the polyGSD tail. Pulling caused a gradual current decrease (Fig. 3.3b,iii), followed by a sudden drop to a median current of ~9.6 pA (Fig. 3.3b,iv) characterized by high variance (s.d.= 2 pA). This pattern was quantitatively consistent with previous work in Chapter 2 which correlated states iii and iv with pre-unfolding dwell of the Smt3 atop the pore orifice, followed by unfolding and translocation. Because a second Smt3 domain was included near the amino terminus of S2-GT, we predicted that this ionic current pattern would be repeated at the end of each complete translocation event. This prediction was supported by our data. Notably, the last state prior to return to open channel current (Fig. 3.3b xi), shared nearly identical characteristics with state iv (mean currents: iv = 9.6 ± 1.7 pA, xi = 9.3 ± 1.6 pA; s.d.: iv = 2.0 ± 0.4 pA, xi = 2.0 ± 0.4 pA; dwell time: iv = 720 ± 320 ms, xi = 540 ± 290 ms).

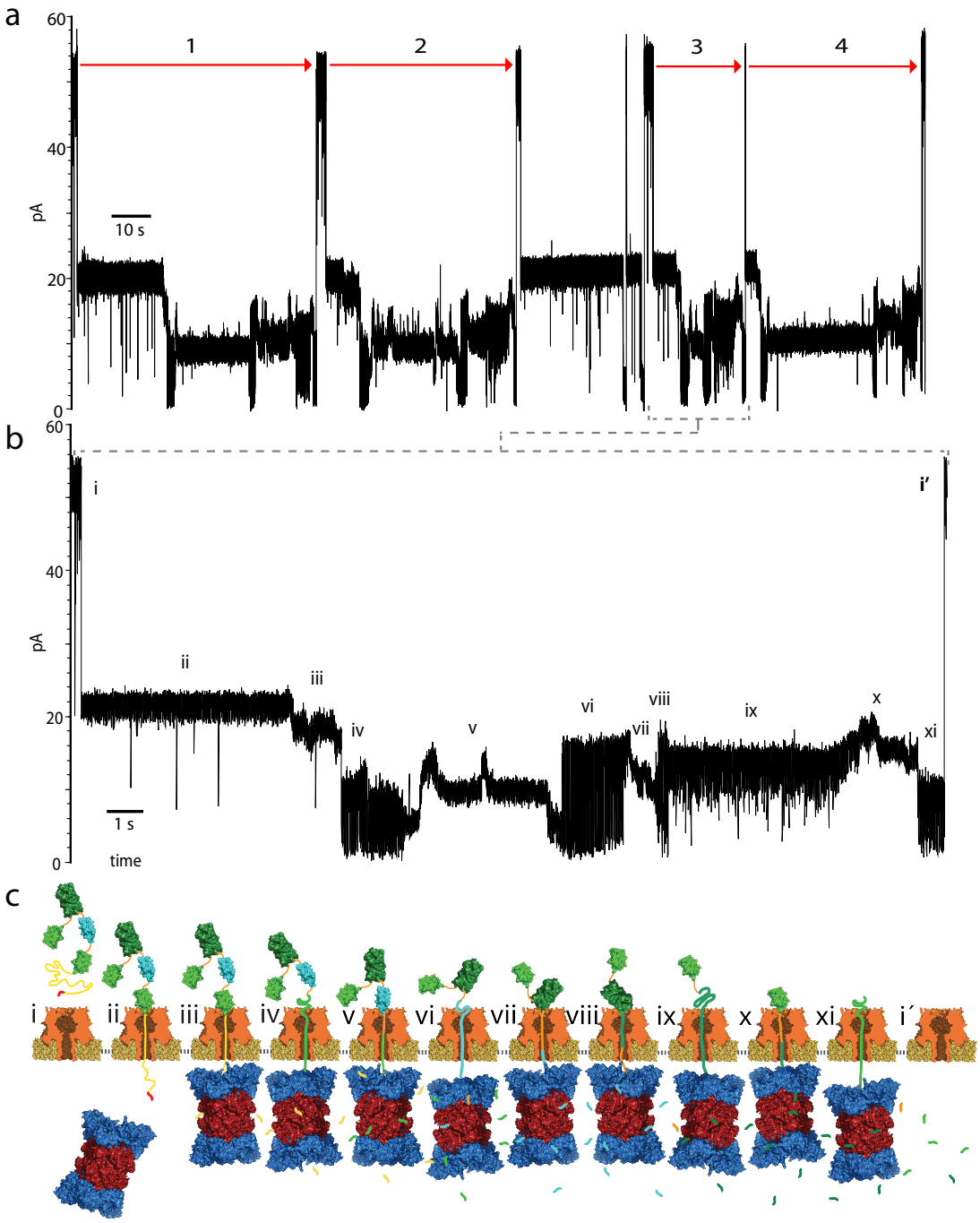


Figure 3.3: Ionic current traces during ClpXP-mediated protein S2-GT translocation. (a) Four consecutive S2-GT translocation events. The gap between the third and fourth events corresponds to protein captures that were ejected from the nanopore by briefly reversing voltage polarity. (b) Expanded view of ionic current

states during S2-GT translocation. Open channel current through the αHL nanopore under standard conditions (mean ~53 pA, current s.d. 1.2 pA) (i). Initial capture of the S2-GT substrate (mean current ~22 pA, current s.d. 0.7 pA) (ii). ClpXP-mediated C-terminal Smt3 pre-unfolding (mean current ~19.1 pA, current s.d. 1.7 pA) (iii). C-terminal Smt3 domain unfolding and translocation through the nanopore (mean current ~9.6 pA, current s.d. 2.0 pA) (iv). Ionic current transition into the titin I27 V15P pre-unfolding state. Several discrete current levels are typically observed (mean current ~9.5 pA, current s.d. 2.3 pA) (v). Unfolding and translocation of the titin I27 V15P domain through the nanopore (mean current ~14 pA, current s.d. 4.6 pA) (vi). The GFP pre-unfolding state. Several discrete current levels are typically observed (mean current ~14 pA, current s.d. 2.0 pA) (vii). Extraction of the C-terminal beta strand 11 of GFP (mean current ~11 pA, current s.d. 3.7 pA) (viii). Global unfolding and translocation of GFP (mean current ~15 pA, current s.d. 1.4 pA) (ix). N-terminal Smt3 pre-unfolding state (mean current ~15 pA, current s.d. 1.7 pA) (x). N-terminal Smt3 domain unfolding and translocation through the nanopore (mean current ~9.3 pA, current s.d. 2.0 pA) (xi). Return to open channel current upon completing translocation of the entire S2-GT protein to the *trans* compartment (i′). (c) Working model of ClpXP-mediated S2-GT translocation. Roman numerals assigned to each panel correspond to ionic current states in b.

Given these Smt3-dependent ionic current 'bookends', it was logical that intervening ionic current states v-ix would correlate with processing of the titin I27 and GFP domains. Thus, we developed a model in which these two domains also contributed unique pre-unfolding and translocation current states (Figure 3.3c). If correct, the characteristics of each of these states should be domain-dependent, and variations within those domains should cause predictable changes in their ionic current signatures enabling their discrimination from the reference protein.

3.2 S2-GT Variants

3.2.1 Nanopore analysis of the titin I27 domain and a destabilized mutant

In our model, we assigned ionic current states v and vi to titin I27 because it is the next domain along the S2-GT polypeptide that would contact the nanopore following translocation of the C-terminal Smt3 domain (ionic current state iv). Quantitative evidence supports this assignment. Ionic current state v (putative pre-unfolding of titin I27) had an average dwell time of 26 ± 20 s. This long dwell time is consistent with bulk phase studies which showed that titin I27 is resistant to ClpXP-mediated unfolding[58]. Further, state v exhibited ionic current level repeats (arrows in Appendix B) suggestive of small steps and slips of ClpXP as it attempted to advance along the polypeptide strand against a significant energy barrier. Immediately following a successful unfolding attempt, the ionic current shifted abruptly to state vi. Analysis of state vi dwell time (1.4 s ± 0.6 s) suggests that ClpXP pulls the unfolded titin I27 domain through αHL at an average rate of 64 aa/s. This rate is similar to previous studies that established a maximum ClpXP translocation rate of 60-70 aa/s[59].

If these assignments are valid, it follows that changes in the stability of the titin I27 domain would result in detectable changes in ionic current state v. As a test, we constructed an S2-GT variant (S2-GT^{EE}) where two buried cysteines (C47, C63) were mutated to glutamic acid residues (Figure 3.4a). These side chain alterations are similar to carboxymethylation of C47/C63 and mutation of those cysteines to aspartic

acid that are known to destabilize the titin I27 domain[58,60]. As anticipated, S2-GT^{EE} state v dwell times were several orders of magnitude shorter than were S2-GT state v dwell times (Figure 3.4b and Appendix B). The other ionic current states remained relatively unchanged between the two constructs (Figures 3.4c and 3.5).

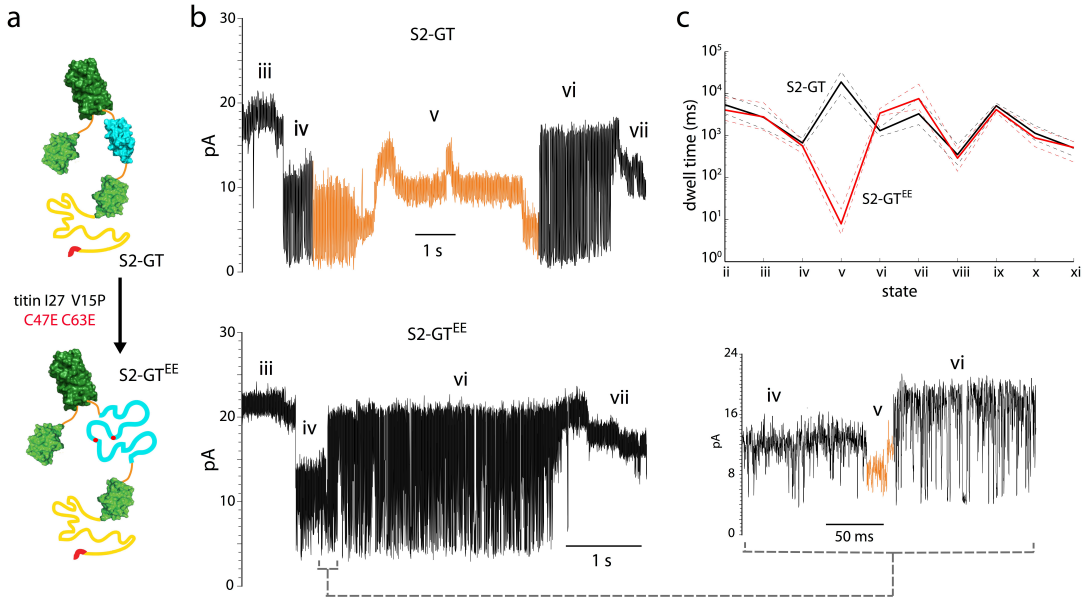


Figure 3.4: Ionic current state v is dramatically changed by two point mutations in the titin I27 V15P domain of S2-GT. (a) Cartoon depiction of the two proteins that were compared in this experiment. S2-GT is at the top and a modified version bearing two point mutations within the titin I27 V15P domain (S2-GT^{EE}: C47E C63E) is at the bottom. Modifications at C47 and C63 are known to destabilize titin I27 tertiary structure (b) Ionic current states iii-vii of representative S2-GT (top) and S2-GT^{EE} (bottom) translocation events. State v for each event is colored orange. (c) A parallel coordinates plot comparing median dwell times for ionic current states ii-xi of S2-GT (black, n=91 translocations) and S2-GT^{EE} (red, n=93 translocations). The median state v dwell time of S2-GT^{EE} is ~3.5 orders of magnitude shorter than the comparable state v median dwell time of S2-GT (7.9 ms and 22.5 s, respectively). Dashed lines represent the 1st and 3rd quartile medians.

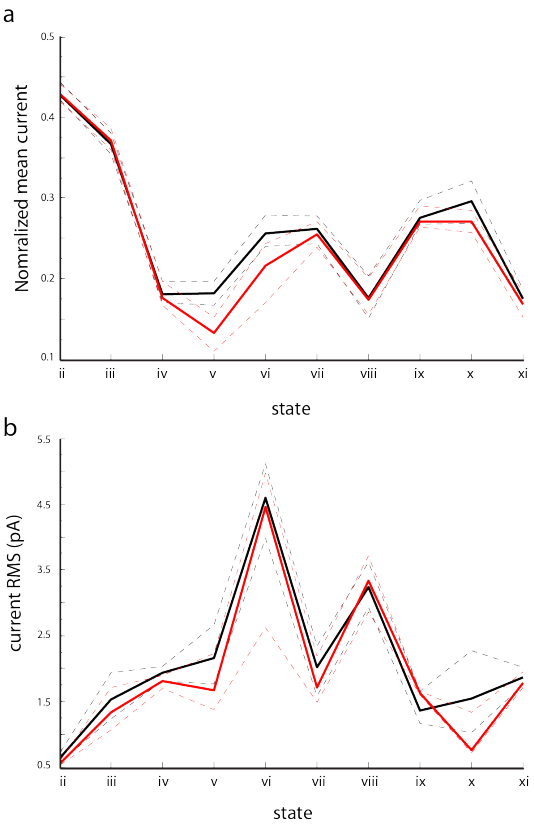


Figure 3.5: Comparison of ionic current state characteristics of S2-GT and S2-GT^{EE}. Parallel coordinates plots comparing (a) normalized mean current and (b) current s.d. for ionic current states ii-xi of S2-GT (black, n=91 translocations) and S2-GT^{EE} (red, n=93 translocations). Dashed lines represent the 1st and 3rd quartile medians.

3.2.2 Nanopore analysis of the GFP domain and a 'superfolder' variant

After translocation of titin I27, the next domain along the S2-GT strand is GFP. Accordingly, we predicted that states vii-ix would correlate with pre-unfolding and translocation of GFP (Fig. 3.3b and 3.3c). As was the case for ionic current state v (titin pre-unfolding), ionic current state vii often contained repeated current levels suggestive of small steps and slips of ClpXP as it attempted to unfold GFP. This state (mean ionic current 13.8 pA, ionic current s.d. 2.0 pA, and median dwell time 3.1 s)

ended with an abrupt and irreversible transition to state viii (mean ionic current 11.0 pA, ionic current s.d. 3.1 pA, and median dwell time of 380 ms). This state is consistent with a step along the GFP unfolding pathway that corresponds to extraction of the 11th beta-strand that precedes global GFP unfolding[59,61]. State viii was followed by a distinct shift to state ix characterized by a higher mean ionic current (14.8 pA) with relatively low noise (s.d. = 1.4 pA). We reasoned that state ix corresponds to translocation of the unfolded GFP domain through αHL at 34-58 aa/s following successful ClpXP-mediated GFP unfolding.

As an initial test to determine if altering the GFP domain would change ionic current states vii-ix, we engineered an S2-GT construct (S2-G^{SF}T) in which the GFP domain was replaced by a 'super-folding' GFP variant (Figures 3.6a and Appendix A). Superfolder GFP (GFP^{SF}) contains 11 point mutations which increase its resistance to chemical denaturants and which help maintain GFP fluorescence when beta-strands that form the functional core are permuted[62]. As anticipated, the characteristics of S2-G^{SF}T events were altered relative to S2-GT events (Figures 3.6b and 3.7, and Appendix B). Most notably, S2-G^{SF}T events exhibited a three level ionic current pattern within state ix (states ix-a, ix-b, and ix-c) that was absent in S2-GT events. We argue that this pattern reflects additional pre-unfolding (ix-b) and translocation (ix-c) states arising from an additional GFP unfolding intermediate. This argument is based on two facts: 1) Single-molecule optical tweezer experiments have revealed a short-lived GFP unfolding intermediate in which beta-strands 6→1 maintain their

tertiary structure following initial ClpXP-mediated unfolding of GFP beta-strands 11→7[41]. 2) The combined dwell times of states ix-a (mean = 1.7 s) and ix-c (mean = 3.7 s) is 5.4 s, which is similar to the mean dwell time of state ix (5.3 s) for translocation of GFP beta-strands 11→1 in S2-GT events (Figure 3.8). Thus, ionic current states ix-a and ix-c are consistent with sequential translocation of GFP^{SF} beta-strands 11→7 and 6→1 separated by an intermediate pre-unfolding state (ix-b) that is not observable in the S2-GT strand.

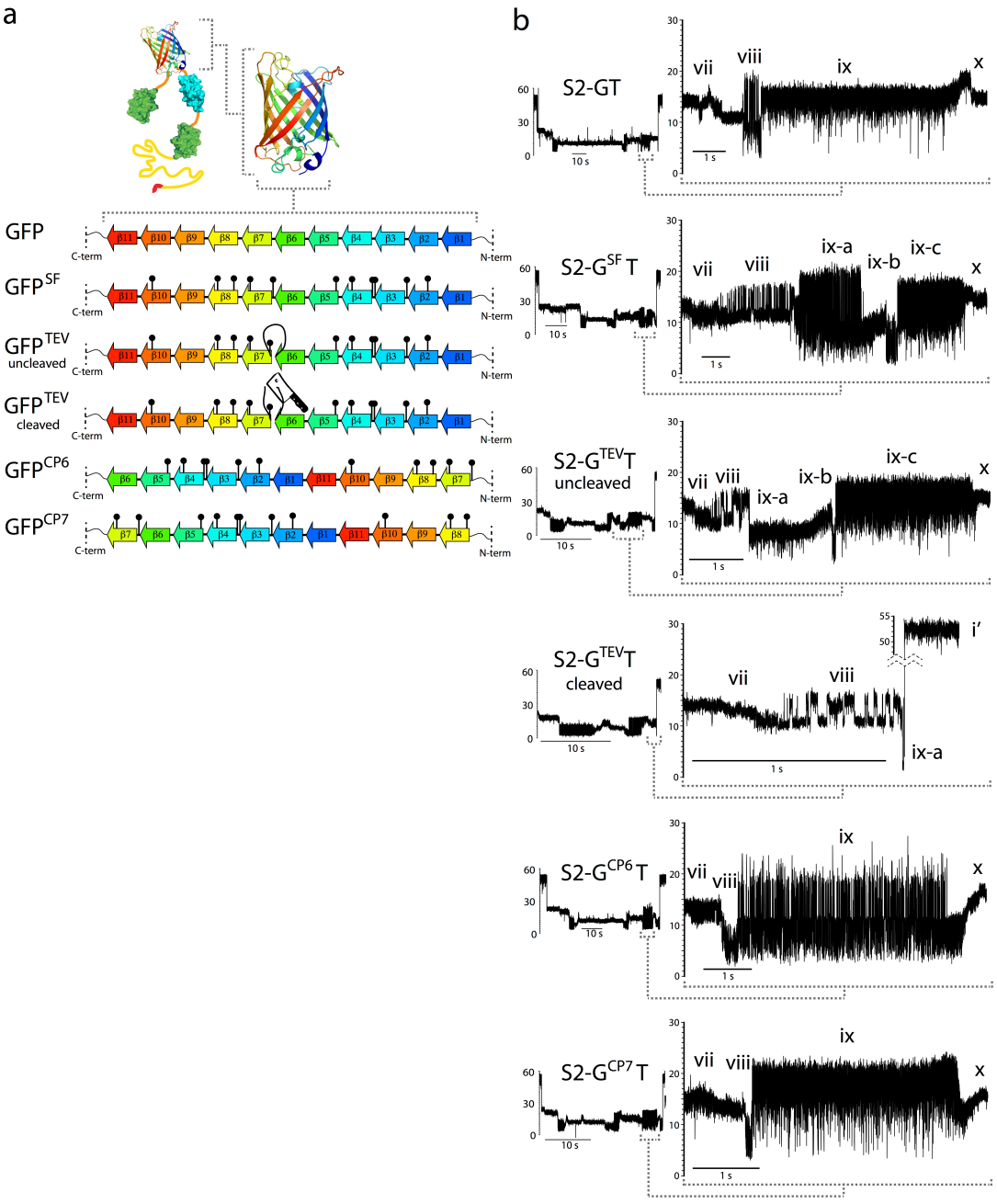


Figure 3.6: Ionic current signatures for S2-GT GFP variants. (a) Beta strand connectivity of the GFP domain within proteins S2-GT (GFP), S2-G^{SF}T (GFP^{SF}), cleaved and uncleaved S2-G^{TEV}T, S2-G^{CP6}T (GFP^{CP6}), and S2-G^{CP7}T (GFP^{CP7}). Each colored arrow represents a beta strand. GFP^{SF} (superfolder GFP) contains 11 point mutations (black markers) compared to GFP. GFP^{TEV} contains a TEV protease cleavage site between the 6th and 7th beta strands of GFP^{SF}. GFP^{CP6} and GFP^{CP7} are

circular permutations of GFP^{SF} between the $6^{th}/7^{th}$ and $7^{th}/8^{th}$ beta strands, respectively. (b) Representative ClpXP-mediated GFP variant translocation events with expanded views of GFP-dependent ionic current states vii-ix. Ionic current states vii and viii are pre-unfolding of GFP. State ix is translocation of the unfolded GFP domain. Ionic current states ix for S2-G^{SF}T and S2-G^{TEV}T include three unique sub-states (ix-a, ix-b, and ix-c) that correspond to translocation of unfolded beta strands 11→7 (ix-a), pre-unfolding of an intermediate (ix-b), and translocation of the unfolded intermediate beta strands 6→1 (ix-c). Cleavage of S2-G^{TEV}T with TEV protease terminates the event following a brief ionic current state ix-a.

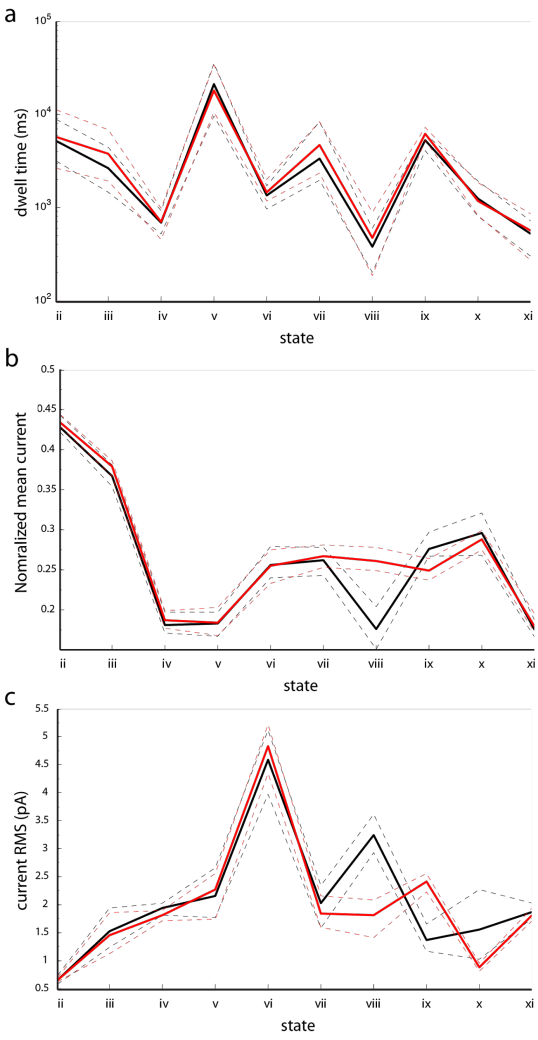


Figure 3.7: Comparison of ionic current state characteristics of S2-GT and S2-G^{SF}T. Parallel coordinates plots comparing (**a**) median dwell time, (**b**) normalized

mean current, and (**c**) current s.d. for ionic current states **ii-xi** of S2-GT (black, n=91 translocations) and S2-G^{SF}T (red, n=78 translocations). Dashed lines represent the 1st and 3rd quartile medians.

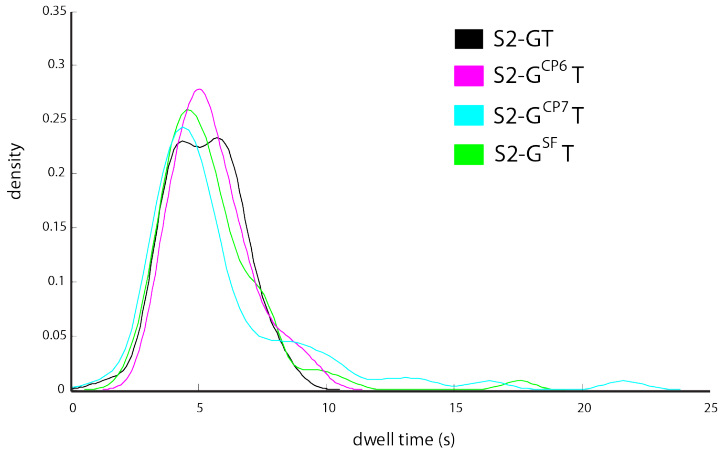


Figure 3.8: GFP domain translocation dwell times. The data are kernel density distributions for ionic current state **ix** dwell times for four GFP variants. The S2-G^{SF}T values are the sum of states **ix-a** and **ix-c**. S2-GT n=91; S2-GT^{EE} n=93; S2-G^{SF}T n=78; S2-G^{CP6}T n=73; S2-G^{CP7}T n=82, where n is the number of translocation events for a given protein variant.

3.2.3 Protease cleavage of 'superfolder' GFP

If ionic current states ix-a and ix-b/c are dependent on GFP^{SF} beta-strands 11→7 and 6→1, respectively, then cleaving the polypeptide chain between the two regions should terminate translocation events at state ix-a. To test this, we inserted a Tobacco Etch Virus (TEV) protease cleavage site between the 6th and 7th beta-strands of GFP^{SF} (protein S2-G^{TEV}T, Figure 3.6a and Appendix A). Consistent with our prediction, the uncleaved protein retained states ix-b→xi, while cleavage with TEV protease resulted in events terminating at state ix-a (Fig. 3.6b and Appendix B).

Surprisingly, the cleaved GFP^{SF} beta-strands 11→7 translocation rate (~6,000 aa/s) was much faster than expected for ClpXP-mediated translocation (~50 aa/s). A likely explanation is that cleavage and separation of the 275 N-terminal amino acids in the *cis* compartment reduced hydrodynamic drag on the translocating strand[63]. This would allow the relatively weaker electrophoretic force to drive translocation of the cleaved strand at a high rate absent ClpXP activity.

3.2.4 Structural rearrangements of 'superfolder' GFP

If ionic current state ix is sensitive to GFP sequence as we assert, then rearrangement of beta-strand order should cause the state ix current pattern to change as well. To examine this, we constructed variants of the S2-G^{SF}T protein in which the GFP^{SF} domain was circularly permuted between the 6^{th} and 7^{th} beta-strands (protein S2-G^{CP6}T) and between the 7^{th} and 8^{th} beta-strands (S2- G^{CP7}T) (Figure 3.6a and Appendix A)[62]. Representative ClpXP-mediated ionic current traces for proteins S2-G^{CP6}T and S2-G^{CP7}T are shown in Figure 3.6b (see also Appendix B). State ix of S2-G^{CP6}T and S2-G^{CP7}T events differed from the three state current pattern observed in S2-G^{SF}T events, displaying a single nearly-homogeneous current state similar to S2-GT events in dwell time, but differing in current mean and current variance (Figures 3.6b, 3.9, and 3.10). These results are consistent with a model where ionic state ix is

sensitive to the sequence topology of GFP translocation. State viii characteristics for S2-G^{CP6}T and S2-G^{CP7}T events also differed from S2-GT and S2-G^{SF}T events (Figure 3.6b). This is not surprising because these structural rearrangements altered the identity of the first beta-strand extracted during unfolding.

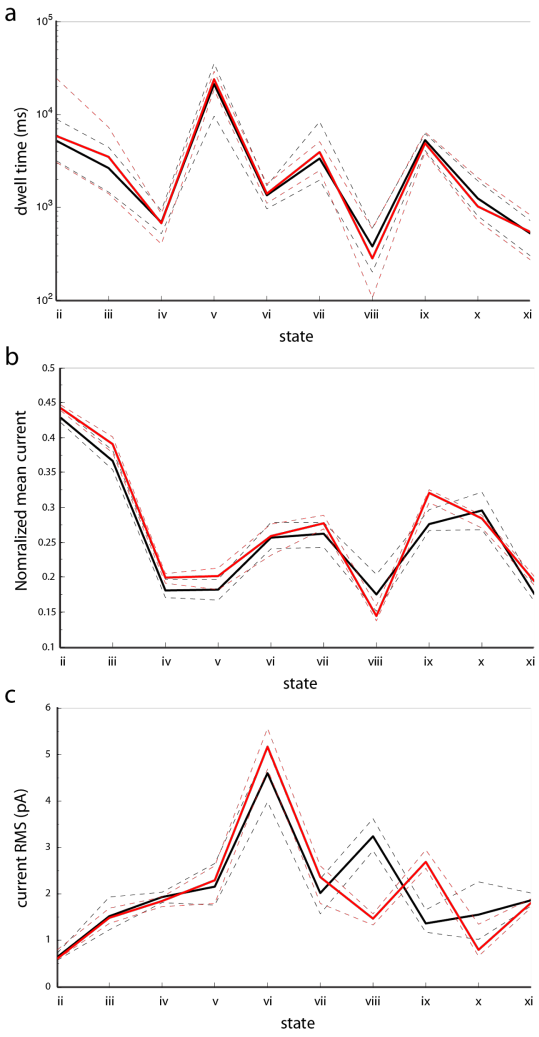


Figure 3.9: Comparison of ionic current state characteristics of S2-GT and S2-G^{CP6}T. Parallel coordinates plots comparing (**a**) median dwell time, (**b**) normalized mean current, and (**c**) current RMS for ionic current states **ii-xi** of S2-GT (black,

n=91) and S2-G^{CP6}T (red, n=73), where n is the number of translocation events. Dashed lines represent the 1st and 3rd quartile medians.

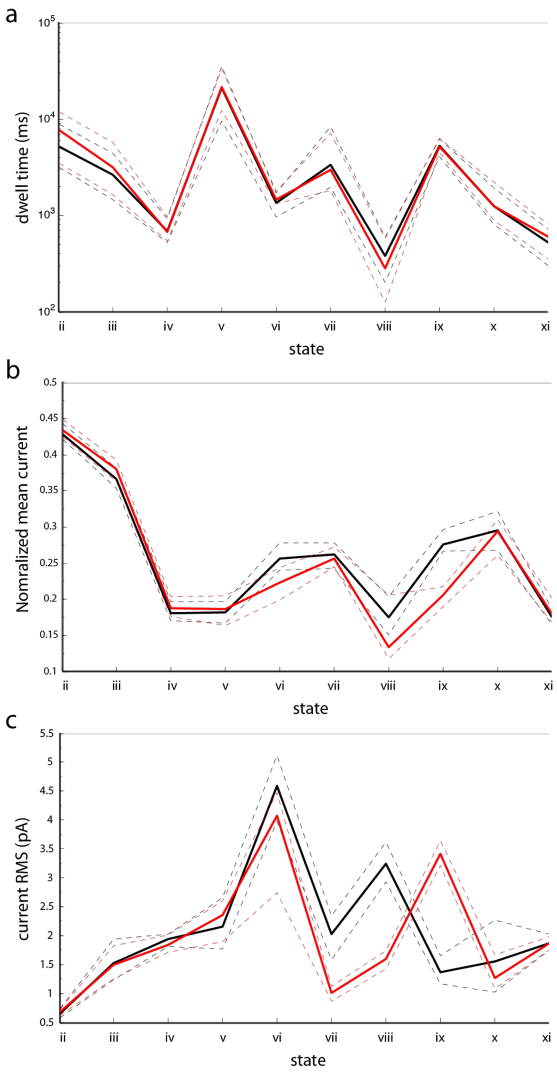


Figure 3.10: Comparison of ionic current state characteristics of S2-GT and S2-G^{CP7}T. Parallel coordinates plots comparing (**a**) median dwell time, (**b**) normalized mean current, and (**c**) current RMS for ionic current states **ii-xi** of S2-GT (black, n=91) and S2-G^{CP7}T (red, n=82), where n is the number of translocation events. Dashed lines represent the 1st and 3rd quartile medians.

3.3 Discrimination among protein variants using Naive Bayes classifiers

One motivation for this research is to develop a nanopore device that uses sequential ionic current measurements to identify individual proteins. As a test of the ClpXP-αHL prototype, we quantified the accuracy of calls among five of the S2-GT variants examined in this study.

3.3.1 Identifying features important for variant discrimination

Values for three parameters (dwell time, average current amplitude, and standard deviation of the current amplitude) were collected for states ii-to-xi within each complete translocation event (approximately 80 events for each of the five variants). To determine which of these 30 features were useful for protein classification, we performed a random forest analysis (see Appendix C). First, pairwise comparisons were performed between the null construct, S2-GT, and each of the four variants independently (Figure 3.11). As expected, the most important feature for distinguishing between S2-GT and S2-GT^{EE} was dwell time for the titin I27 pre-unfolding (state v). Also, as expected, features that were important for distinguishing between S2-GT and its GFP variants centered on states vii-ix (pre-unfolding and translocation of GFP). In these cases, average current amplitude and current standard deviation proved to be important features.

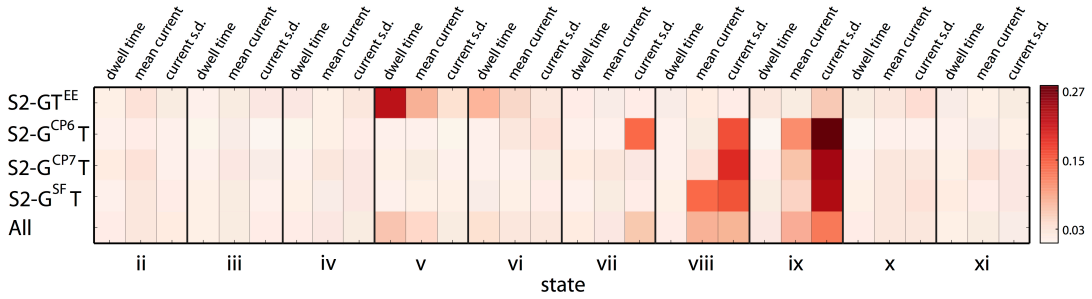


Figure 3.11: Identification of ionic current states important for discriminating between S2-GT variants. Each column represents a feature (dwell time, mean current, or current s.d.) for each current state (ii-xi). Each row is one of the S2-GT variants compared against S2-GT using a binary comparison, with the exception of the last row, which is a multi-class comparison of all the S2-GT constructs together. Each row is normalized and sums to 1. The "heat" of each square (scale at right) represents the relative importance of that feature as determined by a forest of extremely randomized trees (see Materials and Methods). State v dwell time was the most important feature for discriminating between proteins S2-GT and S2-GT^{EE}. The current s.d. of states viii and ix were the most important for discriminating between S2-GT and GFP variant proteins S2-G^{SF}T, S2-G^{CP6}T, and S2-G^{CP7}T.

We then reframed the question and asked which features were important for classification of a given translocation event when comparing all five protein constructs against one another simultaneously. Predictably, this analysis yielded eight useful features (row labeled 'All' in Figure 3.11) that were a composite of the features identified by pairwise comparisons.

3.3.2 Assessing discrimination accuracy

To estimate the accuracy with which we could call a given translocation event, we used these eight pertinent features and a Naive Bayes classifier[64] to establish a

confusion matrix for S2-GT and the four variants compared in Figure 3.11. Naive Bayes classifiers are suitable for this data set because they do not require tuning of parameters and thus avoid unnecessary complexity for preliminary tests. Upon building a confusion matrix using maximum *a posteriori* estimates, we found that there was an 86.4 to 98.7% chance of making an accurate call for each protein variant (Table 3.1).

	Predicted Class				
Actual Class	S2-GT	S2-GT^{EE}	S2-G^{CP6}T	S2-G^{CP7}T	S2-G^{SF}T
S2-GT	98.68	0.05	0.05	1.15	0.05
S2-GT^{EE}	1.13	96.57	1.13	1.13	0.05
S2-G^{CP6}T	0.07	0.07	95.63	2.80	1.43
S2-G^{CP7}T	0.06	0.06	12.22	86.38	1.28
S2-G^{SF}T	2.62	0.06	0.06	3.90	93.35

Table 3.1: Confusion matrix for discriminating between S2-GT variants using a multi-class Naive Bayes classifier. Each cell represents the percent probability of classifying a particular S2-GT variant (left column labels) as any of the five variants (top row labels). The diagonal (light gray boxes) represents the correct classification.

3.4 Discussion of protein variant discrimination

This study was motivated by two questions pertaining to sequential protein analysis using nanopores. First, can we distinguish between different protein domains in series along individual strands as they are driven through the αHL pore by an enzyme motor? Two lines of evidence indicate that we can. i) Pre-unfolding states differed among Smt3, GFP, and titin I27. For example, the pre-unfolding dwell time for titin I27 was substantially longer than pre-unfolding dwell times for GFP and Smt3 (Figure 3.4c). This is consistent with titin I27's characteristic resistance to mechanical denaturation[58]. Further, GFP displayed a three-state unfolding pathway, distinct from titin I27 and Smt3 two-state pathways (Figure 3.6b); ii) Overall, individual domain translocation states had ionic current signatures that were quantitatively distinguishable from one another based on current mean and s.d., while dwell times were consistent with domain size (Figures 3.3b, 3.5, 3.7, 3.8, 3.9 and 3.10).

Second, can we detect variants of these domains in single proteins, e.g. structural modifications arising from point mutations, truncations, and rearrangements? A number of results demonstrate that we can. i) Destabilizing point mutations within titin I27 caused predictable changes to its ionic current pattern (S2-GT^{EE}, Fig. 3.4); ii) Eleven point mutations within GFP^{SF} (S2-G^{SF}T) modified its unfolding dynamics relative to GFP (S2-GT), and allowed us to identify a second unfolding intermediate within GFP^{SF} (Fig. 3.6b, state ix-b); iii) Proteolytic cleavage of S2-G^{TEV}T truncated

S2-G^{TEV}T translocation events at a predicted position within the GFP-dependent ionic current state (ix, Fig. 3.6b); iv) Circular permutations of GFP (S2-G^{CP6}T and S2-G^{CP7}T) resulted in ionic current signatures that differed from the reference protein, and that did not display unfolding intermediates found in S2-G^{SF}T (Fig. 3.6b, states xiii and ix-b); and v) A Naive-Bayes classifier demonstrated our ability to discriminate between these variants (Table 3.1).

3.4.1 Relevance to disease-related protein detection

Three of the protein variants we tested are representative of variant classes commonly associated with disease states[52,53]: i) The destabilizing point mutations we analyzed in titin I27 (S2-GT^{EE}) are similar to point mutations within titin Ig domains that cause cardiomyopathy[65,66]; ii) Mutations that stabilize protein unfolding intermediates (similar to the intermediate observed in superfolder GFP (GFP^{SF})) contribute to diseases such as amyloidosis[67]; and iii) a truncated variant (cleaved S2-G^{TEV}T) demonstrate that the ClpXP-nanopore device can detect truncations derived from early termination or proteolytic processing that are each associated with disease[55]. Additionally, the circular permutants (S2-G^{CP6}T and S2-G^{CP7}T) suggest that we could also detect chimeric oncoproteins[68] and isoforms associated with cancer[54].

3.4.2 Current limitations and summary

One limitation of the current technology is that the polyGSD-ssrA tag needed for capture and ClpXP binding was engineered into the expressed proteins we analyzed. Practical applications will require a method to conjugate the tag to endogenous proteins[69,70]. A second limitation of our current approach is that we are not able to predict a piori the ionic current pattern for a given protein, and must instead rely upon patterns established empirically in each case. However, as the number of analyzed proteins increases, we expect that ionic current patterns will emerge that are characteristic of domain classes. This could facilitate assembly of composite patterns for de novo protein identification.

In summary, an unfoldase-coupled nanopore sensor can discriminate among distinct protein domains and among variants of those domains. Compared to protein mass spectrometry, the ClpXP-nanopore device has the advantages of single-molecule resolution[71], and analysis of unfragmented protein strands. These could be important because an estimated 2/3rds to 4/5ths of eukaryotic proteins are comprised of multiple domains[72], with each domain having potentially many unique modified forms[73].

Chapter 4

PROTEIN TAGGING, BARCODING, AND HIGH-THROUGHPUT NANOPORE ANALYSIS

As mentioned in Chapter 3, several additional steps are required for this method of protein analysis to become generally practical. Principally, these are a technique to tag endogenous proteins with the polyGSD-ssrA tag, and nanopore devices capable of high-throughput single-molecule analysis. Fortunately, solutions to these problems may currently exist (or will soon be available).

4.1 Chemical tagging

Protein analysis with the described unfoldase-nanopore technique requires that target proteins be tagged at the N or C-terminal with a nanopore/ClpX-targeting motif (e.g. the polyGSD-ssrA tag). This was previously accomplished by genetic manipulation

of synthetic protein genes that created fusions containing the polyGSD-ssrA sequence. However, the ultimate goal of this work is to enable analysis of endogenously expressed native proteins. To accomplish this, we have started to develop a generic tagging strategy for proteins derived from biological samples (e.g. cell lysates). This method will post-translationally modify the N-terminal of target proteins with a tag analogous to the polyGSD-ssrA tag used previously.

Specific modification of native protein N-termini will be accomplished via a recently described pyridoxal phosphate (PLP)-mediated reaction in which N-terminal amines are specifically modified to reactive ketones/aldehydes[69,70]. This ketone/aldehyde is then an orthogonal chemical handle that can be specifically conjugated to a synthetic alko-oxyamine-tagged polypeptide (similar in design to the polyGSD-ssrA tag). See Figure 4.1.

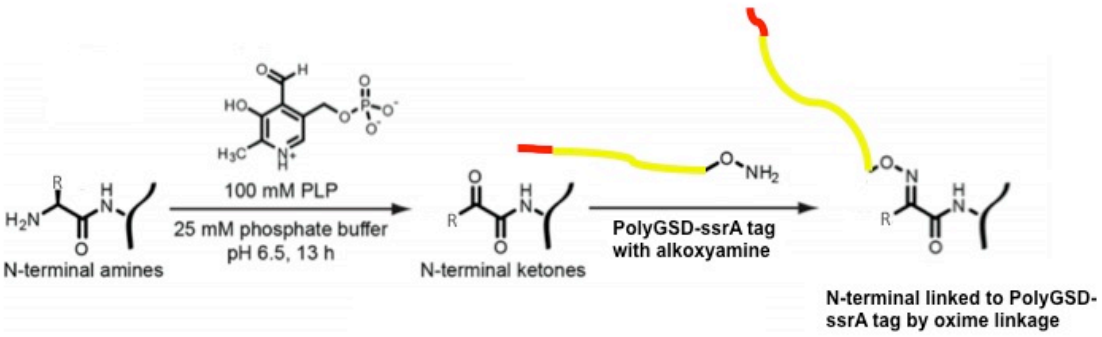


Figure 4.1: **Tagging strategy for endogenous protein analytes.** Endogenous proteins (e.g. proteins from a cell lysate) are reacted with a pyridoxal phoshphate (PLP) solution, transforming native N-terminal amines to reactive ketones or aldehydes. A synthetic PolyGSD-ssrA polypeptide tag containing an alko-oxyamine moiety is then specifically conjugated to the N-termini of transformed endogenous proteins by formation of a stable oxime linkage. Figure adapted from ref. 74.

Work on this tagging method is currently ongoing in the lab.

4.2 Barcode tags for sample multiplexing

Multiplexed assays are advantageous because they analyze multiple species and samples in a single experiment, saving cost and time. The development of multiplex DNA sequencing through tagged or "barcoded" oligo sequences has significantly increased the throughput of sequencing efforts, and is now a widely used concept in molecular sensing technologies[75]. To facilitate multiplexing in this unfoldase-nanopore protein analysis method, I reasoned the polyGSD-ssrA tag could be barcoded. Tagging different protein samples (as discussed previously) with unique polyGSD barcodes would enable sample multiplexing.

4.2.1 PolyGSD barcodes

When a tagged protein is initially captured in the nanopore by voltage, the polyGSD region is held static within the pore. This generates an amino acid sequence-specific ionic current state (e.g. ionic current state ii in the sample traces appearing in the previous chapters). By mutating the amino acid sequence of the region that sits in the pore (Figure 4.2), a specific and distinguishable ionic current state can be observed.

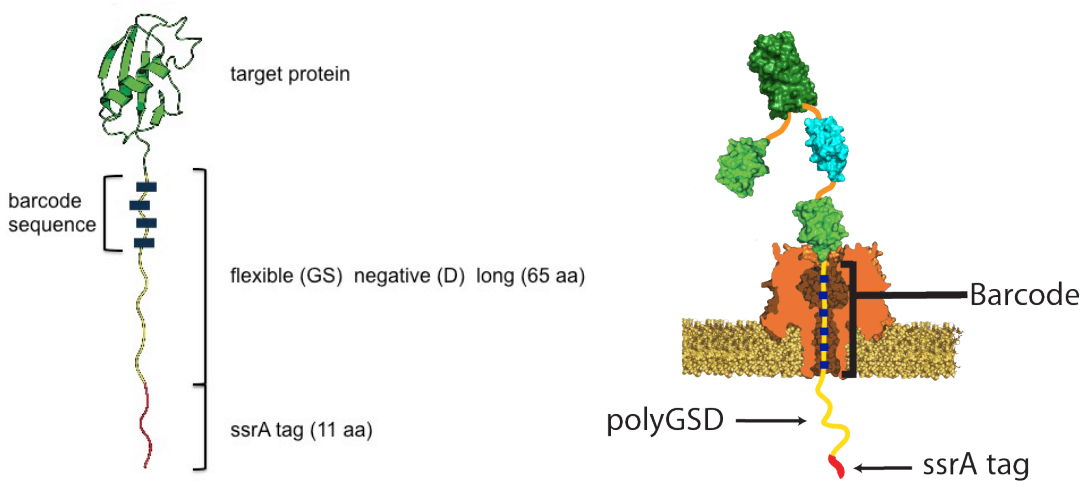


Figure 4.2: Barcoding the polyGSD tag. The polyGSD tag is composed of glycine, serine, and aspartic acid residues, the exact sequence of which is arbitrary. When a polyGSD-tagged protein is captured in the nanopore by voltage, a region of the tag is held static within the pore vestibule (the "barcode sequence"). By mutating the amino acids in this region, sequence-specific ionic current levels can be attained. Different sequences can be considered unique "barcodes." This enables sample multiplexing through conjugation of unique barcode sequences to different samples.

As an initial proof-of-principle, barcoding was implemented by changing the amino acid sequence in the PolyGSD tag at various sites (Figure 4.3).

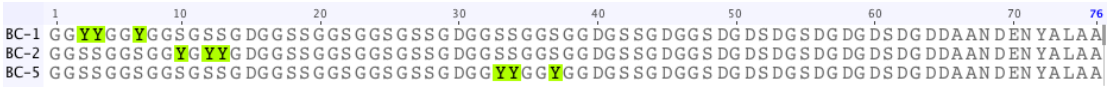


Figure 4.3: Multiple sequence alignment of polyGSD barcodes. The polyGSD tag was mutated (three serine residues were changed to tyrosines) at three different positions to generate three unique barcode sequences (BC-1, BC-2, and BC-5). Each of these barcodes manifests a unique ionic current when it is captured in an αHL nanopore, enabling discrimination among the barcodes.

Three different barcodes (BC-1, BC-2, and BC-5) were generated and tested on the nanopore. Figure 4.4 shows characteristic ionic current traces of voltage-mediated captures of protein S2-GT tagged with the three unique barcodes.

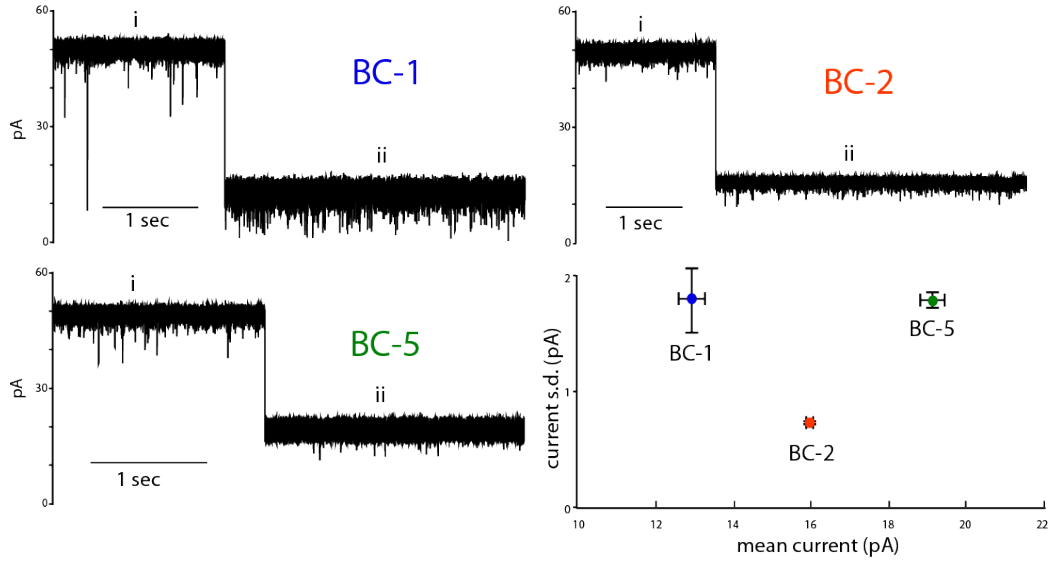


Figure 4.4: PolyGSD barcodes captured in the nanopore generate specific and distinguishable ionic current states. Open channel current through αHL (~50 pA) (i). Capture of the barcoded polyGSD tags in the nanopore (ii). The barcodes can be distinguished from one-another based on the mean ionic current and s.d. of capture state ii.

4.3 High-throughput nanopore analysis

The utility of multiplexing protein samples for analysis using our current nanopore apparatus is marginal; the throughput on our setups is relatively low for these types of experiments (~30 events per experiment hour). However, the rate of data collection can be significantly increased with nanopore array devices that collect data from multiple single channels simultaneously.

4.3.1 Unfoldase-mediated protein translocation through αHL using a MinION

Oxford Nanopore Technologies (ONT), a collaborator with the UCSC Nanopore group, is currently developing nanopore sensor arrays that run hundreds to thousands of protein nanopores in parallel. One of these devices (called a MinION) has recently been released to academic labs for DNA sequencing applications. ONT is also interested in developing protein analysis methods using their nanopore array technology. Towards this end, I performed preliminary experiments in their Oxford-based labs. During this visit, I replicated the unfoldase-αHL system on their nanopore array platform. Example traces for ClpXP-mediated protein translocation of several model proteins using a MinION device are shown in Figure 4.5.

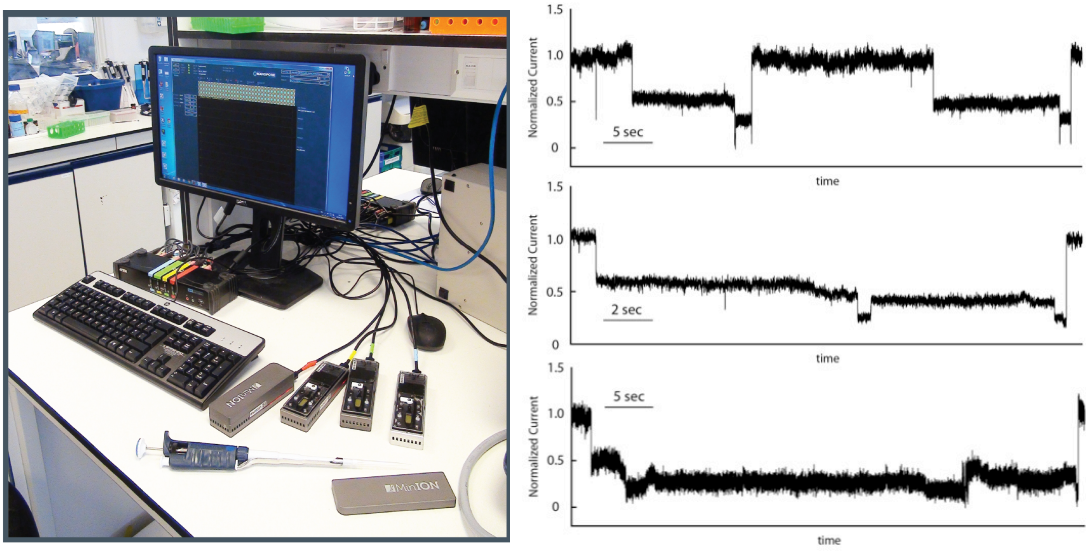


Figure 4.5: ClpXP-mediated nanopore protein analysis using a MinION device. Left: Several MinION devices are shown on an ONT lab bench. Right: Example ionic current traces arising from ClpXP-mediated protein translocation of proteins S2-148 (top and middle), and S2-GT (bottom).

4.4 Towards *de novo* single-molecule protein sequencing

The research described in this dissertation outlines a general method that could be used to identify and count individual proteins in mixtures which is not possible using ensemble methods such as mass spectrometry and immuno-staining. A more ambitious goal would be *de novo* sequencing of single proteins. To achieve this, three technical milestones must be met: 1) more precise control of the protein translocation step size. The ClpX motor typically steps in 1-4nm bursts[61]; 2) a shorter pore lumen that could read 3-to-5 amino acid 'words' as has been employed to read DNA sequence with the MspA pore[17]. The sensitive stem of the αHL pore lumen is about 5 nm long, thus approximately fifteen amino acids contribute to the ionic current impedance that reports strand composition; and 3) bioinformatic algorithms that could identify each of twenty or so amino acid possibilities at each position along a protein strand. This would be a much more difficult problem than identifying the four canonical DNA bases and a small number of epigenetic base variants required for nanopore DNA sequencing.

APPENDICES

APPENDIX A

PROTEIN SEQUENCES USED IN THIS WORK

Model substrate protein sequences used in Chapter 3 are below. Green: Smt3 domains; yellow: charged tail; red: ssrA tag; orange: linker region; black: affinity purification tag regions; blue: additional residues added to obscure the ssrA tag.

S1:

MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKT
TPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGGGSSGGS
GGSGSSGDGGSSGGSGGSGSSGDGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDAAN
DENYALAA

S2-35:

MGHHHHHHGSLQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFA
KRQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQIGGGGSGSGGSGSGGSGSQNE
YRSGGSGSGGSGSGGSGMGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHI
NLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDI
IEAHREQIGGGGSSGGSGGSGSSGDGGSSGGSGGSGSSGDGGSSGGSGGDGSSGDGGSDGDSD
GSDGDGDSDGDDAANDENYALAA

S2-148:

MGHHHHHHGSLQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAK
RQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQIGGGGSGSGGSGSGGSGSQNEYR
SGGGGSGSAGSGASGSSGSEGSGASGSAGSGSAGSRGSGASGSAGSGSAGSGGAEAAKEAAKE
AAKEAAKEAAKAGGSGSAGSAGSASSGSDGSGASGSAGSGSAGSKGSGASGSAGSGSSGSSGG
SGMGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKK
TTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGGGSSGGS
GGSGSSGDGGSSGGSGGSGSSGDGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDAAND
ENYALAA

S1-RQA:

MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKT
TPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGGGSSGGS
GGSGSSGDGGSSGGSGGSGSSGDGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDAAN
DENYALAARQA

Amino acid sequences of substrate proteins used in Chapter 4 are below. Light green, Smt3; dark green, GFP; cyan, titin I27 V15P; yellow, polyGSD tail; red, ssrA tag; black, affinity purification tag or linker regions; blue, TEV protease site.

S2-GT/S2-GT^{EE}:

MGHHHHHHGSLQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFA
KRQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQIGGGGSGSGGSGSGGSGSQNE
YRSGGMRKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPW
PTLVTTFGYGVQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTL
VNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHY
QQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGIEGTASGLIEVEKPLYG
VEVFPGETAHFEIELSEPDVHGQWKLKGQPLAASPD(C/E)EIIEDGKKHILILHN(C/E)QLGMTG
EVSFQAANTKSAANLKVKELGHHHHHHHGAANDENYALAASGGSGMGSSHHHHHHHGSGLVP
RGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKE
MDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGGGSSGGSGGSGSSGDGGSSGGSGGS
GSSGDGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDAANDENYALAA

S2-G^{SF}T:

MGHHHHHHGSLQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFA
KRQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQIGGGGSGSGGSGSGGSGSQNE
YRSGGMRKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPW
PTLVTTLTYGVQCFARYPDHMKQHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKFEGDTL
VNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADHY
QQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAAGIEGTASGLIEVEKPLYG
VEVFPGETAHFEIELSEPDVHGQWKLKGQPLAASPDCEIIEDGKKHILILHNCQLGMTGEVSFQ
AANTKSAANLKVKELGHHHHHHHGAANDENYALAASGGSGMGSSHHHHHHHGSGLVPRGSAS
MSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSL
RFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGGGSSGGSGGSGGSGSSGDGGSSGGSGGSGSSG
DGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDAANDENYALAA

S2-G^{TEV}T:

MGHHHHHHGSLQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFA
KRQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQIGGGGSGSGGSGSGGSGSQNE
YRSGGMRKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPW
PTLVTTLTYGVQCFARYPDHMKQHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKFEGDTL
VNRIELKGIDFKEGGTESGENLYFQGGSGESGGSDGNILGHKLEYNFNSHNVYITADKQKNGI
KANFKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEF
VTAAGIEGTASGLIEVEKPLYGVEVFPGETAHFEIELSEPDVHGQWKLKGQPLAASPDCEIIED
GKKHILILHNCQLGMTGEVSFQAANTKSAANLKVKELGHHHHHHHGAANDENYALAASGGSG
MGSSHHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKT
TPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGGGSSGGS
GGSGSSGDGGSSGGSGGSGSSGDGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDAAN
DENYALAA

S2-G^{CP6}T:

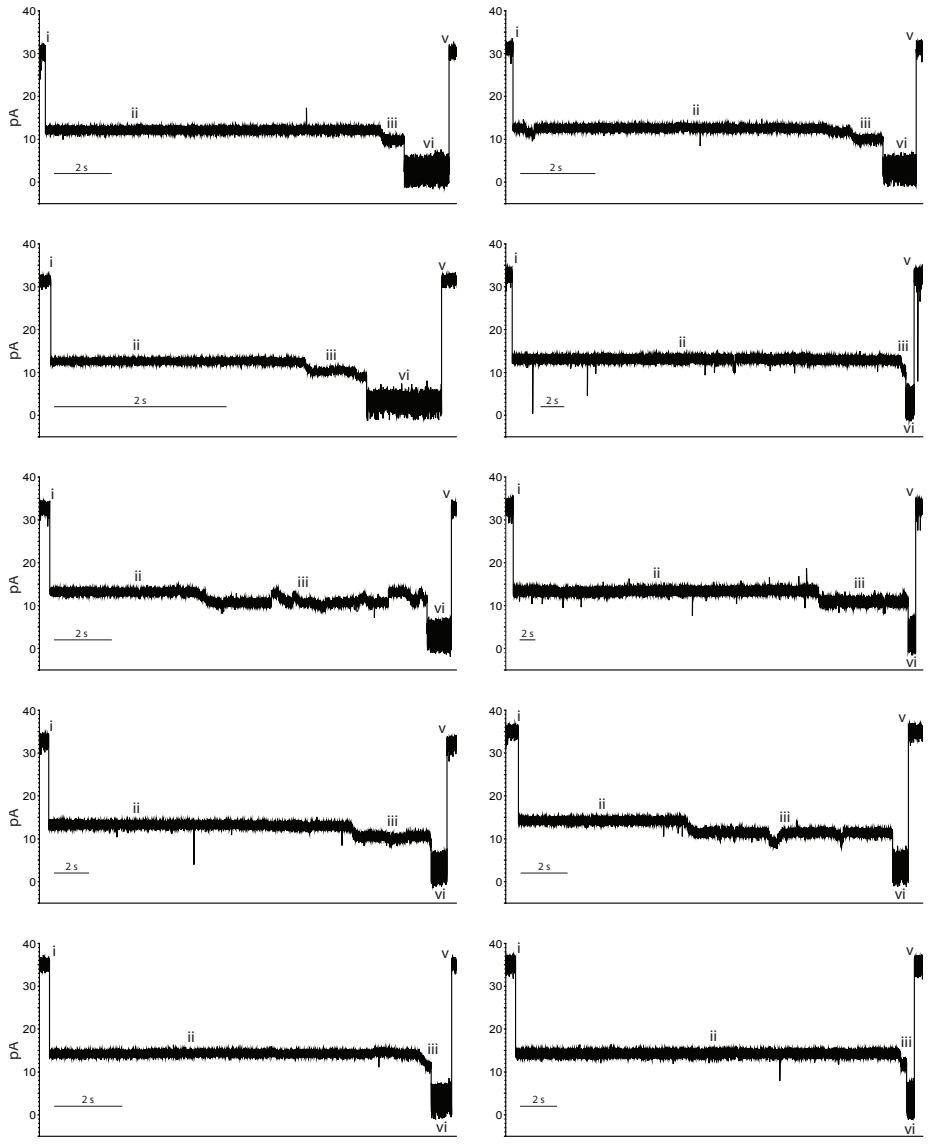
MGHHHHHHGSLQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFA
KRQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQIGGGGSGSGGSGSGGSGSQNE
YRSGGMNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYL
STQSVLSKDPNEKRDHMVLLEFVTAAGITHGMDELYKGGTGGSMRKGEELFTGVVPILVELD
GDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFARYPDHMKQ
HDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEY
NLEGTASGLIEVEKPLYGVEVFPGETAHFEIELSEPDVHGQWKLKGQPLAASPDCEIIEDGKKH
ILILHNCQLGMTGEVSFQAANTKSAANLKVKELGHHHHHHHGAANDENYALAASGGSGMGSS
HHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRR
LMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGGGSSGGSGGSGS
SGDGGSSGGSGGSGSSGDGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDAANDENYA
LAA

S2-G^{CP7}T:

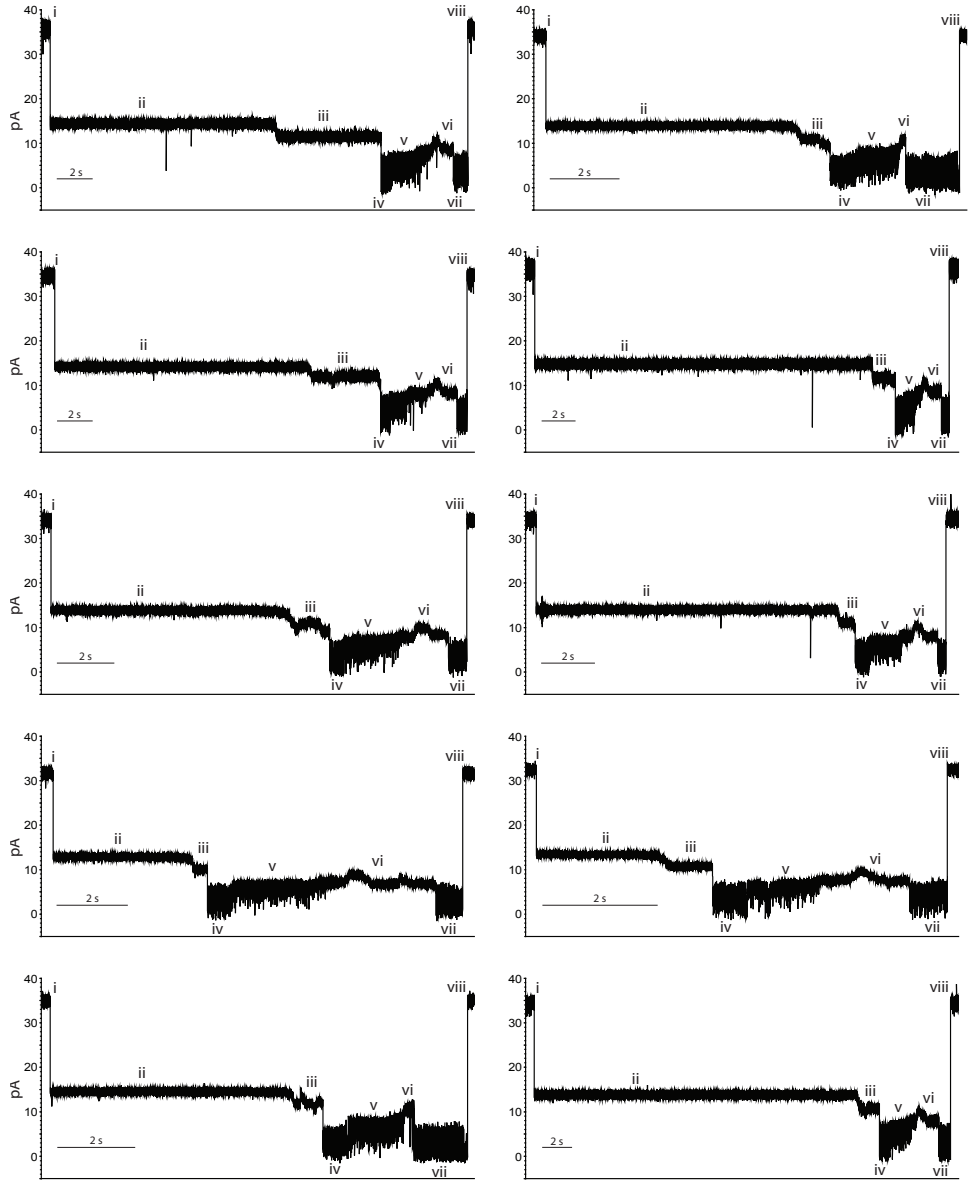
MGHHHHHHGSLQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFA
KRQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQIGGGGSGSGGSGSGGSGSQNE
YRSGGIKQKNGIKANFKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPN
EKRDHMVLLEFVTAAGITHGMDELYKGGTGGSMRKGEELFTGVVPILVELDGDVNGHKFSV
RGEGEGDATNGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFARYPDHMKQHDFFKSAMPEG
YVQERTISFKDDGTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITA
DKEGTASGLIEVEKPLYGVEVFPGETAHFEIELSEPDVHGQWKLKGQPLAASPDCEIIEDGKKH
ILILHNCQLGMTGEVSFQAANTKSAANLKVKELGHHHHHHHGAANDENYALAASGGSGMGSS
HHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRR
LMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGGGSSGGSGGSGS
SGDGGSSGGSGGSGSSGDGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDAANDENYA
LAA

APPENDIX B

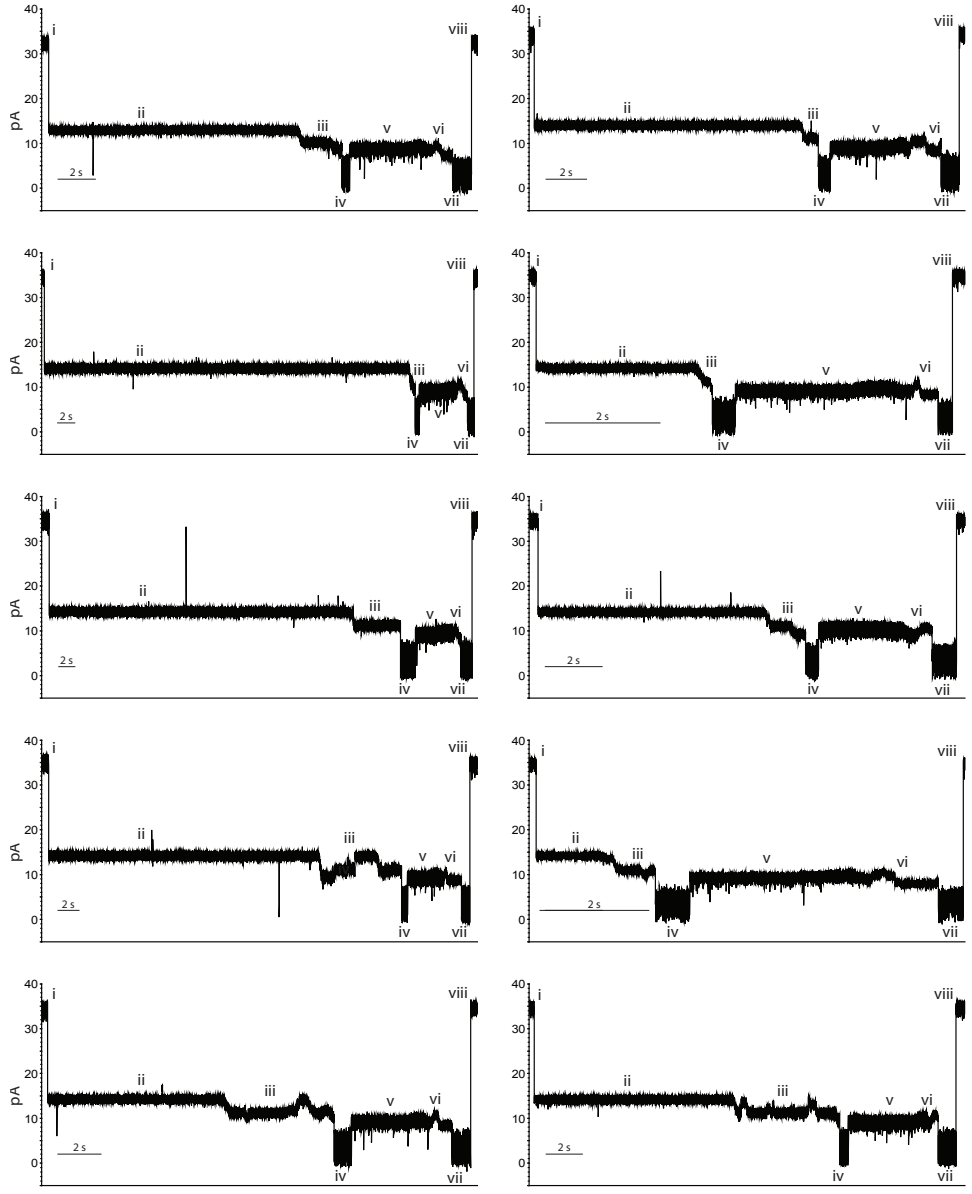
EXAMPLE TRACES



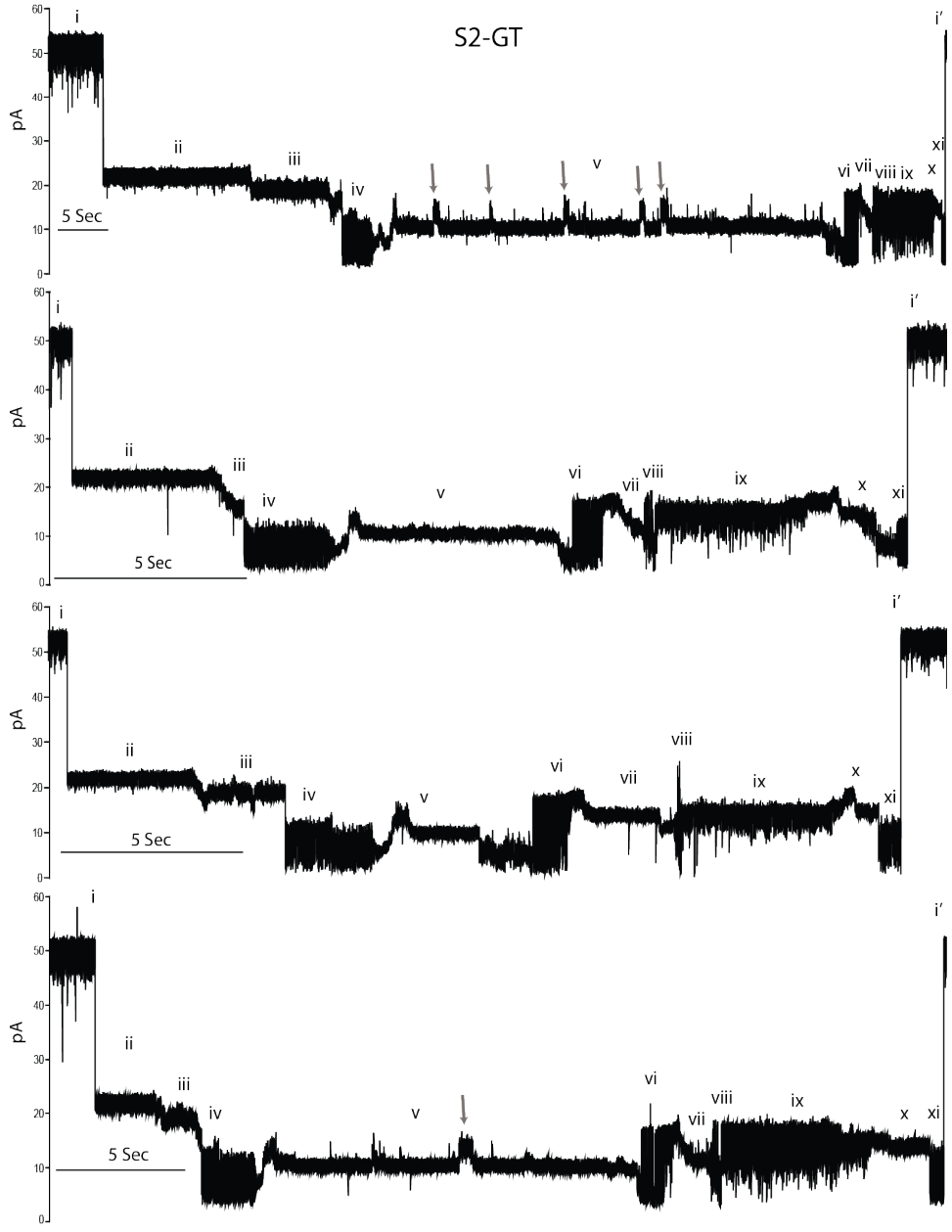
Ten representative ClpX-mediated S1 translocation events with ionic current states i-v labeled.



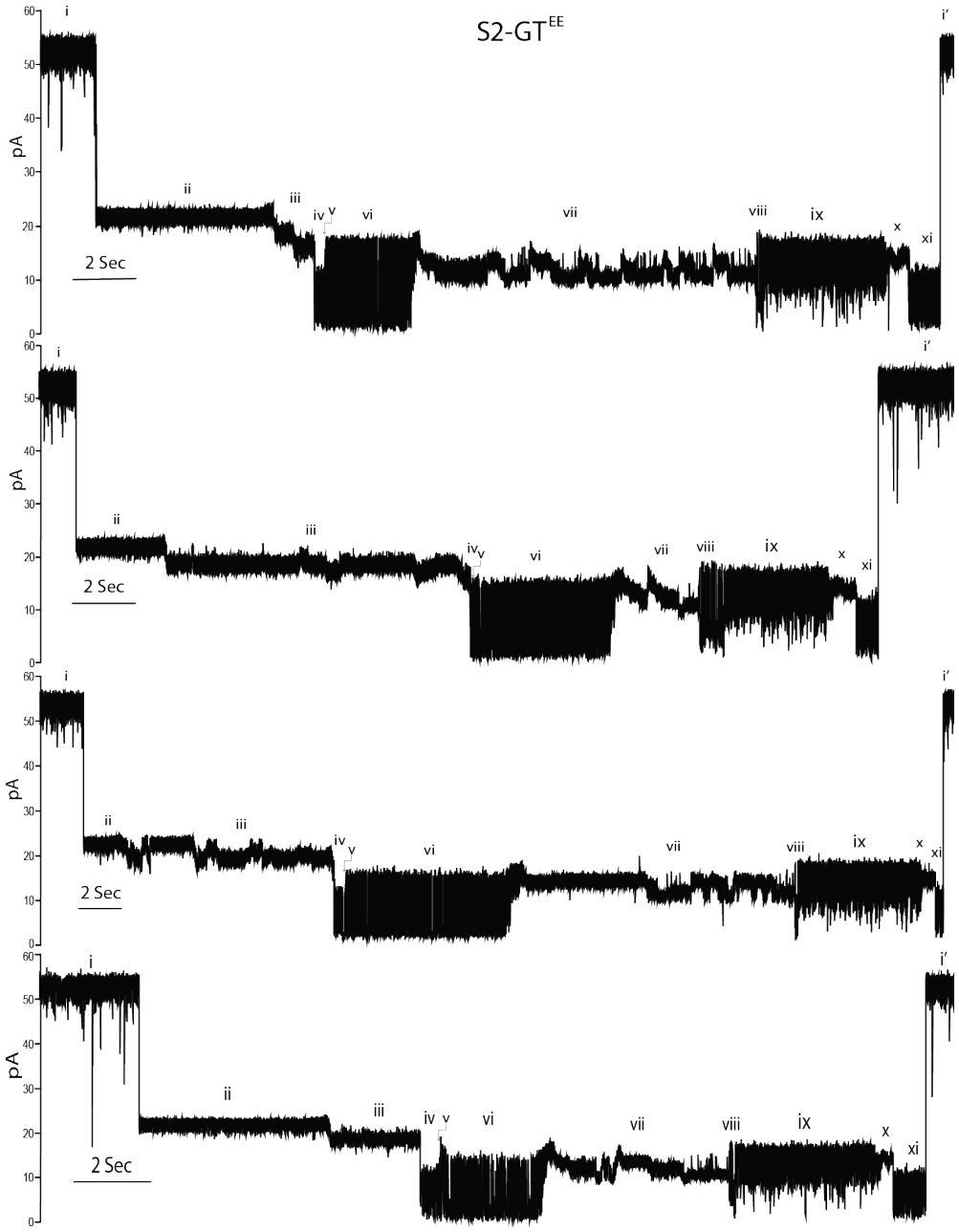
Ten representative ClpX-mediated S2-35 translocation events with ionic current states i-vii labeled.



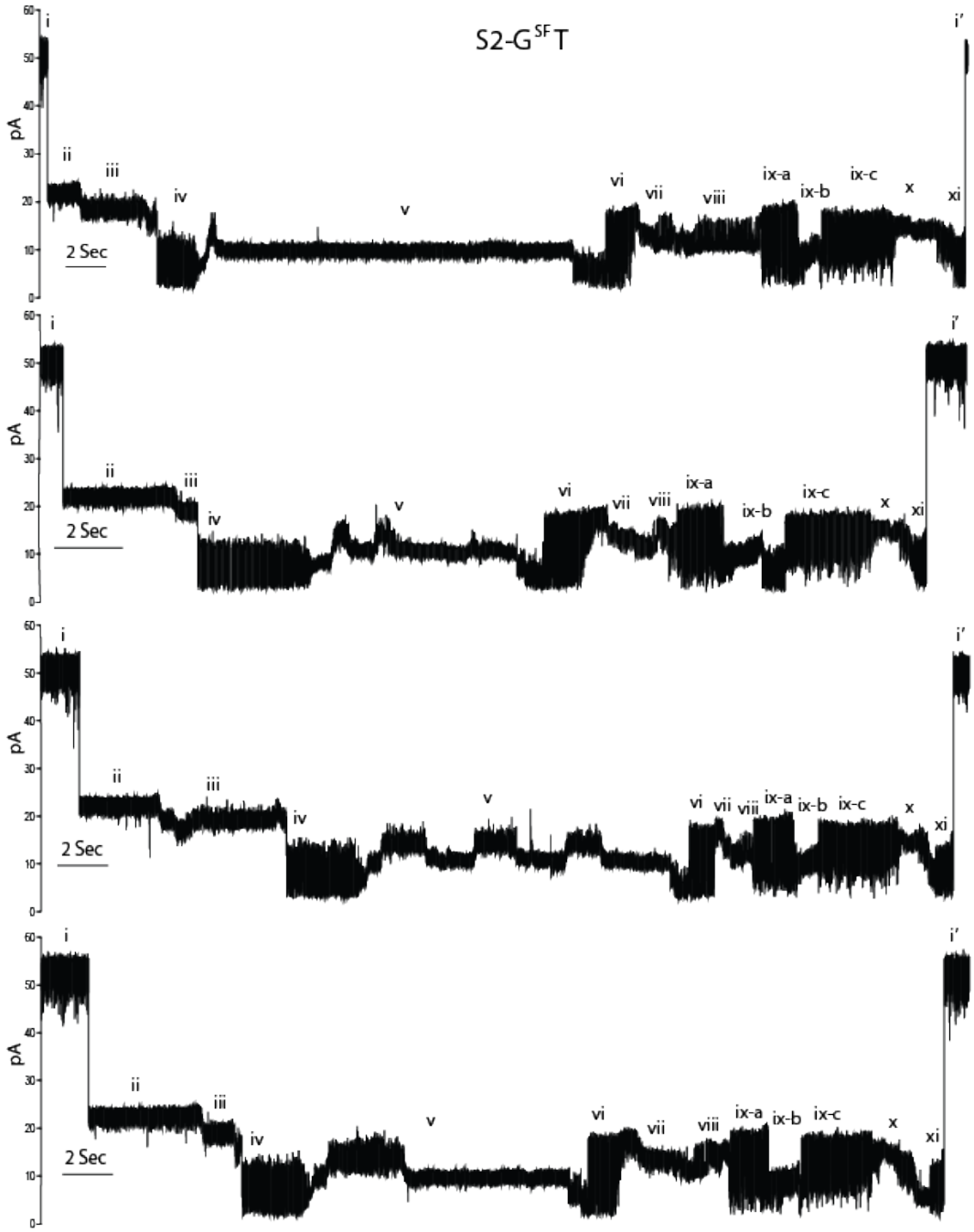
Ten representative ClpX-mediated S2-148 translocation events with ionic current states i-vii labeled.



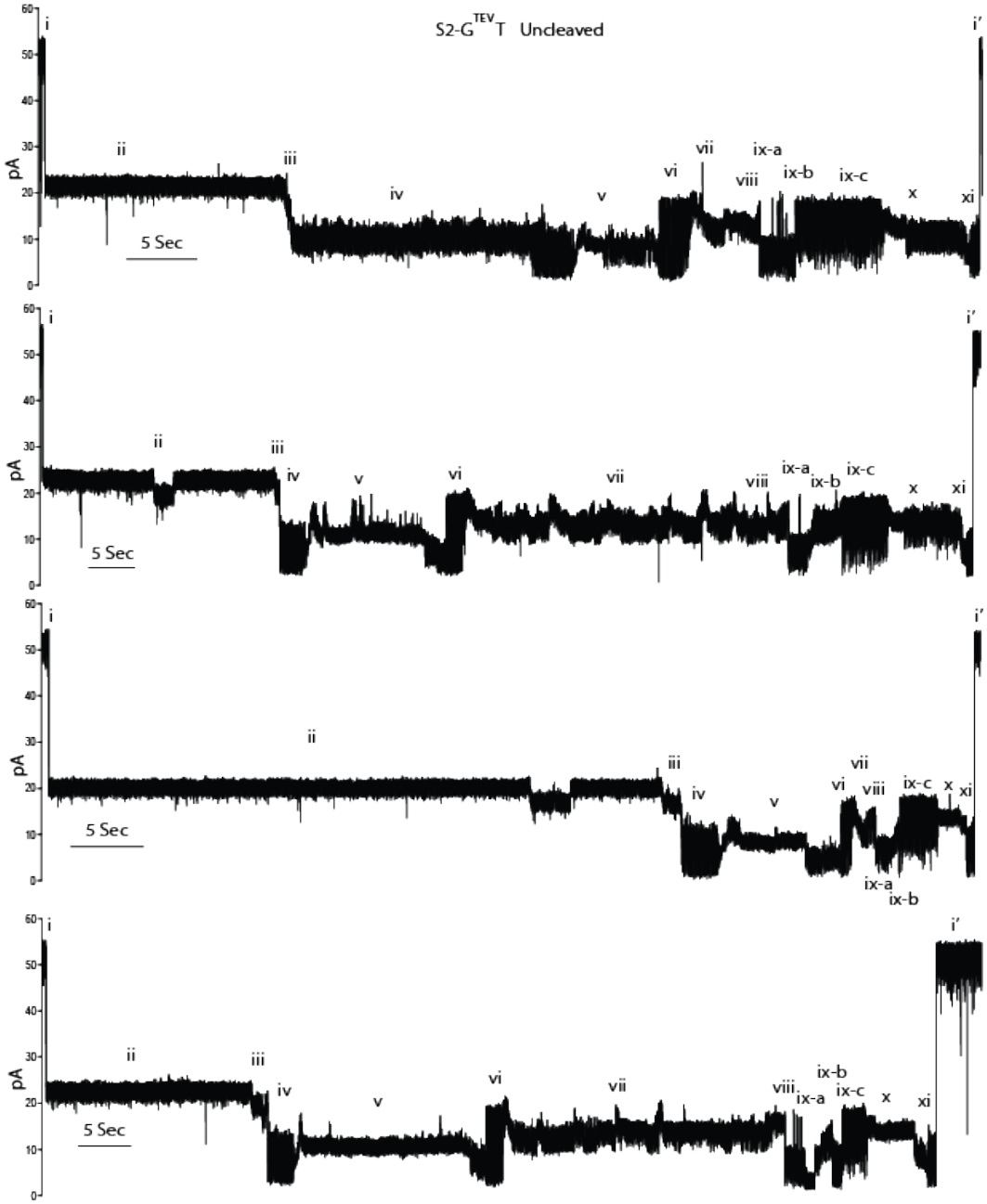
Four representative ClpXP-mediated S2-GT translocation events with ionic current states i-i′ labeled. Open channel (**i**), S2-GT capture (**ii**), Smt3 pre-unfolding (**iii**), Smt3 translocation (**iv**), titin pre-unfolding (**v**), titin translocation (**vi**), GFP pre-unfolding (**vii**), GFP unfolding (**viii**), GFP translocation (**ix**), Smt3 pre-unfolding (**x**), Smt3 translocation (**xi**), and return to open channel (**i′**). Gray arrows highlight current level changes within state **v** that are likely due to enzyme backslips.



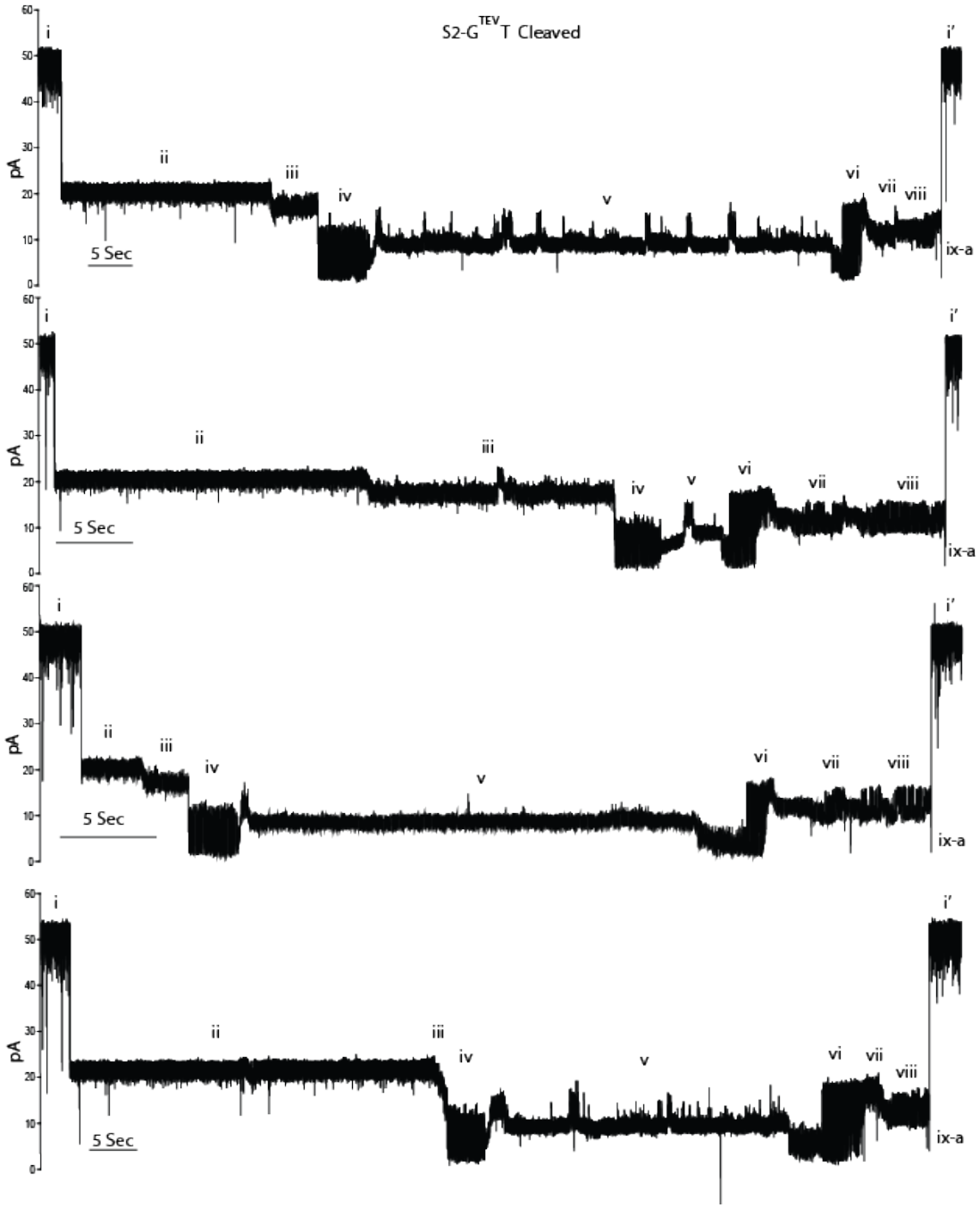
Four representative ClpXP-mediated S2-GT^{EE} translocation events with ionic current states i-i′ labeled. Open channel (**i**), S2-GT^{EE} capture (**ii**), Smt3 pre-unfolding (**iii**), Smt3 translocation (**iv**), titin pre-unfolding (**v**), titin translocation (**vi**), GFP pre-unfolding (**vii**), GFP unfolding (**viii**), GFP translocation (**ix**), Smt3 pre-unfolding (**x**), Smt3 translocation (**xi**), and return to open channel (**i′**).



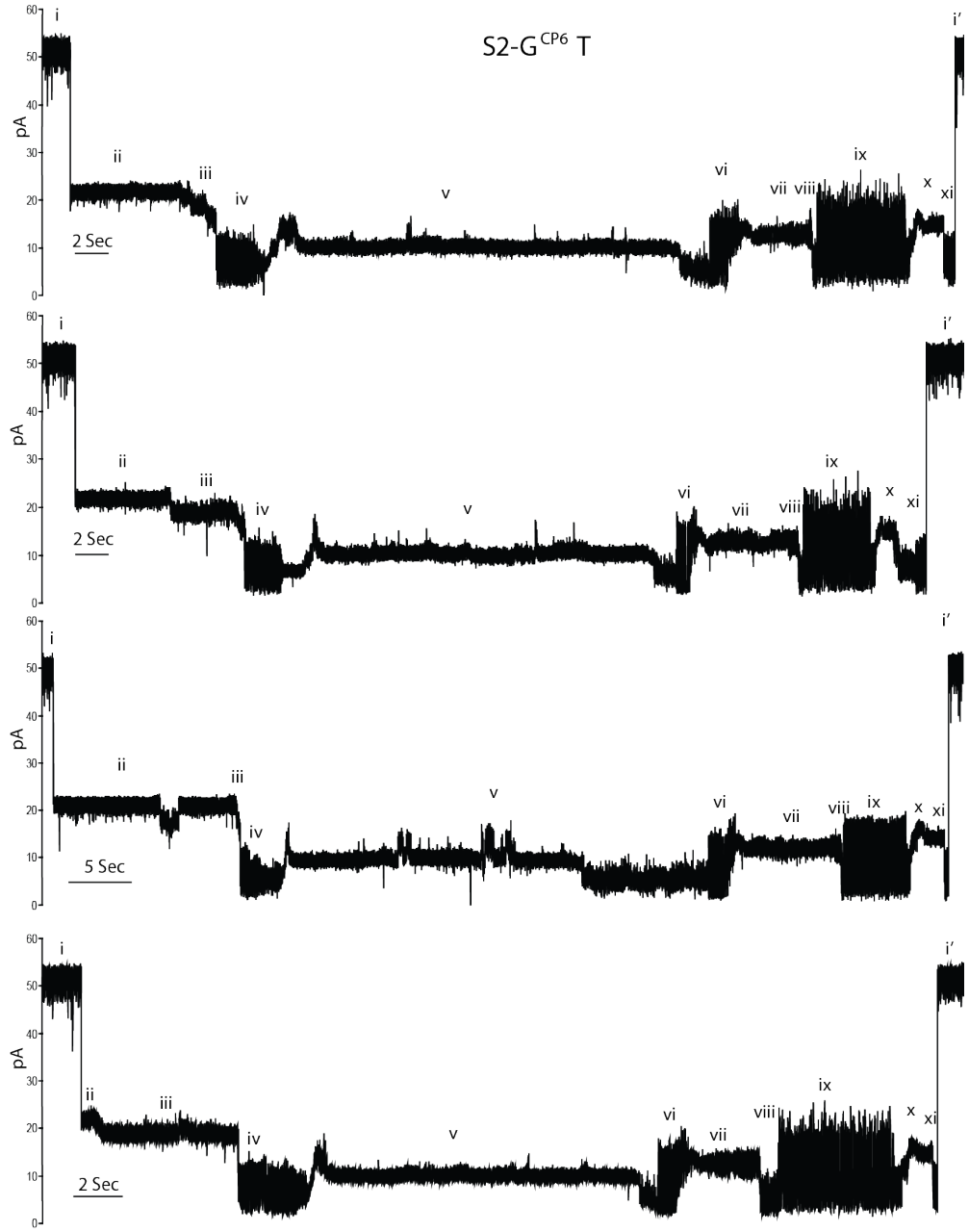
Four representative ClpXP-mediated S2-G^{SF}T translocation events with ionic current states i-i′ labeled. Open channel (**i**), S2-G^{SF}T capture (**ii**), Smt3 pre-unfolding (**iii**), Smt3 translocation (**iv**), titin pre-unfolding (**v**), titin translocation (**vi**), GFP pre-unfolding (**vii**), GFP unfolding (**viii**), GFP translocation (**ix-a**,**b**,**c**), Smt3 pre-unfolding (**x**), Smt3 translocation (**xi**), and return to open channel (**i′**).



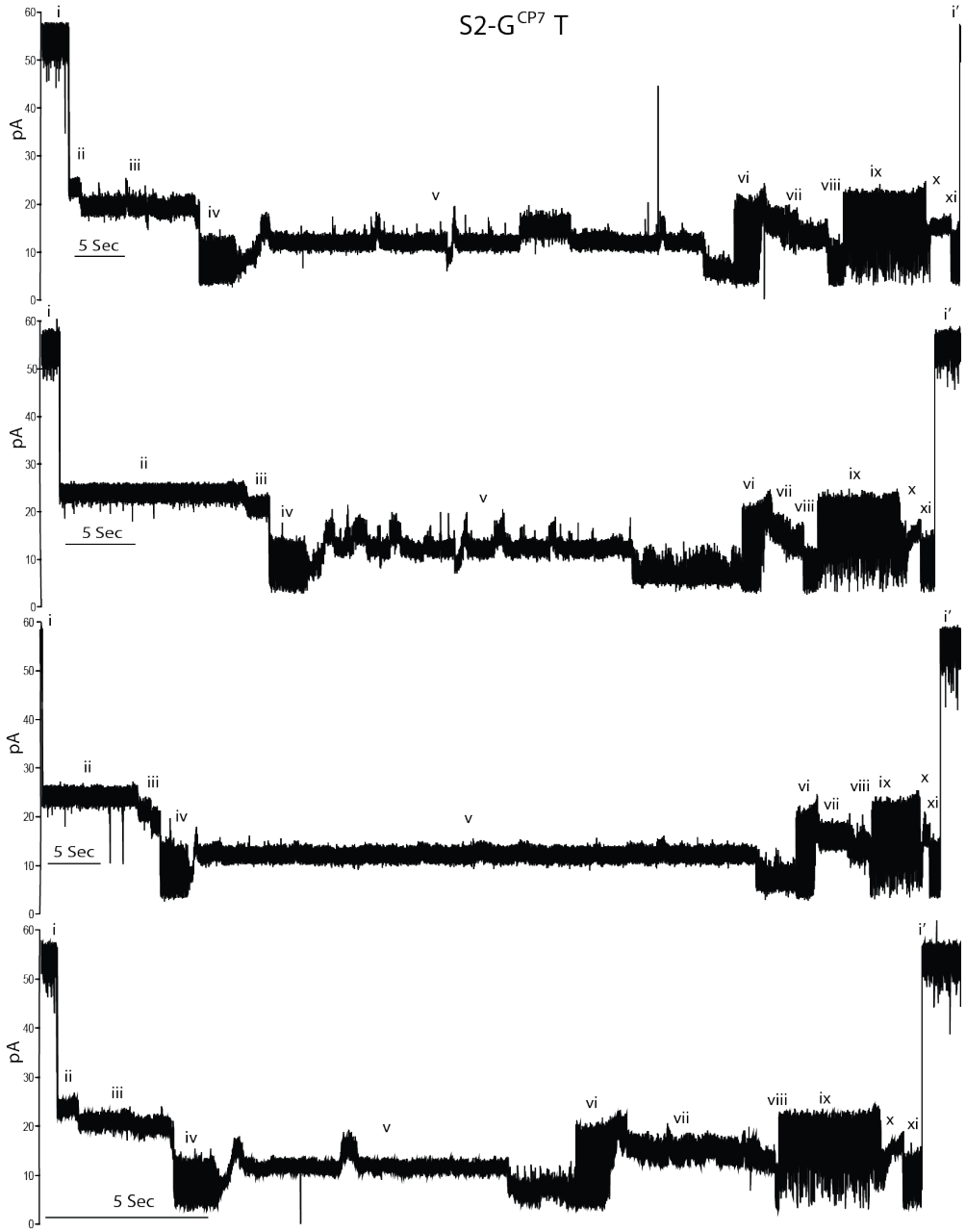
Four representative ClpXP-mediated S2-G^{TEV}T (uncleaved) translocation events with ionic current states i-i′ labeled. Open channel (**i**), S2-G^{TEV}T capture (**ii**), Smt3 pre-unfolding (**iii**), Smt3 translocation (**iv**), titin pre-translocation (**v**), titin translocation (**vi**), GFP pre-unfolding (**vii**), GFP unfolding (**viii**), GFP translocation (**ix-a**,**b**,**c**), Smt3 pre-unfolding (**x**), Smt3 translocation (**xi**), and return to open channel (**i′**).



Four representative ClpXP-mediated S2-G^{TEV}T translocation events (TEV protease cleaved) with ionic current states i-i′ labeled. Open channel (**i**), S2-G^{TEV}T capture (**ii**), Smt3 pre-unfolding (**iii**), Smt3 translocation (**iv**), titin pre-translocation (**v**), titin translocation (**vi**), GFP pre-unfolding (**vii**), GFP unfolding (**viii**), cleaved GFP translocation (**ix-a**), and return to open channel (**i′**).



Four representative ClpXP-mediated S2-G^{CP6}T translocation events with ionic current states i-i′ labeled. Open channel (**i**), S2-G^{CP6}T capture (**ii**), Smt3 pre-unfolding (**iii**), Smt3 translocation (**iv**), titin pre-unfolding (**v**), titin translocation (**vi**), GFP pre-unfolding (**vii**), GFP unfolding (**viii**), GFP translocation (**ix**), Smt3 pre-unfolding (**x**), Smt3 translocation (**xi**), and return to open channel (**i′**).



Four representative ClpXP-mediated S2-G^{CP7}T translocation events with ionic current states i-i′ labeled. Open channel (**i**), S2-G^{CP7}T capture (**ii**), Smt3 pre-unfolding (**iii**), Smt3 translocation (**iv**), titin pre-unfolding (**v**), titin translocation (**vi**), GFP pre-unfolding (**vii**), GFP unfolding (**viii**), GFP translocation (**ix**), Smt3 pre-unfolding (**x**), Smt3 translocation (**xi**), and return to open channel (**i**).

APPENDIX C

ADDITIONAL MATERIALS AND METHODS

ClpX Expression and Purification

A covalently linked trimer of an N-terminal truncated ClpX variant (ClpX-ΔN3) was used for all ClpX nanopore experiments. ClpX protein expression was induced at an A_{600} of ~0.6 by addition of 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and incubated at 23 °C with shaking for 3 h. Cultures were pelleted, resuspended in lysis buffer (50 mM NaH_2PO_4 pH 8, 300 mM NaCl, 100 mM KCl, 20 mM imidazole, 10% glycerol, 1 mM dithiothreitol (DTT)) and lysed by vortexing with glass beads. After centrifugation and filtration of the lysate, the protein was purified on a Ni^{2+}-NTA affinity column (Thermo) and an Uno-Q anion exchange column (Bio-Rad). ClpX was then flash frozen in small aliquots and stored at −80 °C.

ClpP and S2-GT Constructs Expression and Purification

DNA for the GFP and titin I27 domains of fluorescent protein S2-GT were extracted by PCR from a GFP-titin-I27V15P-ssrA expression vector obtained from A. Martin (UC Berkeley), and cloned into the S2-35 vector by Gibson assembly. GFP^{SF}, GFP^{CP6} and GFP^{CP7} DNA was obtained from A. Nager and K. Schmitz (MIT) in the form of expression plasmids and subsequently PCR-extracted and cloned into the S2-GT expression plasmid by replacement of the S2-GT GFP domain via Gibson assembly.

The S2-GT^{EE} and S2-G^{TEV}T mutants were constructed by Gibson assembly using mutagenic oligos and PCR. These engineered proteins and a his-tagged ClpP were expressed in BL21 (DE3)*. Expression was induced at ~0.6 A_{600} by addition of 0.5 mM IPTG, and incubated at 30 °C with shaking for 3-4 h. Cultures were pelleted, resuspended in lysis buffer and lysed via vortexing with glass beads. After centrifugation and filtration of the lysate, the protein was purified on a Ni^{2+}-NTA affinity column (Thermo) and an SD200 size exclusion column (GE). Cleaved protein S2-G^{TEV}T was digested with TurboTEV (Nacalai USA) for 24 hours at room temperature. The proteins were flash frozen in small aliquots following purification and stored at −80 °C.

Feature Selection*

For every event, the mean, standard deviation, and duration of states ii through xi were taken, resulting in thirty features and 417 events. The identity of each event is also known, since data on each protein variant were collected individually. The Gini importance was calculated for each feature using a forest of extremely randomized trees. Features that performed higher than a null model assuming equal importance for each feature were kept. The open-source scikit-learn[43] (version 0.14.1) command used to build the forest of extremely randomized trees was as follows:

ExtraTreesClassifier (n_estimators=500, max_features=6, max_depth=None, min_samples_split=1, random_state=42)

Naive Bayes Classification*

Using feature selection, eight features were selected automatically as being useful in a five-class classification, with each protein variant being a class. We estimated the error rate using stratified five-fold cross validation with a Gaussian Naive Bayes classifier. Simply, the classifier is trained on 80 percent of the data, and predicts labels for the remaining 20 percent. This process is repeated four more times using a unique 20 percent each time, predicting labels for each event. This strategy was performed a total of 20 times to ensure the accuracy measurement was not an outlier. A confusion matrix is generated by comparing the predicted labels to the known identity of each event and using Maximum A Posteriori (MAP) estimates with a symmetric Dirichlet prior in which one is added to each of the counts before normalization to get probabilities for each cell.

* Jacob Schreiber implemented the random forest analysis and the Naive Bayes classifier

REFERENCES

1. Anderson, N.L. & Anderson, N.G. Proteome and proteomics: New technologies, new concepts, and new words. *Electrophoresis* **19**, 1853-1861 (1998).

2. Horgan, R.P. & Kenny, L.C. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *TOG* **13**, 189-195 (2011).

3. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat Reviews* **5**, (2004).

4. Altelaar, A.F., Munoz, J., & Heck, A.J. Next-generation of proteomics: towards an integrative view of proteome dynamics. *Nat Reviews* **14**, (2013).

5. Ma, B. & Johnson, R. De novo sequencing and homology searching. *Mol and Cell Proteomics* **11.2**, (2012).

6. Han, X., Aslanian, A., & Yates, J.R. Mass spectrometry for proteomics. *Curr Opin in Chem Biol* **12**, (2008).

7. Bandeira, N. *et al*. Automated de novo protein sequencing of monoclonal antibodies. *Nat Biotechnol.* **26**, (2008).

8. Lubec, G. & Afjehi-Sadat, L. Limitations and pitfalls in protein identification by mass spectrometry. *Chem. Rev*. **107**, (2007).

9. Hood, L.E. *et al*. New and improved proteomics technologies for understanding complex biological systems: Addressing a grand challenge in the life sciences. *Proteomics* **12**, (2012).

10. Coulter, W.H. Means for counting particles suspended in a fluid. US patent 2,646,508 (1953).

11. Church, G.M., Deamer, D.W., Branton, D., Baldarelli, R. & Kasianowicz, J. Characterization of individual polymer molecules based on monomer-interface interaction. US patent 5,795,782 (1998).

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

13. Hayden, E.C. Nanopore genome sequencer makes it debut. *Nature*, published online 17 February.

14. Pennisi, E. Search for pore-fection. *Science* **336**, (2012).

15. Cherf, G.M. *et al*. Automated forward and reverse ratcheting of DNA at 5-A precision. *Nat. Biotechnol.* **30**, (2012).

16. Howorka, S. & Siwy, Z. Nanopore analytics: sensing of single molecules. *Chem. Soc. Rev.* **38**, (2009).

17. Manrao, E.A. *et al*. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* **30**, 349-353 (2012).

18. Laszlo, A.H. *et al*. Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol*. **32**, 829-833 (2014).

19. Olasagasti, F. *et al*. Replication of individual DNA molecules under electronic control using a protein nanopore. *Nat. Nanotechnol.* **5**, 798-806 (2010).

20. Movileanu, L.; Howorka, S.; Braha, O.; Bayley, H. Detecting Protein Analytes That Modulate Transmembrane Movement of a Polymer Chain Within a Single Protein Pore. *Nat. Biotechnol.* **18**, 1091–1095 (2000).

21. Mohammad, M. M.; Prakash, S.; Matouschek, A.; Movileanu, L. Controlling a Single Protein in a Nanopore through Electrostatic Traps. *J. Am. Chem. Soc.* **130**, 4081–4088 (2008).

22. Jin, Q. *et al*. Base-Excision Repair Activity of Uracil-DNA Glycosylase Monitored Using the Latch Zone of α-Hemolysin. *J. Am. Chem. Soc.* **135**, 19347–19353 (2013).

23. Lieberman, K. R. *et al*. Processive Replication of Single DNA Molecules in a Nanopore Catalyzed by Phi29 DNA Polymerase. *J. Am. Chem. Soc.* **132**, 17961–17972 (2010).

24. Harrington, L.; Cheley, S.; Alexander, L. T.; Knapp, S.; Bayley, H. Stochastic Detection of Pim Protein Kinases Reveals Electrostatically Enhanced Association of a Peptide Substrate. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4417–26 (2013).

25. Rodriguez-Larrea, D.; Bayley, H. Multistep Protein Unfolding During Nanopore Translocation. *Nat. Nanotechnol.* **8**, 288–295 (2013).

26. Rosen, C. B.; Rodriguez-Larrea, D.; Bayley, H. Single-Molecule Site-Specific Detection of Protein Phosphorylation with a Nanopore. *Nat. Biotechnol.* **32**, 179–181 (2014).

27. Talaga, D. S.; Li, J. Single-Molecule Protein Unfolding in Solid State Nanopores. *J. Am. Chem. Soc.* **131**, 9287–9297 (2009).

28. Merstorf, C.; Cressiot, B.; Pastoriza-Gallego, M.; Oukhaled, A.; Betton, J.-M.; Auvray, L.; Pelta, J. Wild Type, Mutant Protein Unfolding and Phase Transition Detected by Single-Nanopore Recording. *ACS Chem. Biol.* **7**, 652–658 (2012).

29. Bell, N.A.W. *et al.* DNA origami nanopores. *Nano Lett.* **12**, 512-517 (2012).

30. Dekker, C. Solid-state nanopores. *Nat. Nanotechnol.* **2**, 209-215 (2007).

31. Maglia, G., Restrepo, M.R., Mikhailova, E., & Bayley, H. Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 19720-19725 (2008).

32. Bhakdi, S. & Tranum-Jensen, J. Alpha-toxin of Staphylococcus aureus. *Microbiol. Rev.* **55**, 733-751 (1991).

33. Kasianowicz, J., Brandin, E., Branton, D., & Deamer, D. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 1377-13773 (1996).

34. Stoddart, D. *et al*. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7702-7707 (2009).

35. Butler, T.Z. *et al*. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 20647-20652 (2008).

36. Manrao, E. *et al.* Nucleotide discrimination with DNA immobilized in the MspA nanopore. *PLoS ONE*. **6**, (2011).

37. Baker, T.A. & Sauer, R.T. ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochim. Biophys. Acta* **1823**, 15-28 (2012).

38. Keiler, K.C. Biology of trans-translation. *Annu. Rev. Microbiol*. **62**, 133-151 (2008).

39. Neuwald, A.F., Aravind, L., Spouge J.L., & Koonin, E.V. AAA+: A class of chaperone-like ATPases associated with assembly, operation, and disassembly of protein complexes. *Genome Res.* **9**, 27-43 (1999).

40. Aubin-Tam, M.E. *et al.* Single-molecule protein unfolding and translocation by an ATP-fueled proteolytic machine. *Cell* **145**, 257-267 (2011).

41. Maillard, R.A. *et al*. ClpX(P) generates mechanical force to unfold and translocate its protein substrates. *Cell* **145**, 459-469 (2011).

42. Johnson, E.S., Schweinhorst, I., Dohmen, R.J. & Blobel, G. The ubiquitin-like protein Smt3p is activated for conjugation to other proteins by an Aos1p/Uba2p heterodimer. *EMBO J*. **16**, 5509-5519 (1997).

43. Sheng, W. & Liao, X. Solution structure of a yeast ubiquitin-like protein Smt3: the role of structurally less defined sequences in protein-protein recognitions. *Protein Sci*. **11**, 1482-1491 (2002).

44. Gottesman, S., Roche E., Zhou, Y. & Sauer, R.T. The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes Dev.* **12**, 1338-1347 (1998).

45. Flynn, J.M. *et al*. Overlapping recognition determinants within the ssrA degradation tag allow modulation of proteolysis. *Proc. Natl. Acad. Sci. USA* **98**, 10584-10589 (2001).

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

47. Talaga, D.S. & Li, J. Single-molecule protein unfolding in solid state nanopores. *J. Am. Chem. Soc.* **131**, 9287-9297 (2009).

48. Merstorf, C. *et al*. Wild type, mutant protein unfolding and phase transition detected by single-nanopore recording. *ACS Chem. Biol.* **7**, 652-658 (2012).

49. Christensen, C. *et al*. Effect of charge, topology and orientation of the electric field on the interaction of peptides with the α-hemolysin pore. *J. Pept. Sci.* **17**, 726-734 (2011).

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

51. Oukhaled, G. *et al*. Unfolding of proteins and long transient conformations detected by single nanopore recording. *Phys. Rev. Lett.* **98**, 158101 (2007).

52. Bross, P.; Corydon, T. J.; Andresen, B. S.; Jørgensen, M. M.; Bolund, L.; Gregersen, N. Protein Misfolding and Degradation in Genetic Diseases. *Hum. Mutat.* **14**, 186–198 (1999).

53. Ferrer-Costa, C.; Orozco, M.; de la Cruz, X. Characterization of Disease-Associated Single Amino Acid Polymorphisms in Terms of Sequence and Structure Properties. *J. Mol. Biol. 315*, 771–786 (2002).

54. Wei, J.; Zaika, E.; Zaika, A. P53 Family: Role of Protein Isoforms in Human Cancer. *J. Nucleic Acids* **2012**, 687359 (2012).

55. García-Sierra, F.; Mondragón-Rodríguez, S.; Basurto-Islas, G. Truncation of Tau Protein and Its Pathological Significance in Alzheimer's Disease. *J. Alzheimers. Dis.* **14**, 401–409 (2008).

56. Improta, S.; Politou, A. S.; Pastore, A. Immunoglobulin-Like Modules from Titin I-Band: Extensible Components of Muscle Elasticity. *Structure* **4**, 323–337 (1996).

57. Ormö, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. Crystal Structure of the Aequorea Victoria Green Fluorescent Protein. *Science* **273**, 1392–1395 (1996).

58. Kenniston, J. A.; Baker, T. A.; Fernandez, J. M.; Sauer, R. T. Linkage Between ATP Consumption and Mechanical Unfolding During the Protein Processing Reactions of an AAA+ Degradation Machine. *Cell* **114**, 511–520 (2003).

59. Martin, A.; Baker, T. A.; Sauer, R. T. Protein Unfolding by a AAA+ Protease Is Dependent on ATP-Hydrolysis Rates and Substrate Energy Landscapes. *Nat. Struct. Mol. Biol.* **15**, 139–145 (2008).

60. Gur, E.; Sauer, R. T. Recognition of Misfolded Proteins by Lon, a AAA(+) Protease. *Genes Dev.* **22**, 2267–2277 (2008).

61. Sen, M.; Maillard, R. A.; Nyquist, K.; Rodriguez-Aliaga, P.; Pressé, S.; Martin, A.; Bustamante, C. The ClpXP Protease Unfolds Substrates Using a Constant Rate of Pulling but Different Gears. *Cell* **155**, 636–646 (2013).

62. Pédelacq, J.-D.; Cabantous, S.; Tran, T.; Terwilliger, T. C.; Waldo, G. S. Engineering and Characterization of a Superfolder Green Fluorescent Protein. *Nat. Biotechnol.* **24**, 79–88 (2006).

63. Storm, A. J.; Storm, C.; Chen, J.; Zandbergen, H.; Joanny, J.-F.; Dekker, C. Fast DNA Translocation through a Solid-State Nanopore. *Nano Lett.* **5**, 1193–1197 (2005).

64. Zhang, H. The Optimality of Naive Bayes. In *FLAIRS Conference*; AAAI Press: Miami Beach; 562–567 (2004).

65. Herman, D. S.; Lam, L.; Taylor, M. R. G.; Wang, L.; Teekakirikul, P.; Christodoulou, D.; Conner, L.; DePalma, S. R.; McDonough, B.; Sparks, E.; *et al.* Truncations of Titin Causing Dilated Cardiomyopathy. *N. Engl. J. Med.* **366**, 619–628 (2012).

66. LeWinter, M. M.; Granzier, H. L. Titin Is a Major Human Disease Gene. *Circulation* **127**, 938–944 (2013).

67. Ma, B.; Nussinov, R. The Stability of Monomeric Intermediates Controls Amyloid Formation: Abeta25-35 and Its N27Q Mutant. *Biophys. J.* **90**, 3365–3374 (2006).

68. Croce, C. M. Oncogenes and Cancer. *N. Engl. J. Med.* **358**, 502–511 (2008).

69. Gilmore, J. M.; Scheck, R. A.; Esser-Kahn, A. P.; Joshi, N. S.; Francis, M. B. N-Terminal Protein Modification through a Biomimetic Transamination Reaction. *Angew. Chem. Int. Ed. Engl.* **45**, 5307–5311 (2006).

70. Scheck, R. A.; Francis, M. B. Regioselective Labeling of Antibodies through N-Terminal Transamination. *ACS Chem. Biol.* **2**, 247–251 (2007).

71. Lubec, G.; Afjehi-Sadat, L. Limitations and Pitfalls in Protein Identification by Mass Spectrometry. *Chem. Rev.* **107**, 3568–3584 (2007).

72. Han, J.-H.; Batey, S.; Nickson, A. A.; Teichmann, S. A.; Clarke, J. The Folding and Evolution of Multidomain Proteins. *Nat. Rev. Mol. Cell Biol.* **8**, 319–330 (2007).

73. Prabakaran, S.; Lippens, G.; Steen, H.; Gunawardena, J. Post-Translational Modification: Nature's Escape from Genetic Imprisonment and the Basis for Dynamic Information Encoding. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **4**, 565–583 (2012).

74. Carrico, Z.M. *et al*. N-terminal labeling of filamentous phage to create cancer marker imaging agents. *ACS Nano* **6**, 6675-6680 (2012).

75. Church, G.M. & Kieffer-Higgins, S. Multiplex DNA Sequencing. *Science* **240**, 185-188 (1988).