Peptide Fingerprinting Using Single-Molecule Fluorescence

van Ginkel, Jetty

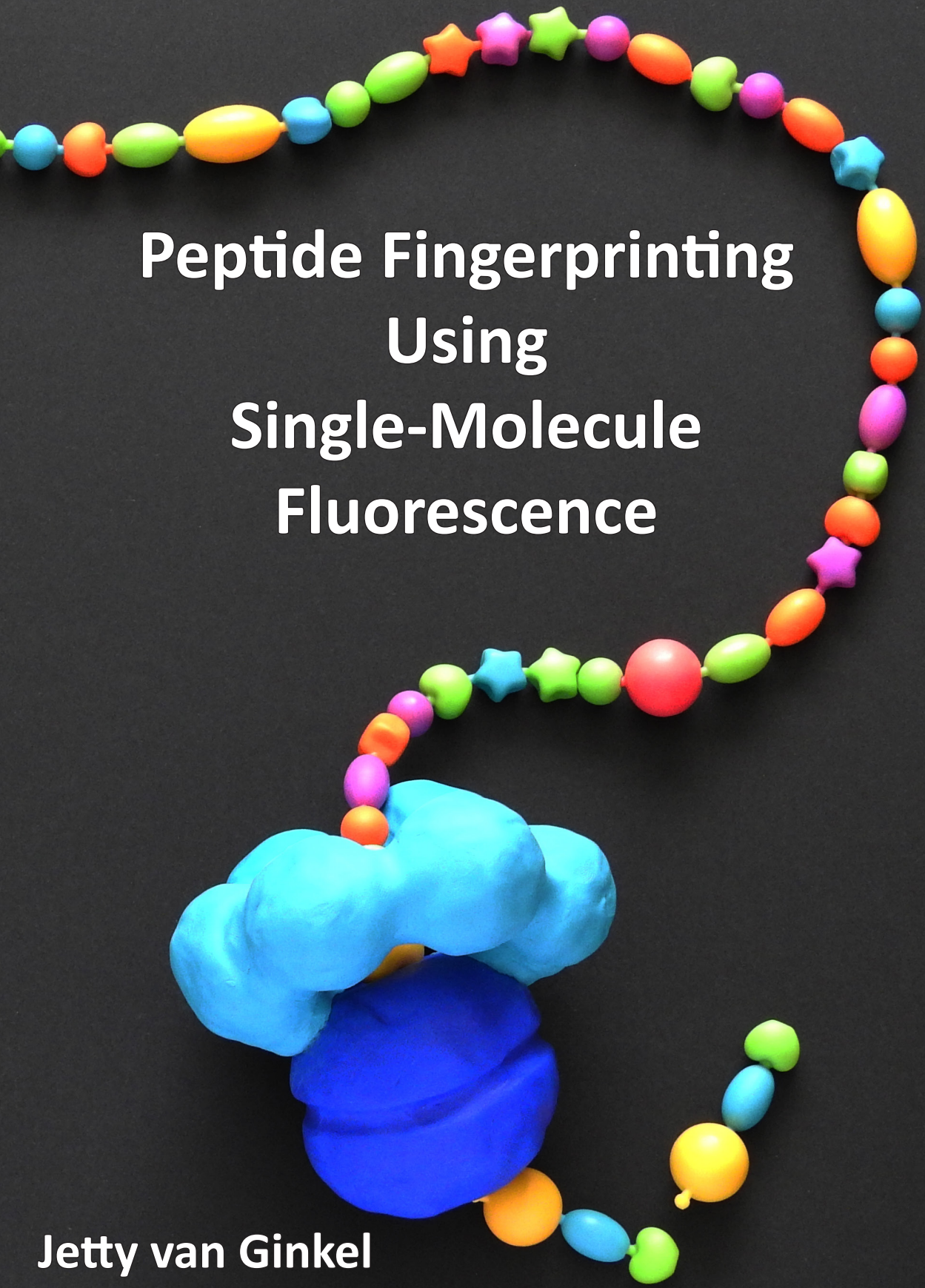
DOI
10.4233/uuid:894b72df-6a38-4e6b-b2d2-1dd740ba92db

Publication date
2016

Document Version
Final published version

Citation (APA)
van Ginkel, J. (2016). *Peptide Fingerprinting Using Single-Molecule Fluorescence*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:894b72df-6a38-4e6b-b2d2-1dd740ba92db

Peptide Fingerprinting Using Single-Molecule Fluorescence



Jetty van Ginkel

Peptide Fingerprinting Using Single-Molecule Fluorescence

Hendrika Geertruida Theodora Maria VAN GINKEL

Peptide Fingerprinting Using Single-Molecule Fluorescence

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.ir. K.C.A.M. Luyben;
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
dinsdag 20 december 2016 om 15:00 uur

door

Hendrika Geertruida Theodora Maria VAN GINKEL

Master of Science in de Biomedische Technologie
Universiteit Twente, Nederland
geboren te Gouda, Nederland

Dit proefschrift is goedgekeurd door:

Promotor:
Prof. dr. C. Dekker Technische Universiteit Delft
Copromotor:
Dr. C. Joo Technische Universiteit Delft

Samenstelling promotie commissie:
Rector Magnificus, Voorzitter
Prof. dr. C. Dekker Promotor
Dr. C. Joo Copromotor

Onafhankelijke leden:
Prof. dr. G. Maglia Rijksuniversiteit Groningen
Prof. dr. M. Wuhrer Leids Universitair Medisch Centrum
Prof. dr. A.H. Engel Technische Universiteit Delft
Dr. D. Dulin Friedrich-Alexander Universität Erlangen-Nührnberg
Dr. A.S. Meyer Technische Universiteit Delft
Prof. dr. M. Dogterom Technische Universiteit Delft, reservelid

Bionanoscience Department
Think big about life at the smallest scale

Keywords: Protein sequencing; proteomics; single-molecule fluorescence; ClpXP
Printed by: Gildeprint
Front & Back: Bart van Manen & Jetty van Ginkel

Casimir PhD Series, Delft - Leiden 2016-37
ISBN: 978-90-8593-281-9

An electronic version of this dissertation is available at:
http://repository.tudelft.nl/

Contents

Chapter 1

The Road to Single-Molecule Protein Sequencing

1.1 Introduction

Proteins are vital building blocks to maintain life; consequently critical information on biological processes is hidden in the proteome. Proteomics can provide valuable information on molecular pathways and state of health. The past two decades the proteomics field was propelled by a combination of two ionization techniques, matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI), and advances in database search algorithms such as SEQUEST and MASCOT (1) leading to shotgun proteomics. Two independent research groups profited from these developments and drafted the first maps of the human proteome (2,3). This map provides tremendous insight for the biological community and ultimately has the potential to transform personalized medicine (4).

With current mass spectrometry based techniques – the method of choice for characterizing complex protein samples – these types of large scale studies remain an enormous effort. The requirement of large amount of sample and the limit of detection precludes detection of low abundant proteins and renders single-cell analysis impossible. To give an example: typically a cerebral spinal fluid sample has a volume of 3 mL and contains less than 65 fmol of the combined splice variants and many different modified forms of tau, a marker protein for Alzheimer's disease, resulting in concentrations very close to the detection limit of mass spectrometry for each of the molecular species (5). In addition, the trade-off between resolution and dynamic range prohibits simultaneous identification and quantitation of complex protein samples (6). With conventional mass spectrometry charge states can only be resolved if the number of masses present is much greater than the number of ions sampled (7). Multiple efforts have pushed mass spectrometry towards single protein (8) and single-molecule resolution (reviewed by Keifer *et al.* (7)). However, due to limited detection speed these techniques appear unsuitable for proteomics and other high-throughput applications and more suitable for the analysis of large

objects such as viruses, cells and organelles, which are beyond the mass limit of conventional mass spectrometry.

Whereas genomics benefits from high-throughput technologies as developed by Illumina (9,10), Roche (11), Applied Biosystems (12) and emerging single-molecule techniques from Helicos (13), Pacific Biosciences (14,15) and Oxford Nanopores (16), development of highly sensitive, deep protein sequencing solutions lags behind. Where DNA and RNA consist of four unique building blocks, proteins are built from 20 distinctive amino acids. Independent of the read out method of choice, this requires the detection of 20 distinguishable signals, a non-trivial challenge. Aforementioned DNA sequencing techniques can utilize polymerase enzymes to amplify sample. Protein sequencing platforms lacks this advantage since such copying machinery has not been discovered or engineered for proteins. Therefore, protein sequencing techniques will only be commercially successful if they can detect very low protein numbers.

Single-molecule techniques deal extremely well with samples containing target molecules with low copy numbers. In this chapter we will review recent efforts to establish single-molecule protein sequencing based on nanopores, tunneling current measurements and fluorescence.

1.2 Protein sequencing with nanopore technology

The concept of using nanopores for sequencing purposes was proposed over two decades ago (17). After applying a voltage over an artificial membrane containing a nanometer-sized pore, biopolymers can be driven through the nanopore by diffusion, electrophoretic and electro-osmotic flow (18). Compared to biological nanopores, nanopores fabricated from solid-state materials, such as silicon nitride, silicon dioxide and graphene, allow for controllable nanopore formation with increased stability and potential adjustment of surface properties (19). Nonetheless, approaches using biological nanopores, such as α-hemolysin, have currently been more successful for sequencing and have recently led

to the first commercial, nanopore-based, DNA sequencer (20).

Amino acids residues vary widely in charge distribution, unlike DNA which is essentially uniformly charged. Therefore, electrophoresis driven unidirectional translocation of polypeptides through nanopores is not self-evident. Tagging protein substrates with an oligonucleotide tail has shown to facilitate potential driven translocation of proteins through an α-hemolysin pore (21,22). Using this technique unphosphorylated, monophosphorylated and diphosphorylated thioredoxin could be distinguished based on the level of current blockage and noise (23).

Most natively folded proteins are too large to translocate through an α-hemolysin pore without unfolding. Unfolded proteins, on the other hand, translocate through in milliseconds, too fast for sequencing purposes. Motor proteins such as ClpX can be used to control unfolding and translocation speed through a nanopore (24). Again based on repetitive patterns in current blockage and noise levels, different protein domains could be identified (**Figure 1.1**) (25). Using machine learning algorithms point mutations, truncations and strand rearrangements could be detected based on their unique current patterns. Although the detection of individual amino acids seems ostensibly far away, this study proves the feasibility of using nanopores in combination with motor proteins for sequencing purposes.

Even though with materials such as graphene (26) nanopores of 1 Å thickness can be produced, the detection area remains a spherical volume dependent on the size of the pore, rendering reading single amino acids without detecting their neighboring residues impossible (27). Even with pores with diameters in the sub-nanometer range, the ionic-current blockage signal reported on "words" of four amino acids (28). The number of fluctuations observed matched the number of amino acids of the model protein. The current blockage, however, corresponded to the expected volume excluded by quadromers. Machine learning algorithms can be used to identify amino acids matching a

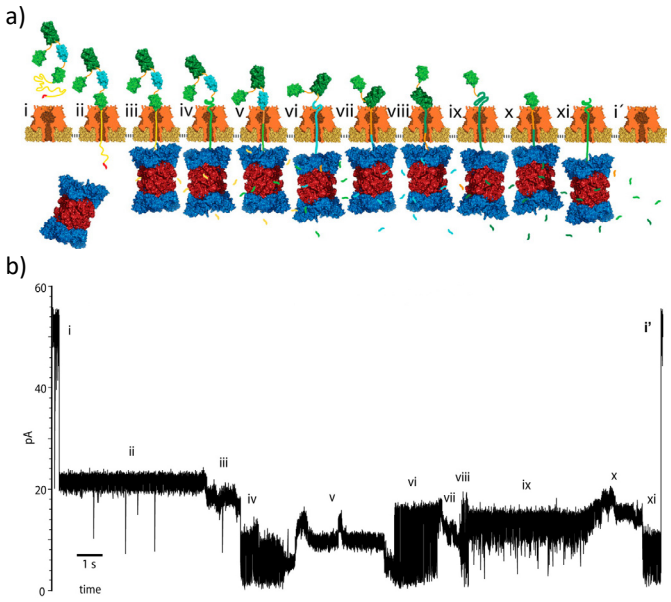


Figure 1.1: (**a**) Working model of ClpXP mediated protein translocation through a biological nanopore. (**b**) Representative current blockage profile reveiling domain specific signals during ClpXP mediated translocation. (Figures adapted from (25)).

certain nanospectrum (29). Computational analysis was simplified by grouping the amino acids in four size categories. The machine learning algorithm was applied to existing nanospectra from denatured and charge linearized proteins translocating through sub-nanometer pores. Proteins could be identified with typical p-values of 0.001-0.05, making it difficult to identify proteins from large databases with current signal-to-noise ratios.

Theoretically, the tasks of unfolding and sequencing of the polypeptide could be divided over two nanopores using a tandem electrolytic cell (30,31). The main advantage of this approach is the potential to detect individual amino acids by preventing the presence of multiple amino acids in the detection volume. The upstream pore could translocate the polypeptide into the trans1/cis2 void where an exopeptidase cleaves off one amino acid at the time. Driven by diffusion, the amino acids will pass through the downstream pore where they are identified based on current blockage and dwell time. Although the computational results

show some amino acids could be distinguished with high fidelity, to distinguish the sequence of all 20 naturally occurring amino acids with a confidence interval of 90%, ~70.000 identical proteins must be sequenced.

1.3 Protein sequencing with tunneling currents

Zhao *et al.* (32) described the implementation of recognition tunneling, a technique based on scanning tunneling microscopy (STM) developed in their lab (33), for single-molecule protein sequencing. Two metal electrodes, coated with 4(5)-(2-mercaptoethyl)-1*H*-imidazole-2-carboxamide (ICA), are placed with a ~2 nm gap. ICA is a chemical reagent that was developed to interact with DNA bases (34), additionally Zhao *et al.* (32) showed these recognition molecules also form weak, non-covalent, hydrogen bounds with amino acids, trapping them for ~0.2 s. Applying a small voltage across the electrode gap resulted in clustered peaks with features characteristic for the chemical composition of the molecule trapped in the gap. A machine-learning algorithm was trained to identify unique peak features in the complex signals derived from each amino acid (Gly, mGly, L-Asn, D-Asn, Leu, Ile and Arg). The trained machine-learning algorithm was then applied to signals from L-Asn and D-Asn mixed in different ratios. Although a linear trend was observed, the stoichiometric ratio was either under- or overestimated, depending on the method of quantification, due to preferential binding of L-Asn. This technique could be with combined nanopores to reduce the concentration of analyte needed, typically 1-100 μM, and an exopeptidase to feed the system amino acids sequentially.

By reducing the gap between the electrodes further, to 0.55 or 0.7 nm, single amino acids could be identified even in the absence of a layer of recognition molecules (**Figure 1.2**) (36). Nanogap electrodes were produced from nanofabricated mechanically controllable break junctions. Depending on the gap size, electron tunneling currents were measured for twelve out of twenty amino acids, remaining amino acids did not generate a detectable signal. The level of

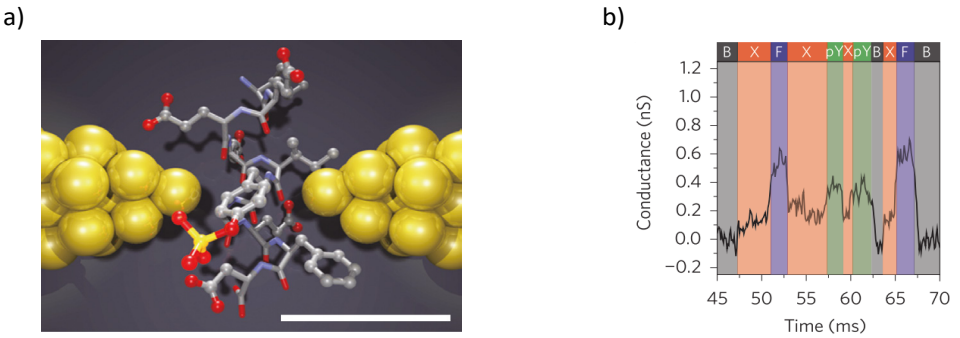


Figure 1.2: (**a**) Schematic representation of tunneling current measurements (Figure adapted from (35)). (**b**) Example trace with conductance levels assigned to individual amino acids (Figure adapted from (36)).

conductance and peak duration were sufficiently unique to be used as identifiers for the twelve detectable amino acids. In addition, phosphorylated tyrosine, an important post-translational modification in signal transduction and growth regulation (37), could be distinguished from its unphosphorylated form. A mixture of tyrosine and phosphotyrosine resulted in a conductance histogram with two peaks, areas under the peak corresponded to tyrosine:phosphotyrosine ratios. Adding model peptides with either tyrosine or phosphotyrosine to the system resulted in similar conductance profiles as were observed for single amino acids. Three single conductance peaks could be assigned to specific amino acids including (phospho)tyrosine, the remaining peak resembled a blend of three other amino acids that could not be resolved.

The major disadvantage of the approach described above is the need for different sized nanogap electrodes; a single gap is unable to distinguish the twelve amino acids mentioned here. In addition, analyte concentrations needed for tunneling experiments are currently in the 1-100 μM range (32,36). These limitation could be overcome by introducing transverse ionic transport, an intersection of two nanochannels is created where an ionic current flows through the transverse channel and the polypeptide is threaded through the longitudinal channel (38). Continues longitudinal threading of the polypeptide could be assisted by electrophoretic or electroosmotic force, as was proposed for DNA sequencing (39), by

using optical tweezers (40) or by use of a molecular motor such as ClpXP studied in this thesis and by others (24,25). The ionic current distributions obtained from transvers ionic current measurements will vary depending on chemical structure and size of the residue. Molecular dynamics simulations suggest current distributions will be unique for each residue. Currently, a sequencing device based on transverse ionic current has not been build and therefore this approach has not yet been experimentally demonstrated.

1.4 Protein sequencing with fluorescence techniques

Edman degradation, developed in the 1940s by Pehr Edman, is a well-known method to determine primary amino acid sequences from short, purified peptides without the need of a database reference (41). In short, the N-terminal amino acid is labeled and subsequently cleaved from the polypeptide chain. The released amino acid can in turn be identified by chromatography or electrophoresis. Two independent research groups proposed to combine Edman degradation chemistry with single-molecule fluorescence (42,43). Target proteins are fragmented into short peptides in predictable locations by a protease, subsequently amino acids are labeled with fluorophores and labeled peptides are immobilized on a surface (**Figure 1.3**) (44). Using Total Internal Reflection Fluorescence (TIRF) microscopy, the fluorescent signal of the fluorophores can be monitored. A drop in fluorescence after a cycle of Edman degradation reports on the release of a specific amino acid. Although the principle describe here appears straightforward, practical implementation is non-trivial. Edman degradation creates a harsh chemical environment, not compatible with many commercially-available fluorophores. In addition, the method might suffer from inefficient fluorophore conjugation, photobleaching or inefficiency of Edman degradation. Due to lack of specific chemistry and insufficient numbers of spectrally distinguishable fluorescent probes, the number of distinct amino acids labeled is limited. Computational analysis has shown that, while taking method specific errors into account, retrieving a proteins fingerprint, in other words identifying a subset of the 20

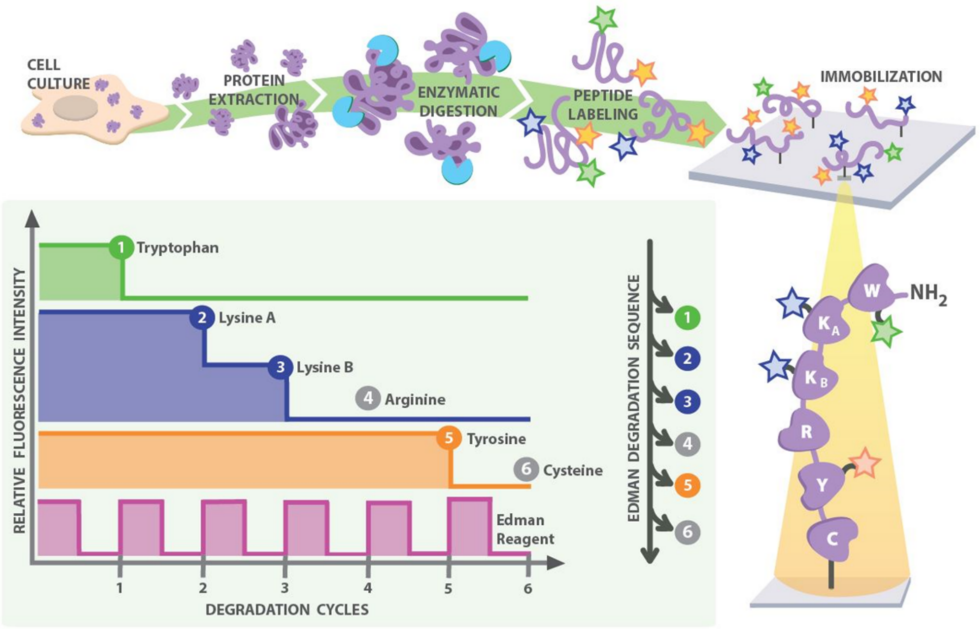


Figure 1.3: Schematic representation of fluorescence-based protein sequencing with Edman degradation. (Figure adapted from (46)).

naturally occurring amino acids, is sufficient to identify proteins from an existing database (45,46).

In this thesis we present a single-molecule protein sequencing platform we have developed by combining single-molecule FRET (Förster Resonance Energy Transfer) with the AAA+ protease ClpXP from Escherichia coli. ClpXP, a molecular motor protein from the AAA+ family (47,48), unfolds and degrades proteins with specific recognition tags that are used in vivo for protein degradation and remodeling. ClpX is a homohexameric ring that can exercise mechanical force of roughly 20-30pN on a folded protein using ATP hydrolysis (49,50). ClpX partners with ClpP, a homotetradecameric protease that strongly self-assembles into a barrel shape, shielding its fourteen cleavage sites from its surroundings (48). Together, ClpXP can bind, unfold, translocated and degrade proteins in a highly processive manner, making it a perfect candidate to scan full length sequencing substrates (51). In short, target proteins are labeled with acceptor fluorophores

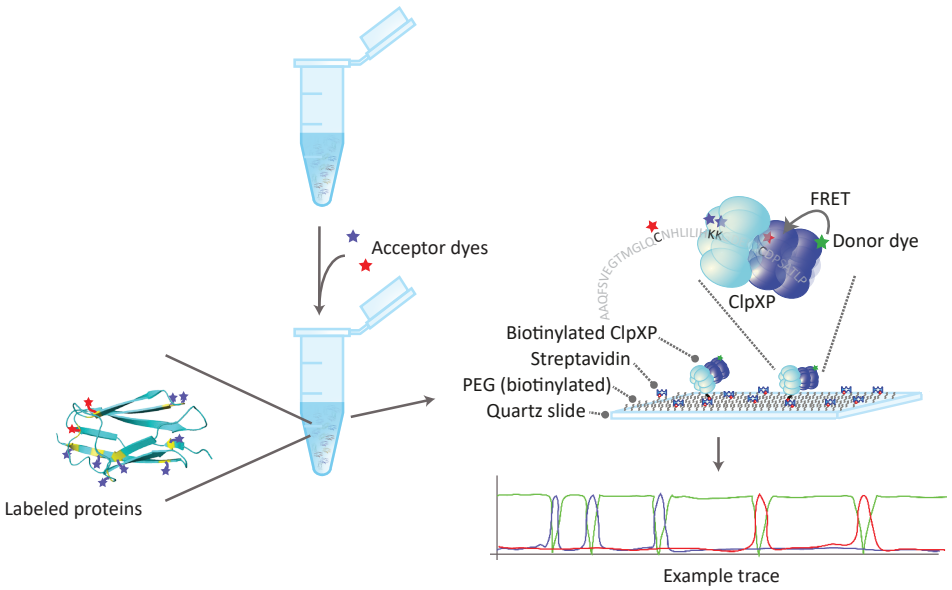


Figure 1.4: Schematic representation of the single-molecule protein fingerprinting technique further described throughout this thesis.

on cysteine and lysine residues and introduced to ClpXP, which is immobilized through biotin-streptavidin interaction and equipped with a donor fluorophore. After binding of the target protein to ClpXP, ATP dependent translocation is initiated, carrying the acceptor fluorophores into the Förster radius (2-8 nm) (52), inducing a drop in intensity of the donor signal and a concurrent increase in acceptor signal (**Figure 1.4**).

1.5 Outline of this thesis

In **Chapter 2** we discuss the feasibility of predicting full protein sequences from fingerprinting data. An algorithm was designed to compare simulated fingerprinting data to proteins from a database. Errors arising from sample handling and data collection were modeled into the simulated data to monitor their effects on detection precision. This study showed that incorporating additional information, such as distance between the cysteines and lysines detected, can greatly improve the performance of our method.

In **Chapter 3**, the highlight of this thesis, we show proof-of-concept of the single-molecule protein sequencing platform briefly described above. We successfully immobilized ClpXP on our single-molecule surface and observed FRET between donor-labeled ClpP and acceptor labeled substrates. We could detect peptides carrying two distinct acceptor dyes and determine the order of the dyes in C-terminal to N-terminal direction.

In **Chapter 4** we share the insight we gained in substrate recognition by ClpXP. We studied the effect of nucleotide composition on the binding and translocation efficiency of ClpX. Interestingly, only up to 8% of all the substrate binding events result in successful translocation. We observed AMP-PNP allows for ClpXP complex formation, but inhibits substrate binding. ATPγS on the other hand, allows for substrate binding, but partly inhibits translocation and reduces translocation speed. In addition, we observed major changes in ClpXP processivity for mutated versions of the ssrA degradation tag.

In **Chapter 5** we give a detailed description of the development of ClpP mutants that enable introduction of a donor fluorophore as used in Chapter 3 and 4. We introduced cysteines in several positions for site specific labeling and evaluated their performance both in bulk and in our single-molecule assay.

In **Chapter 6** we discuss the potential market impact of our research. We systematically explored the market need, performed a technology assessment, identified lead users and performed a preliminary market analysis. Our analysis provides a framework for a strategic technology roadmap.

1.6 References

1. Yates III, J. R. A century of mass spectrometry: from atoms to proteomes. *Nat. Methods* **8,** 633–637 (2011).
2. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–81 (2014).
3. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–7 (2014).

4. Muñoz, J. & Heck, A. J. R. From the Human Genome to the Human Proteome. *Angew. Chem. Int. Ed. Engl.* 2–5 (2014).

5. Portelius, E. *et al.* Characterization of tau in cerebrospinal fluid using mass spectrometry. *J. Proteome Res.* **7,** 2114–20 (2008).

6. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **28,** 710–21 (2010).

7. Keifer, D. Z. & Jarrold, M. F. Single-molecule mass spectrometry. *Mass Spectrom. Rev.* **47,** (2016).

8. Hanay, M. S. *et al.* Single-protein nanomechanical mass spectrometry in real time. *Nat. Nanotechnol.* 1–7 (2012).

9. Balasubramanisn, S. & Bentley, D. Polynucleotide arrays and their use in sequencing. WO patent 2001057248A2. (2001).

10. Bentley, D. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

11. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–80 (2005).

12. Valouev, A. *et al.* A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18,** 1051–1063 (2008).

13. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–9 (2008).

14. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–8 (2009).

15. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–6 (2003).

16. Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat. Biotechnol.* **30,** 295–6 (2012).

17. Deamer, D. W. & Akeson, M. Nanopores and nucleic acids: Prospects for ultrarapid sequencing. *Trends Biotechnol.* **18,** 147–151 (2000).

18. Ho, C. *et al.* Electrolytic transport through a synthetic nanometer-diameter pore. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 10445–50 (2005).

19. Dekker, C. Solid-state nanopores. *Nat. Nanotechnol.* **2,** 209–15 (2007).

20. Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat. Biotechnol.* **30,** 295–6 (2012).

21. Rodriguez-Larrea, D. & Bayley, H. Multistep protein unfolding during nanopore translocation. *Nat. Nanotechnol.* **8,** 288–95 (2013).

22. Rodriguez-Larrea, D. & Bayley, H. Protein co-translocational unfolding

depends on the direction of pulling. *Nat. Commun.* **5,** 4841 (2014).

23. 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).
24. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* 1–5 (2013).
25. Nivala, J., Mulroney, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among protein variants using an unfoldase-coupled nanopore. *ACS Nano* **8,** 12365–75 (2014).
26. Heerema, S. J. & Dekker, C. Graphene nanodevices for DNA sequencing. *Nat. Nanotechnol.* **11,** 127–136 (2016).
27. Lindsay, S. The promises and challenges of solid-state sequencing. *Nat. Nanotechnol.* **11,** 109–111 (2016).
28. Kennedy, E., Dong, Z., Tennant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm3 resolution using a subnanometre-diameter pore. *Nat. Nanotechnol.* (2016).
29. Kolmogorov, M., Kennedy, E., Dong, Z., Timp, G. & Pevzner, P. Single-Molecule Protein Identification by Sub-Nanopore Sensors. 1–10 (2016).
30. Sampath, G. Amino acid discrimination in a nanopore and the feasibility of sequencing peptides with a tandem cell and exopeptidase. *RSC Adv.* **5,** 30694–30700 (2015).
31. Sampath, G. A tandem cell for nanopore-based DNA sequencing with exonuclease. *RSC Adv.* **5,** 167–171 (2015).
32. Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* **9,** 466–73 (2014).
33. Lindsay, S. *et al.* Recognition tunneling. *Nanotechnology* **21,** 262001 (2010).
34. Liang, F., Li, S., Lindsay, S. & Zhang, P. Synthesis, Physicochemical Properties, and Hydrogen Bonding of 4(5)-Substituted 1-H-Imidazole-2-carboxamide, a Potential Universal Reader for DNA Sequencing by Recognition Tunneling. *Chem. - A Eur. J.* **18,** 5998–6007 (2012).
35. Di Ventra, M. & Taniguchi, M. Decoding DNA, RNA and peptides with quantum tunnelling. Nat. Nanotechnol. 11, 117–126 (2016).
36. Ohshiro, T. et al. Detection of post-translational modifications in single peptides using electron tunnelling currents. Nat. Nanotechnol. 9, 835–840 (2014).
37. Hunter, T. & Cooper, J. A. Protein-tyrosine kinases. *Annu. Rev. Biochem.* **54,** 897–930 (1985).
38. Boynton, P. & Di Ventra, M. Sequencing proteins with transverse ionic transport in nanochannels. 1–10 (2015).

39. Wilson, J. & Di Ventra, M. Single-base DNA discrimination via transverse ionic transport. *Nanotechnology* **24,** 415101 (2013).

40. Keyser, U. F. *et al.* Direct force measurements on DNA in a solid-state nanopore. *Nat. Phys.* **2,** 473–477 (2006).

41. Edman, P. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **4,** 283–293 (1950).

42. Marcotte, E., Swaminathan, J., Wllington, A. & Anslyn, E. Identifying peptides at the single molecule level. US patent 20140349860. (2014).

43. Hesselberth, J. R. Peptide identification and sequencing by single-molecule detection of peptides undergoing degradation. US patent US20150087526. (2013).

44. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **11,** e1004080 (2015).

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

46. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. *A theoretical justification for single molecule peptide sequencing*. (2014).

47. Sauer, R. T. & Baker, T. a. AAA+ proteases: ATP-fueled machines of protein destruction. *Annu. Rev. Biochem.* **80,** 587–612 (2011).

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

49. Aubin-Tam, M.-E., Olivares, A. O., Sauer, R. T., Baker, T. a & Lang, M. J. Single-Molecule Protein Unfolding and Translocation by an ATP-Fueled Proteolytic Machine. *Cell* **145,** 257–67 (2011).

50. Maillard, R. A. *et al.* ClpX(P) Generates Mechanical Force to Unfold and Translocate Its Protein Substrates. *Cell* **145,** 459–69 (2011).

51. Joo, C., Dekker, C., Ginkel, H. G. T. M. van & Meyer, A. S. Single molecule protein sequencing, WO patent 2014014347. (2014).

52. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5,** 507–16 (2008).

Chapter 2

Single-Molecule Protein Sequencing Through Fingerprinting: Computational Assessment

This chapter has been published as:

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

Proteins are vital in all biological systems as they constitute the main structural and functional components of cells. Recent advances in mass spectrometry have brought the promise of complete proteomics by helping draft the human proteome. Yet, this commonly used protein sequencing technique has fundamental limitations in sensitivity. Here we propose a method for single-molecule protein sequencing. A major challenge lies in the fact that proteins are composed of 20 different amino acids, which demands 20 molecular reporters. We computationally demonstrate that it suffices to measure only two types of amino acids to identify proteins and suggest an experimental scheme using single-molecule fluorescence. When achieved, this highly sensitive approach will result in a paradigm shift in proteomics, with major impact in the biological and medical sciences.

2.1 Introduction

In 2014 two international teams produced the first draft of the human proteome, using mass spectrometry (MS) (1,2). By opening a new chapter in proteomics, these large scale studies will help us understand complex cellular processes. Yet, MS—the most widely used protein sequencing technology—requires a large amount of sample. This hampers quantification, precludes detecting many proteins of interest that are present only in low concentrations in the cell, and renders single-cell analysis impossible.

Single-molecule (SM) protein sequencing would bring about "protein deep sequencing" (3–5). However, unlike DNA sequencing that needs to read out only 4 nucleotides, protein sequencing demands differentiation of 20 amino acids, far beyond what current SM techniques can offer (3). SM protein sequencing has therefore not followed up SM DNA sequencing that uses fluorescence and nanopores (6–8). Here we propose a novel SM protein sequencing method that overcomes this challenge and assess its feasibility using computational analysis.

2.2 Results and discussion

Unique to protein sequencing is that a protein can be identified using incomplete information with reference to proteomic databases. Consider a 2-bit fingerprinting scheme in which only two types of amino acids are labeled (**Figure 2.1**).

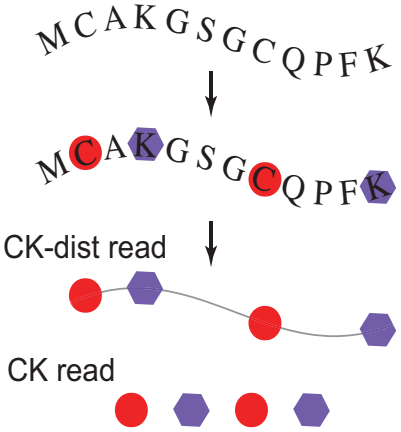


Figure 2.1: A single-molecule read-out. CK fingerprinting read: the order of C's and K's are detected. CK-dist fingerprinting read: the distances between C's and K's are additionally measured.

A consecutive read of 15 labeled amino acids is sufficient to identify up to 2^{15} = 32,768 unique protein sequences. This exceeds the number of (major isoform) protein species that most organisms express. As the median length of a protein ranges from 270 (bacteria) to 350 amino acids (eukaryotes), it is not difficult to choose two amino acid types that appear more than 15 times in each protein (**Supplemental figure S2.1**).

Figure 2.2 describes a SM protein fingerprinting scheme using fluorescence. We chose to label two highly nucleophilic amino acids, lysine (K) and cysteine (C) as they are frequent (**Supplemental figure S2.2**) and can be labeled both efficiently and orthogonally (NHS-ester coupling with lysine and maleimide coupling with cysteine) (9). A similar idea using lysine and arginine for monitoring protein synthesis inside a living cell was patented by *Anima Cell Metrology* (10)**.** Recently, Swaminathan et al discussed fingerprinting schemes that are based on multiple labels including two labels (3). Separately, a work published in 2013 shows how

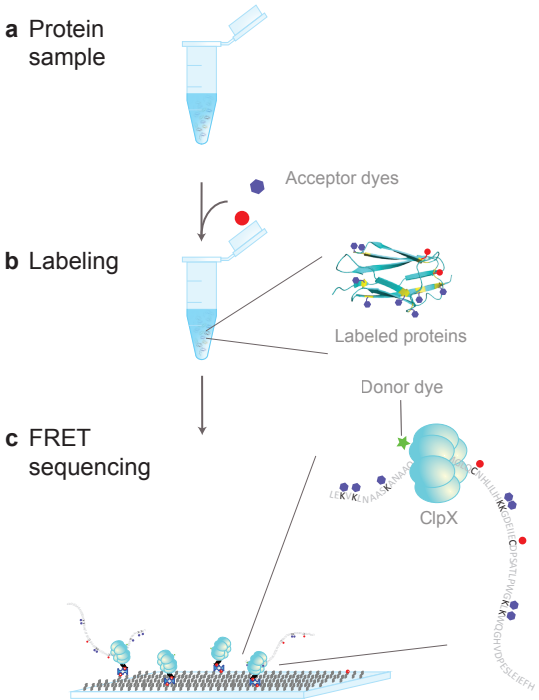


Figure 2.2: The schematic shows how single-molecule fingerprinting is carried out. (**a**) Proteins are obtained from diverse sources including single cells, tissues, and body fluids. (**b**) Extracted proteins are denatured, and cysteines and lysines are labeled with fluorescent dyes. (**c**) An engineered version of a protein translocase (e.g. bacterial ClpX) grabs individual substrate proteins, unfolds them, and translocates them through its nano-channel. Proteins are sequenced using FRET (Förster resonance energy transfer). The translocase is labeled with a donor dye. FRET occurs between the donor on the translocase and the two distinct acceptor dyes on a substrate when the substrate passes through the nanomachine. The FRET signals report the order of the labeled amino acids.

our fingerprinting approach might be implemented using nanopores (4).

To assess the predictive power of fingerprinting, we developed a dedicated search algorithm. In brief, we search CK (cysteine-lysine) fingerprints by combining a filtering strategy to decrease computation time with a dynamic programming-based alignment step, considering a specific set of potential experimental errors (see ***Methods***).

For our analysis, we used a canonical human proteome database based on Uniprot release 2014.04 (11). We simulated 2,000 different read-outs, searched for each of them in the database and measured the detection precision (P), i.e. the probability of retrieving the correct sequence. In an ideal situation with no experimental error, P is 90% (**Figure 2.3a**, blue). Next, we assessed the robustness of the method against inaccuracies that are expected from actual experiments, by iteratively introducing errors into each fingerprint at random (See *Error simulation* in ***Methods*** for details) up to a specified error level (α, a number of errors

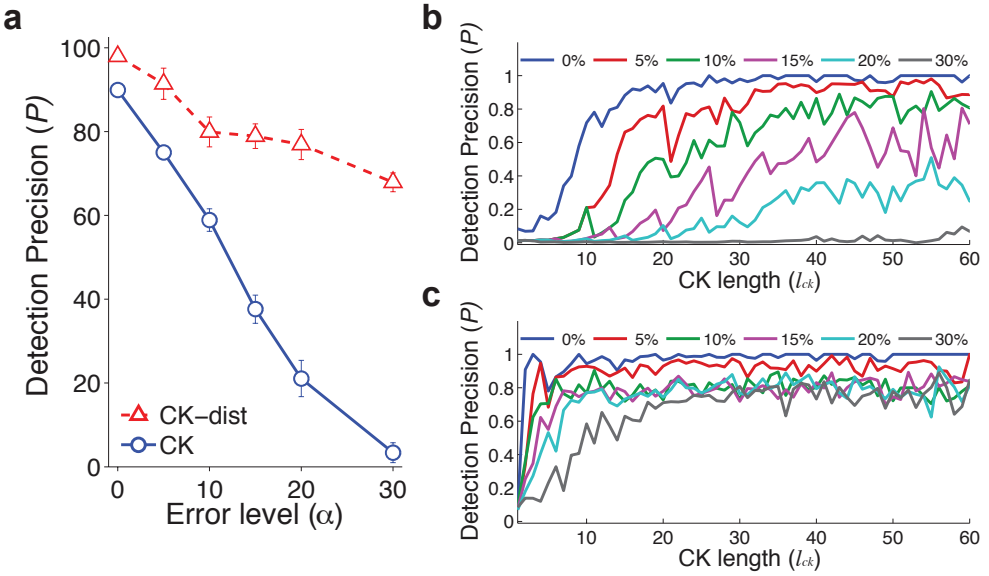


Figure 2.3: (**a**)Detection precision, *P*, at various error levels, α: blue for CK fingerprinting, red for CK-dist fingerprinting. Error bars are the standard deviation from three independent simulations. (**b**) *P* as a function of CK fingerprint length (l_{ck}, the length of CK sequences excluding other amino acids) at various α for CK fingerprinting. (**c**) *P* as a function of l_{ck} at various α for CK-dist fingerprinting.

(*k*) divided by a CK fingerprint length (l_{ck}, the length of CK sequences excluding other amino acids). **Figure 2.3** reports *P* of these computations. As expected, *P* drops when α increases. For example, at α = 10%, half of the sequences are correctly and uniquely retrieved (**Figure 2.3a,** blue).

To improve performance, we considered other information: the distance between C's and K's (**Figure 2.3a**, red). At any α, *P* was dramatically higher with the distance included: at =10%, *P* increased to 85%. In general, *P* increases when l_{ck} becomes longer (**Figure 2.3b-c**). At any l_{ck}, *P* for CK fingerprinting with distance information (named "CK-dist fingerprinting") is higher than or equal to *P* for CK fingerprinting. A similar observation was made when different additional information was considered (**Supplemental figure S2.5**). Taken together, these demonstrate the feasibility of the technique for the application in identifying primary protein sequences.

Another application area could be clinical diagnosis. As an example of detecting infections, we chose human respiratory syncytial virus (HRSV) and tuberculosis (TB). We determined that a set of HRSV and TB proteins contain a unique CK fingerprint and thus can be detected at α as high as 15-20% and be potentially used as markers for HRSV and TB (**Supplemental figure S2.6**).

The CK fingerprinting technique will also enable us to detect post-translational modifications of proteins when it is expanded to a three-color fluorescence measurement. For example, glycosylated amino acids can be labeled with a third acceptor dye using hydrazide-aldehyde coupling chemistry, which is orthogonal to the labeling methods for lysine and cysteine residues. Phosphorylated serine and threonine can be labeled with a third acceptor using another coupling scheme (12). This will advance the proof-of-principle of detecting a post-translationally modified peptide using a nanopore that was reported in 2014 (5).

2.3 Conclusions

We described SM protein fingerprinting, a technique that will provide proteomics with high sensitivity and a large dynamic range. Our computational assessment indicated that, even if we read only two amino acid types, we could correctly identify proteins with reference to proteomic databases. When this entirely new SM protein sequencing approach is achieved, it will become a proteomics tool that complements MS and opens up new avenues in global, high-throughput protein analysis.

2.4 Methods

Here we describe the approaches that we used to simulate errors and find protein fingerprints that match a given query fingerprint pattern.

2.4.1 Error simulation

We simulated 2,000 read-outs, each for a different protein. The proteins are randomly picked from the database and thus contain random amino acids and fingerprint lengths. Next, to assess the robustness of the method against inaccuracies that are expected from actual experiments, errors are iteratively introduced for each read-out up to the error level we want to investigate.

We expect that actual data will be convoluted with poor dye-labeling, photoblinking and photobleaching of dyes, a local structure of a substrate protein, a non-uniform speed of substrate translocation, proximity between dyes etc. The poor labeling, photoblinking, and photobleaching of acceptor dyes will appear as deletion errors (**Figure 2.4**). The non-uniform speed of translocation will introduce insertion and deletion errors to CK-dist fingerprinting. The proximity of acceptor dyes will bring deletion and transposition/substitution errors. If a donor dye is photobleached during a measurement, it will appear as a truncation error. We do not consider this error for fingerprinting analysis since donor photobleaching can be determined from single-molecule time traces and thus can be easily excluded from further analysis. Other complications, such as aggregation of denatured proteins, may also be expected but are not considered in our analysis. See a pseudo-code for simulating these errors (**Supplementary data 2.5.3**).

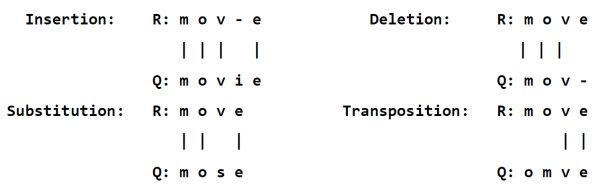


Figure 2.4: Expected experimental errors. 'R' is reference sequence. 'Q' is query sequence.

In **Figure 2.3**, we investigated one combination of errors (70% deletions, 20% insertions, 10% transpositions) for CK fingerprinting, in which we assigned the largest percentage to deletions since this error is the most likely to occur (poor labeling of acceptors due to incomplete denaturation of proteins, photoblinking/ photobleaching of acceptors, and presence of consecutive identical acceptor fluorophores).

For CK-dist fingerprinting analysis, we considered the same combination of errors as for CK fingerprinting but with errors at CK residues and errors of the distance between CK residues equally likely to occur. In **Supplemental figure S2.3**, we expanded the error space that we explored and obtained trends nearly identical to that found in **Supplemental figure S2.3a**.

2.4.2 Overview of CK fingerprinting

The 2,000 simulated readouts are searched for in the database, and the numbers of true positives and the number of matches are recorded. To examine the performance variability of our algorithm in retrieving proteins using fingerprints, three independent repetitions are executed. In each repetition, detection precision (*P*) (**Figure 2.3**) and detection recall (*R*) are calculated based on the outputs (see **Figure 2.4**). *P* is defined as the number of true positives divided by the number of read-outs returned by the algorithm. *R* is the number of true positives divided by the number of conditional matches.

The inputs to our method are a reference database *R* containing fingerprint representations of protein sequences, a query fingerprint *Q* and an error level α. The alphabet is $\Sigma = \{C, K\}$, since we only compare fingerprints of these two amino acids. Let L_Q be the query length and $R_x \in R$ denote the *x*th reference sequence in the database *R* with length L_x^R. The distance $S(R_x, Q)$ between a reference fingerprint R_x and a query *Q* is the minimal number of steps required to transform *Q* into R_x. Formally, given *Q*, *R* and α, the problem is to find all $R_x \in R$ for which $S(R_x, Q)$ is smaller than $k = \alpha \times L_Q$.

Given the inputs, the algorithm takes two steps to retrieve matches: 1) a filtration strategy is applied to identify candidate sequences in *R*; and 2) a verification method is employed to examine all candidates for possible matches.

2.4.3 Filtration: eliminating uninteresting sequences

Dynamic programming is computationally costly, prohibiting direct application on large databases in a high-throughput setting (13). In order to reduce the running time without affecting sensitivity, we use filtration to remove those references that definitely cannot match the query fingerprint *Q* with distance smaller than or equal to *k*. Filtration exploits the fact that it is easier to tell a reference fingerprint that does not match a query fingerprint than to tell one that does match. Typically, it uses a simple and highly efficient filter criterion to analyze the reference sequences, leaving only a small number of $R_x's$ for further (more expensive) analysis. We devised a new filtration method combining two existing algorithms, partial exact matching and -gram counting.

In partial exact matching, the query fingerprint Q is divided into $(k+1)$ pieces $q^0, q^1, ..., q^k$, where k equals $\alpha \times L_Q$. For a match to be possible, there must be at least one piece that appears exactly in a reference sequence R_x (14). If this is not the case, R_x is discarded.

A faster filtration method is -gram counting, which compares the -grams of two fingerprints. An n-gram (15) on the alphabet set $\Sigma = \{C, K\}$ is any string in Σ^n, where Σ^n is the set of all possible strings of length n over Σ. For example, the -grams for $\Sigma = \{C, K\}$ are CC, CK, KC and KK. The n-gram distance is defined as the sum of the absolute differences between the numbers of occurrences of each n-gram. If the n-gram distance exceeds $2nk$, R_x is discarded (15).

We combined the partial exact matching and n-gram counting approaches to decide whether there exists at least one piece in Q that appears with a limited amount of errors as a piece of R_x (16). The distance function between two pieces of Q and R_x, q^j and r_x^j, based on their n-grams was defined as:

$$S_{npm}\left(r_x^j, q^j\right) = \sum_{v \in \Sigma^n} \max\left(G\left(q^j\right)[v] - G\left(r_x^j\right)[v], 0\right)$$

where $[v]$ is an n-gram and $G\left(q^j\right)[v]$ and $G\left(r_x^j\right)[v]$ denote the total number of times $[v]$ occurs in q^j and r_x^j, respectively.

For each piece q^j in the query, the corresponding piece r_x^j contains the same letters in the reference sequence with an additional k letters on both sides, as shown in **Figure 2.5**. It is sufficient to compare the r_x^j in the reference with the q^j in the query to determine whether the piece q^j appears in the reference R_x, since k errors cannot alter more than k positions. Since a query piece is searched in a limited range in the reference, it can discard more entries in the reference database than the partial exact matching method, in which the q^j is compared with the entire reference sequence.

The distance between a piece q^j in query Q and the corresponding piece r_x^j in R_x is computed to determine whether R_x is a candidate match. For each q^j and its corre-

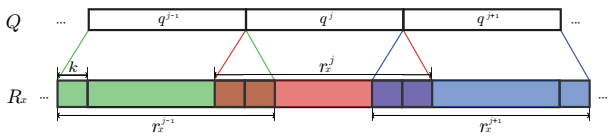


Figure 2.5: Consecutive pieces q^{j-1}, q^j and q^{j+1} of query Q and their corresponding pieces r_x^{j-1}, r_x^j and r_x^{j+1} in reference R_x. Each query piece is compared to a limited range in the reference.

sponding r_x^j, we check whether any -gram occurs more often in q^j than in r_x^j. If not, the $S_{npm}\left(r_x^j, q^j\right)$ is zero, i.e. the *n*-grams in q^j appear exactly in r_x^j. Only if for at least one q^j, $S_{npm}\left(r_x^j, q^j\right)$ is zero, R_x is kept as a candidate.

2.4.4 Verification: finding matches

The remaining candidate matches are examined by a global alignment dynamic programming approach considering a number of possible error types. In our analysis, four types of error may occur: deletion, insertion, mismatching an amino acid with another one (substitution), and swapping (transposition).

The dynamic programming algorithm is designed to provide the optimal gapped alignment between two sequences, i.e. an alignment with long regions of identical amino acid pairs and very few mismatches and gaps (17). As the sequences become more dissimilar, more mismatched amino acid pairs and gaps should appear. To find the optimal alignment, a dynamic programming matrix *M* first needs to be calculated. Each element $M_{i,j}$ represents the maximum score of aligning the substrings $Q[1...i]$ and $R_x[1...j]$. Let *c* denote the scores of the four operations. The base cases, $M_{0,j}$ and $M_{i,0}$, are defined as $\left(c_{del} \times j\right)$ and $\left(c_{ins} \times i\right)$ for all $1 \le j \le L_x^R$ (length of R_x) and $1 \le i \le L_Q$ (length of *Q*) respectively. Then, considering the four possibilities, *M* is updated using the following recursive relation,

$$M_{i,j} = \max \begin{cases} M_{i-1,j-1} + c_{sub} \\ M_{i-1,j} + c_{ins} \\ M_{i,j-1} + c_{del} \\ M_{i-2,j-2} + c_{swap} \end{cases}$$

The score for each operation is set based on the estimation of how likely each error is to occur in our measurements. Currently, deletions caused by low labeling efficiency are the dominating errors, followed by insertions, transpositions and substitutions (i.e. matching C to K or vice versa) (see *Error simulation*). Hence we choose a relatively low score (negative) for deletions and higher scores for transpositions and substitutions. For the matching positions, the score is positive (see **Table S2.2**).

By memorizing the solutions to the subproblems for $1 \le j \le L_x^R$ and $1 \le i \le L_Q$ stored in the dynamic matrix, we can recursively compute the maximum score of aligning R_x and *Q*. Therefore we find the score of the optimal alignment of the two sequences starting from the maximum value in the last row or last column. We maintain a matrix of traceback pointers in the recursion, so that we remember which case was used to calculate every cell $M_{i,j}$, allowing to reconstruct the optimal alignment.

From this alignment the numbers of errors for different types as well as the total number of errors can be calculated. The distance between the query and the refer-

ence $S(R_x, Q)$ is defined as the total number of errors. If this distance is smaller than *k*, the reference sequence R_x is considered as a *match*. Otherwise, it is not a match of the query sequence within the error bound *k*. A match is considered a *true positive match* when the match is the exact query protein. If a match has the same fingerprint but a different amino acid sequence, it is not considered to be a true positive match. In our analysis, this is determined by checking the protein accession codes.

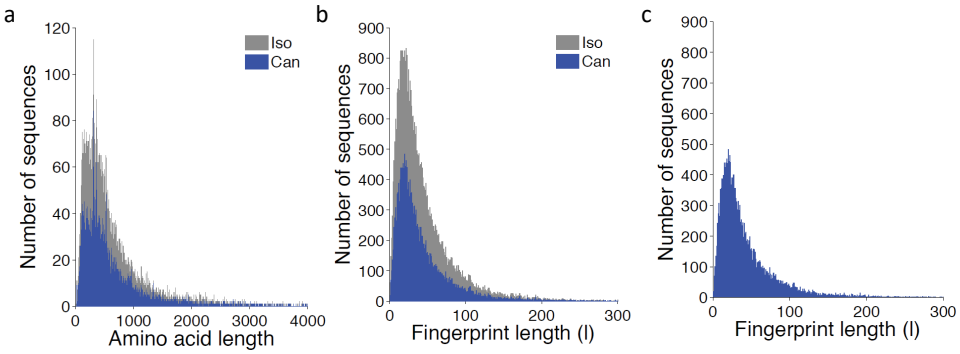
Additional information, such as the distance between two read-outs, can be deduced from the measurements. This distance is the space between two labeled amino acids, which is the number of non-labeled amino acids in between, which show a different pattern in the measurement. For this to be estimated reliably, proteins will have to be sequenced at a relatively constant speed, an assumption which is not *a priori* valid. From the sequencing signals, we cannot easily determine the start or the end of proteins in the time trace if they do not correspond to a labeled amino acid. Thus, the starting and ending non-labeled amino acids are not included when we construct the fingerprint with distance information.

This distance information is added to the original CK fingerprints using an additional symbol (say, 'o'), occurring multiple times (representing the length of distance). Next, two distances between query and reference are calculated to examine whether a reference sequence is a match. One is the $S(R_x, Q)$ between fingerprints with distance information, the other the $S'(R_x, Q)$ between CK fingerprints only. Reference sequence R_x is considered a match if and only if $S'(R_x, Q)$ is smaller than $k' = (\alpha \times L'_Q)$ and $S(R_x, Q)$ is smaller than $k = \alpha \times L_Q$, where the L_Q is the length of the query CK fingerprint, L'_Q is the length of the query fingerprint with distance information and *k'* and *k* represent the numbers of errors allowed. Experimental error on the distance information is also taken into consideration.

2.5 Supplementary data

2.5.1 Database and CK fingerprint length

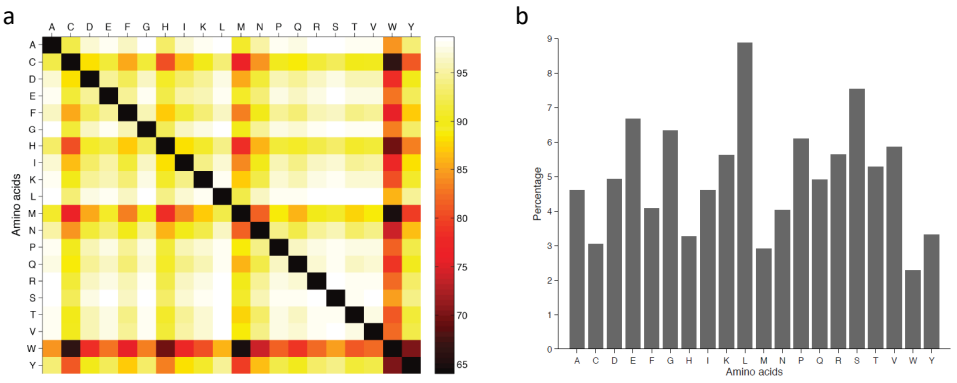
Two human complete proteomes (canonical; and canonical with isoforms) from Uniprot release 2014.04 are used to test our algorithm. There are 20,250 and 39,736 different proteins in the canonical (Can) and isoform (Iso) databases, respectively. Fourteen proteins in the canonical database that have no CK signature are removed from further analysis. In the isoform database, 49 proteins are removed. The length distribution of the amino acid sequences and fingerprints are shown in **Supplemental figure S2.1**. The average fingerprint length is 45 for the canonical database and 46 for the isoform database. The average number of C's is 13 and of K's is 32. Unless explicitly specified otherwise, the results presented were obtained on the 2,000 random proteins selected from canonical database (**Supplemental figure S2.1c**).



Supplemental figure S2.1: The length distribution of (**a**) amino acid sequences and (**b**) CK fingerprints from canonical (Can) and isoform (Iso) databases. (**c**) CK fingerprints of the 2,000 random selected.

2.5.2 Uniqueness of 2-bit fingerprints

To find out how our fingerprinting will perform with other 2-amino acid combinations, we analyzed the uniqueness of all possible choices of 2-amino acid combinations (S**upplemental figure S2.2**). A combination of the most frequent amino acids (L and S) shows the highest percentage of uniqueness (98.7%). A combination of W and M has the lowest (64.6%). The combination of C and K gives 89.8% uniqueness, which is around the average (87.3%). Although a choice for L and S is optimal from a computational point of view, the pair of C and K is chosen since it allows for protein labeling with minimal cross-labeling.



Supplemental figure S2.2: (**a**) Uniqueness of two-amino acid combinations and (**b**) amino acid composition of human proteins.

2.5.3 Pseudo-code for simulating errors

To simulate a dataset of reads containing errors, we proceed as follows. First, a sequence is randomly selected from the database. Specific errors are then iteratively introduced with certain probabilities, until the total number of errors applied exceeds a threshold (which is also a random number smaller than or equal to the maximum number of errors). This gives us simulated read-outs that contain no more than the specified maximum number of errors. We did not allow the errors to occur at the same position. The pseudo-code below shows how errors were introduced into CK fingerprints for simulation.

Input: sequence S, error level α

1) if $\alpha = 0$ do
2) return *S*
3) else
4) *max_no_err* := α * length(*S*)
5) *no_err* := a random integer between 1 and *max_no_err*
6) *pos*[1..*no_err*] := non-overlapping random integers between 1 and length(*S*)
7) sort *pos*[] in descending order
8) for each element *pos*[*i*] do
9) *err_ty* := a random number between 0 and 1
10) if *err_ty* ≤ 0.7 do
11) erase *S*[*pos*[*i*]] % Deletion
12) elseif *err_ty* ≤ 0.9 do

13) insert $S[pos[i]\text{-}1]$ at $S[pos[i]]$ % Insert the left adjacent AA

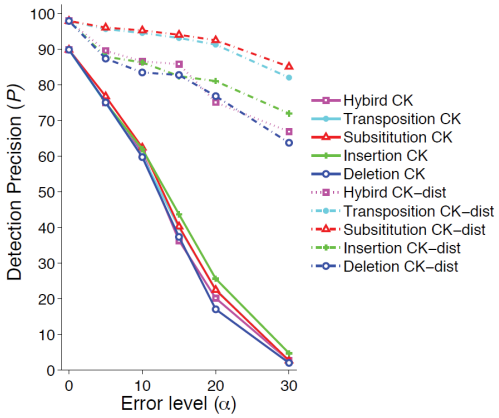
14) else

15) swap $S[pos[i]\text{-}1]$ with $S[pos[i]]$ % Transposition

Output: S

2.5.4 Detection precision (*P*)

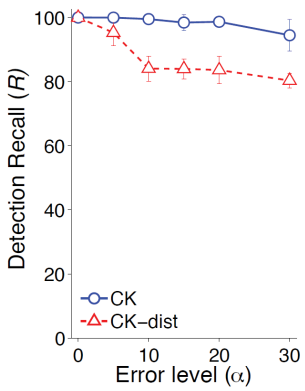
In **Supplemental figure 2.3**, we investigated one combination of errors (70% deletions, 20% insertions, 10% transpositions). We explored a larger error space by considering individual errors separately (**Supplemental figure S2.3**), All of these cases exhibit a trend nearly identical to that found in **Supplemental figure 2.3a**. It suggests that the detection precision that we measured for a particular case (70% deletions, 20% insertions, 10% transpositions) is generally valid for other combinations of errors.



Supplemental figure S2.3: We tested four extreme cases of experimental error. (Light blue) 100% errors are due to transposition. (Red) 100% errors are due to substitution. (Green) 100% errors are due to insertion. (Blue) 100% errors are due to deletion. Solid lines are for CK fingerprinting. Dotted lines are for CK-dist fingerprinting.

Detection recall (R)

R is the number of true positives divided by the number of conditional matches. *R* is an indicator of whether the true positive is retrieved for a query. In our experiment the conditional matches are always 1, and the true positive is one when the searched protein is retrieved; otherwise, the true positive is zero. Thus, *R* equals the number of true positives. When only the searched protein itself is retrieved, the search is optimal and thus both *P* and *R* are 1.



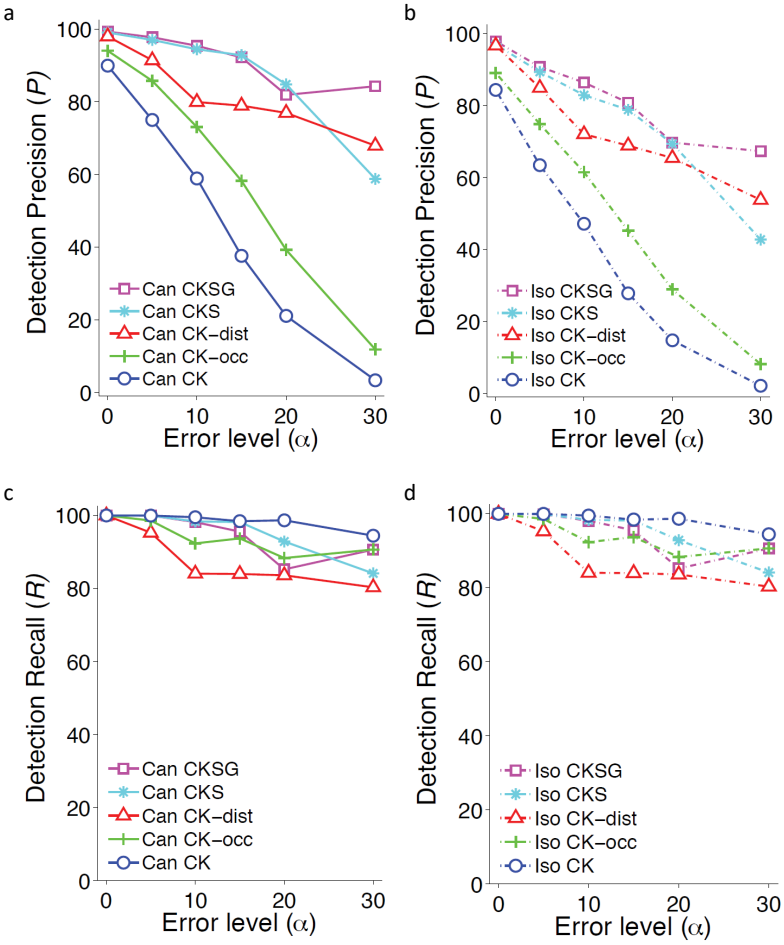
Supplemental figure S2.4: *R* at various error levels: blue for CK fingerprints, red for CK fingerprints with distance information (CK-dist fingerprinting).

Supplemental figure S2.4 presents the recall at various error levels. As error levels increase, we are less able to find the true positive match back, and so recall decreases. There are two reasons for this. First, one of the features of dynamic programing algorithm is that it favors deletions and insertions over substitutions and transpositions, where the latter two are considered as two deletions and/or insertions. Thus, $S(R!,Q)$ becomes bigger, which leads to misidentification. Second, increasingly the length of the true positive match falls outside of the search range $(1-\alpha)\times L! \le l \le (1-\alpha)\times L!$.

We also observe that errors have a larger influence on fingerprints with distance information than on CK fingerprints only. This is because we consider CK information to be more important than distance information, and in dynamic programing thus favor substitution of a distance symbol 'o' with a 'C' or a 'K' instead of a deletion of 'C' or 'K'. This trade-off occurs no matter what scores we choose to use in the verification phase (**Supplemental table S2.1**).

2.5.5 Additional information improves precision

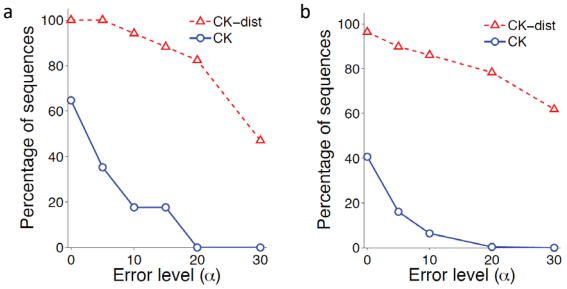
Here we examine the performance for different kinds of readouts (**Supplemental figure S2.5**). First, we measure the performance of fingerprints that consider occurrence of distance but not a length of distance (named CK-occ). We also consider three and four-labeled fingerprints (named CKS and CKSG, where S is for serine and G is for glycine, both randomly chosen.). When additional information is included, the precision increases at any error level. For CKS and CKSG fingerprints, recall drops slightly.



Supplemental figure S2.5: (**a-b**) *P*'s at various error levels and (**c-d**) *R*'s at various error levels. (**a**) and (**c**) are for the canonical database (Can), (**b**) and (**d**) for the isoform database (Iso). Blue for CK fingerprints (CK), yellow for CK fingerprints with occurrence of distance (CK-occ), red for CK fingerprint with distance information (CK-dist), light blue for CKS fingerprints (CKS), and pink for CKSG fingerprints (CKSG).

2.5.6 Clinical diagnosis

As an example of detecting infections, we chose human respiratory syncytial virus (HRSV) and tuberculosis (TB). The UniProt database contains 21 HRSV proteins; four of them have fingerprints shorter than 8. TB has more proteins (6327), 47.0% of which have a fingerprint length of 8 or shorter. These short fingerprints are excluded in further analysis. We searched each HRSV/TB protein in the human database using our algorithm. We computed the percentage of HRSV/TB protein sequences whose CK fingerprints are absent in the human proteome (**Supplemental figure S2.6**). When CK fingerprints of HRSV/TB proteins are used without errors, 65% of HRSV proteins and 41% of TB proteins are not found in human canonical database. When



Supplemental figure S2.6: Percentage of (a) HRSV and (b) TB proteins whose CK fingerprints are absent in human at various error levels. Blue line for CK fingerprints, red line for CK fingerprints with distance information. The higher the percentage, the more HRSV/TB proteins show unique CK fingerprints against human proteins.

HRSV:

Accession Number	Protein name
P03420	Fusion_glycoprotein_F0
O36634	Fusion_glycoprotein_F0
O36635	RNA-directed_RNA_polymerase_L

HRSV:

Accession Number	Protein name
Q02251	Mycocerosic_acid_synthase
P9WNF6	Putative_FAD-containing_monooxygenase_MymA
A1KQG0	Phthioceranic/hydroxyphthioceranic_acid_synthase
P9WQE6	Phthiocerol_synthesis_polyketide_synthase_type_I_Pps A
P9WQE2	Phthiocerol_synthesis_polyketide_synthase_type_I_Pps D
P9WQE0	Phthiocerol_synthesis_polyketide_synthase_type_I_Pps E
P9WN14	Uncharacterized_glycosyl_hydrolase_MT2062

Supplemental table S2.1: Lists of HRSV and TB proteins that are absent in human proteome at α = 15% (HRSV) and α = 20% (TB).

errors are introduced, this percentage drops, but a set of HRSV/TB CK fingerprints are still absent in the human database at error levels as high as 15% - 20% (**Supplemental table S2.1**). If we include distance information, almost all HRSV and TB proteins are correctly found to be non-human.

2.5.7 Score for each operation

In our analysis, four types of error may occur: deletion, insertion, mismatching an amino acid with another one (substitution), and swapping (transposition) (**Supplemental table S2.2**). The score for each operation c is set based on the estimation of how likely each error is to occur in our measurements. Currently, deletions caused by low labeling efficiency are the dominating errors, followed by insertions, transpositions and substitutions (i.e. matching C to K or vice versa). Hence we choose a relatively low score (negative) for deletions and higher scores for transpositions and substitutions. For the matching positions, the score is positive.

The scores used in the verification phase are given in the table below. Here, 'o' represents a distance; a/b (in some cells) gives both the substitution penalty a and the transposition penalty b. The scores for deletion and insertion are $c_{del} = -2$ and $c_{ins} = -5$, respectively.

	'C'	'K'	'o'
'C'	50		
'K'	-50/-45	50	
'o'	-1/-20	-1/-20	2

Supplemental table S2.2: Penalties determined for each error.

2.6 References

1. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–7 (2014).
2. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–81 (2014).
3. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **11,** (2015).
4. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* 1–5 (2013).
5. 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).

6. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–9 (2008).

7. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–8 (2009).

8. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4,** 265–70 (2009).

9. Joo, C., Dekker, C., Ginkel, H. G. T. M. van & Meyer, A. S. Single molecule protein sequencing, WO patent 2014014347. (2014)

10. Preminger, M. & Smilansky, Z. Methods for evaluating ribonucleotide sequences. (2009).

11. UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42,** D191-8 (2014).

12. Kim, J.-S., Kim, J., Oh, J. M. & Kim, H.-J. Tandem mass spectrometric method for definitive localization of phosphorylation sites using bromine signature. *Anal. Biochem.* **414,** 294–6 (2011).

13. Navarro, G. A guided tour to approximate string matching. *ACM Comput. Surv.* **33,** 31–88 (2001).

14. Wu, S. & Manber, U. Fast text searching: allowing errors. *Commun. ACM* **35,** 83–91 (1992).

15. Ukkonen, E. Approximate string-matching with q-grams and maximal matches. *Theor. Comput. Sci.* **92,** 191–211 (1992).

16. Lu, C. W., Lu, C. L. & Lee, R. C. T. A new filtration method and a hybrid strategy for approximate string matching. *Theor. Comput. Sci.* **481,** 9–17 (2013).

17. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

2

Chapter 3

Single-Molecule Peptide Fingerprinting

3.1 Introduction

Proteomics provides essential information on molecular pathways of the cell and the state of a living organism (1). Mass spectrometry is currently the first choice for proteomic analysis. However, the requirement for a large amount of sample renders a small-scale proteomics study such as single cell analysis unfeasible. The limit of detection also precludes proteomic analysis of low abundance proteins. In DNA sequencing, such problems are overcome by utilizing polymerase enzymes to amplify a sample. Protein sequencing platforms lack this advantage since such copying machinery for proteins has not been discovered. Recently developed single-molecule techniques allow for detection of low-copy number molecules (2). We demonstrate the first proof-of-concept of a single-molecule fluorescence peptide analysis, harnessing the AAA+ protease ClpXP that linearizes and scans proteins. Our ClpXP platform exhibits high processivity, uni-directional processing with a constant speed, and two orders of magnitude of dynamic range in sensitivity, making a promising approach to sequence full-length protein substrates.

Single-molecule techniques are cutting-edge detection tools to study biological processes at the nanoscale and are well-suited for analysis of samples containing target molecules with low copy numbers. Several potential single-molecule approaches to protein sequencing have recently been explored. Rosen et al used α-hemolysine nanopores to distinguish between non-phosphorylated and phosphorylated variants of thioredoxin (3). A comparable biological nanopore in combination with a motor protein complex, ClpXP, was used by Nivala et al. to control protein translocation through a nanopore (4,5). Zhao et al. and Ohshiro et al identified electronic signatures of free forms of amino-acid molecules using an electronic trap (6,7). Kennedy et al could draw a profile of a full-length protein using sub-nanometer pores (8,9). Although these methods were able to detect certain protein features, none of them could identify the sequence of a protein.

Genuine single-molecule protein sequencing has not yet been achieved due to the

complexity that arises from primary protein sequences. Whereas DNA consists of four building blocks (A, G, C, T), proteins are built from 20 distinct amino acids. Independent of the readout method of choice, full protein sequencing would require the detection of 20 distinguishable signals, which has so far not been demonstrated in single-molecule detection. Recently, we and others have computationally demonstrated that read-out of only a subset of the 20 building blocks is sufficient to identify proteins at the single-molecule level (10,11). In brief, the number of protein species in an organism is finite. Thus, unlike DNA sequencing that requires detection of every single element, identification of a protein sequence needs detection of only a subset of elements (e.g. two types of amino acids). We name this approach "single-molecule protein fingerprinting." Here we demonstrate the first proof of concept of a single-molecule fingerprinter that scans polypeptides and full-length proteins and detects fluorescently labeled amino acids.

To obtain ordered determination of protein sequences, we needed a molecular probe that can linearize and scan a protein in a processive manner. We adopted a naturally existing molecular machinery, the AAA+ protease ClpXP from *Escherichia coli*. The ClpXP protein complex is an enzymatic motor that unfolds and degrades protein substrates. $ClpX_6$ is a homohexameric ring that can exercise a large mechanical force to unfold proteins using ATP hydrolysis (12,13). $ClpX_6$ translocates substrates in a highly processive manner (14,15), with extensive promiscuity towards unnatural substrate modifications including fluorescent labels (16–18). $ClpX_6$ partners with $ClpP_{14}$, a homotetradecameric protease that contains fourteen cleavage sites and self-assembles into a barrel-shaped complex that encloses a central chamber (19).

3.2 Results

3.2.1 Single-molecule fingerprinting platform

For fingerprinting, we labeled two types of reactive groups, the thiol group of

cysteine and the N-terminal amine group, with two different colors of fluorophores. To detect these fluorescently labeled amino acids with nanometer accuracy, we employed FRET (Förster Resonance Energy Transfer). We constructed a FRET scanner by adding a fluorophore (donor) to the ClpP chamber. The ClpP chamber is ~10 nm away from the substrate entry portal of ClpX (**Supplemental Figure S3.1a**) (20,21), which is longer than the Förster radius of a standard single-molecule FRET pair (~5 nm). This physical dimension enables us to selectively detect signals from only the fluorophores (acceptors) on a protein substrate that have been translocated through a ClpX central channel. Polypeptides and titin (our model protein) were labeled with acceptors, and were also tagged with the 11 amino-acid C-terminal ssrA tag to promote recognition by ClpX.

For continuous single-molecule imaging, we biotinylated ClpXP (22) and tethered it to a PEG-coated quartz surface through biotin-streptavidin conjugation (**Figure 3.1a**). A combination of total internal reflection fluorescence microscopy and Alternating Laser EXcitation (ALEX) imaging (23,24) was used to monitor individual ClpXP complexes binding, translocating and degrading dye-labeled substrates in real time. Notably, immobilized ClpXP remained biochemically active for more than 6 hours (**Supplemental Figure 3.1b**). Multiple conditions were measured in the same observation chamber without compromising the results.

Using a peptide, we obtained FRET time traces reporting on translocation as shown in **Figure 3.1b**. The sudden appearance of acceptor signal during acceptor-directed excitation indicates binding of acceptor-labeled peptide to ClpXP (**Figure 3.1b, middle trace, stage ii**). A gradual increase in FRET between a donor on ClpP and the acceptor represents the translocation of the substrate by ClpX into the ClpP chamber (**Figure 3.1b, top and bottom traces, stage iii**). The high FRET state indicates the retention of a dye-labeled region of a peptide within ClpP (**Figure 3.1c**). The dwell time of the high-FRET state is typically on the

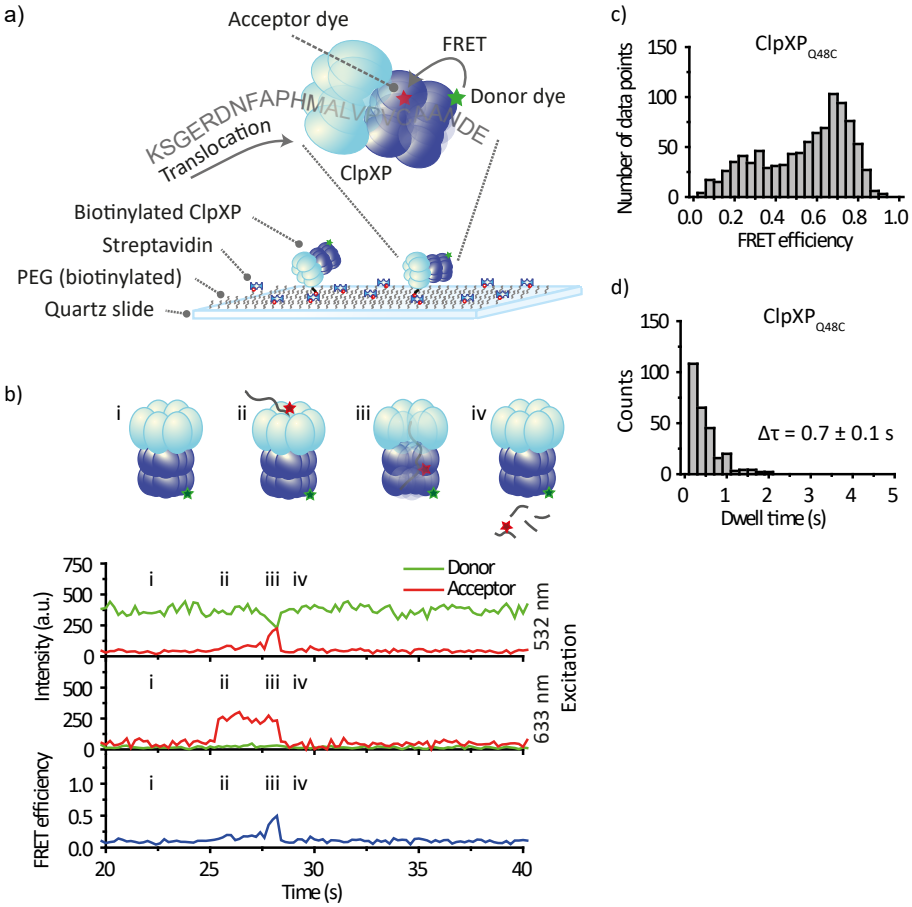


Figure 3.1: (**a**) Schematics of the single-molecule fingerprinting platform. Donor-labeled ClpXP is immobilized on a PEG-coated slide via biotin-streptavidin interactions. Acceptor-labeled substrates (K-38-C-ssrA with Cy5 at Cys, **Supplemental table 3.1**) are introduced to the microfluidic chamber. ClpX recognizes the substrate and translocate it into the ClpP chamber within which FRET occurs between donor and acceptor. (**b**) A typical fluorescence time trace. (**i**) The donor signal is from Cy3-labeled ClpXP (upper trace). (**ii**) The sudden appearance of acceptor signal during acceptor-direct excitation indicates binding of acceptor-labeled substrate to ClpXP (middle trace). (**iii**) A gradual increase in FRET represents the translocation of the substrate by ClpX into ClpP (bottom trace). (**iv**) Loss of signal represents the release of the substrate. (**c**) (Top) Dwell time of region (**iii**). (Bottom) FRET distribution of region (**iii**). Error represents SEM (n = 239).

order of one second (**Figure 3.1d**), which duration allows for unobscured identification of an acceptor molecule by FRET. Loss of signal occurs upon the release of the cleaved peptide fragment (**Figure 3.1b, stage iv**). Time traces displaying

this behavior were considered productive translocation events. See **Chapter 4** of this thesis for a statistical analysis of productive versus abortive translocation events.

3.2.2 ClpP engineering for sequencing scheme

We labeled ClpP with maleimide-functionalized fluorophores by introducing cysteine residues to three different locations in ClpP (**Supplemental Figure S3.1a**, see **Chapter 5** for more details)**.** To select the most promising mutant for our sequencing scheme, we immobilized donor (Cy3) labeled ClpP mutants, in complex with biotinylated ClpX, on a microscope slide and added acceptor (Cy5) labeled Titin. For these experiments we used $Titin_{V13P}$, a less stable mutant of the i27 domain of Titin. In addition, $Titin_{V13P}$ was functionalized with the 11 amino acid long C-terminal ssrA-tag, to promote recognition by ClpX. From all three mutants we could obtain typical time traces as depicted by **Figure 3.1b**, where an event starts with binding of an acceptor (Cy5) labeled substrate to ClpXP, after initial binding the substrate is translocated by ClpX into the ClpP chamber resulting in a high-FRET peak, loss of signal represents release of the substrate. Time traces displaying this type of behavior were considered successful translocation events and the results throughout this paper are based on these events.

In order for our sequencing scheme to work, the FRET events we observe should be short and distinct from the background signal. FRET histograms obtained from the time traces described above show broad distributions for $ClpP_{A139C}$ and $ClpP_{Q48C}$, as we expect from the gradual increase in FRET we observe. For $ClpP_{F31C}$, however, we observe a major population showing low FRET values (**Supplemental Figure S3.1c**). This is most likely a result from the position of the dye being at the far end of the ClpP chamber.

Next, if we compare the width of the high FRET peaks in the individual time traces, $ClpP_{F31C}$ and $ClpP_{Q48C}$ show short dwell times of 1.4 ± 0.2 s and 1.4 ± 0.3 s respectively. The dwell time of the high FRET state of $ClpP_{A139C}$ on the other

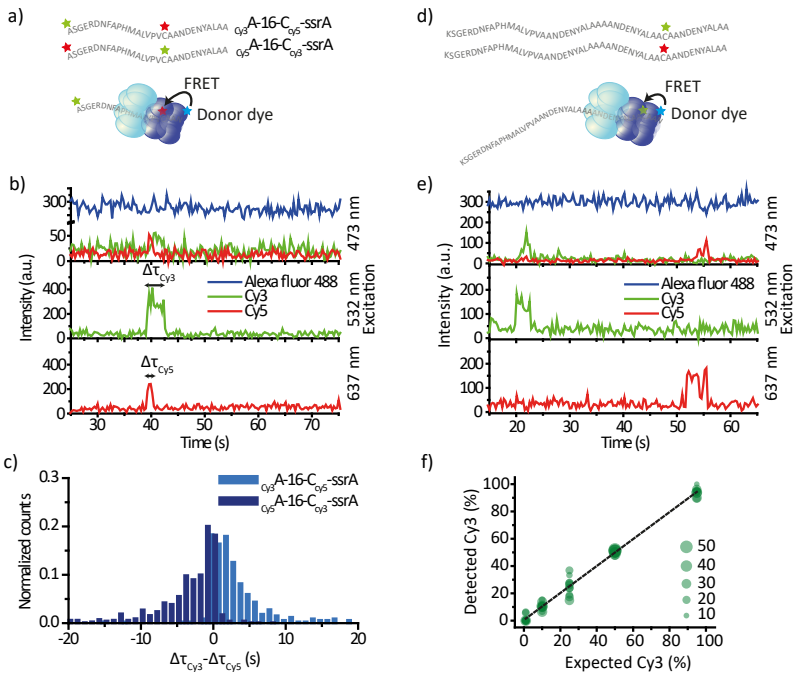


Figure 3.2: (**a**) Substrates (A-16-C-ssrA, **Supplemental table S3.1**) were labeled with Cy3 fluorophore at the N-terminal end and with Cy5 internally. (**b**) A typical time trace obtained from a three-color ALEX measurement showed FRET between Alexa488 and Cy3; and Alexa488 and Cy5 (top trace) and a colocalized signal from Cy3 (middle trace) and Cy5 (bottom trace) excitation. For presentation of multiple traces with clarity, we applied an arbitrary offset of 200 A.U. to the Alexa488 trace (top). We also plotted the sum of the signals from Cy3 and Cy5 upon 532 nm excitation (middle). For the original trace, see Supplemental Figure S2a. (**c**) Comparison of the dwell times of the two acceptor signals. We subtracted the dwell time of Cy5 (ΔτCy5) from the dwell time of Cy3 (ΔτCy3) for each event (blue). The same analysis was made for a construct that had the two acceptors swapped (black). (See also Supplemental Figure S2b) Approximately 900 events from three independent experiments were used for each analysis. (**d**) Substrates (K-38-C-ssrA) were labeled with either Cy3 or Cy5 (acceptors). (**e**) A representative time trace of three-color FRET. Two batches of polypeptides labeled with either Cy3 or Cy5 were mixed in different ratios. Both Cy3 and Cy5 labeled polypeptides showed FRET (top). At t ~ 20 s a Cy3-labeled substrate binds, as reported by Alexa488-Cy3 FRET (top) and 532nm direct excitation (middle). At t ~50 s, a Cy5-labeled substrate binds reported by Alexa488-Cy5 FRET (top) and 637nm direct excitation (bottom). For presentation of multiple traces with clarity, we applied an arbitrary offset of 200 A.U. to the Alexa488 trace. (**f**) The percentage of processed Cy3-labeled substrates compared to the total number of all processed substrates was plotted against the expected percentage of Cy3-labeled substrate in the mixture. Each dot is based on a 100 sec measurement and represents between 8 and 57 translocation events. The number of events is represented by the size of the dot. The solid black line is a linear fit (slope of 0.99 ± 0.02, intercept 0.44 ± 0.77, R2 = 0.99).

hand is much longer, 2.8 ± 0.6 s, and does not show true single exponential behavior (**Supplemental Figure S3.1c**). This suggests that due to the position of the dye, this mutant is not only reporting on the substrates presence in the ClpP chamber, but also on part of the prior translocation process. In addition, labeling of $ClpP_{A139C}$ resulted in 1 dye per $ClpP_{14}$ whereas labeling of $ClpP_{Q48C}$ resulted in ~6 dyes per $ClpP_{14}$. This potentially allows for longer observation time before photobleaching occurs, while not interfering with the quality of the FRET signal. These results led us to continue our work with $ClpP_{Q48C}$.

3.2.3 Single-molecule protein fingerprinting

Our single-molecule protein fingerprinting scheme requires detection of the order of fluorophores on a single substrate. To demonstrate fingerprinting, we functionalized a polypeptide with one type of fluorophore at the N-terminal site and a second type of fluorophore on an internal cysteine residue. We monitored the order in which the two fluorophores (Cy3 and Cy5) passed through Alexa488-labeled ClpP (**Figure 3.2a**). **Figure 3.2b** depicts a typical time trace obtained from a substrate ($_{Cy3}$A-16-C_{Cy5}-ssrA) in which the internal Cy5 fluorophore was processed before the Cy3 fluorophore on the N-terminal site of the polypeptide. The simultaneous appearance of Cy3 and Cy5 signals upon direct excitation with 532 nm and 637 nm (t ~ 39 s) signifies binding of a substrate containing both labels. Thereafter, the increase in FRET between Alexa488 and Cy3/Cy5 dyes in blue excitation represents the translocation of the substrate by ClpX into the ClpP chamber.

The positions of the Cy3 and Cy5 fluorophores relative to the ssrA tag on the substrate should dictate the order of Alexa488-Cy3 FRET and Alexa488-Cy5 FRET signals and also the duration of each acceptor signal. From time traces that showed FRET, we extracted how long Cy3 and Cy5 acceptor fluorophores were engaged with ClpXP ($\Delta\tau_{Cy3}$, $\Delta\tau_{Cy5}$). We observed positive differences in dwell time ($\Delta\tau_{Cy3\text{-}Cy5} = \Delta\tau_{Cy3} - \Delta\tau_{Cy5}$) for a substrate with N-terminal Cy3-labeling and internal Cy5-labeling (**Figure 3.2c, blue,** $_{Cy3}$A-16-C_{Cy5}-ssrA). For a substrate with

exchanged dye positions ($_{Cy5}$A-16-C$_{Cy3}$-ssrA), we observed negative differences (**Figure 3.2c, black** and **Supplemental Figure S3.2b**). Confirming that ClpXP processes ssrA-tagged substrates from the C-terminal to the N-terminal sites, this observation leads us to conclude that the order of detected dyes matches the amino-acid sequence of our substrates. The ordered disappearance of the Cy3 and Cy5 signals further implies that there is no accumulation of uncleaved or partially cleaved substrate within the ClpP chamber, which would otherwise hamper accurate fingerprinting.-

3.2.4 Sensitivity of FRET scanner

For the detection of low abundance proteins, a single-molecule sequencer should perform with high dynamic range. To determine the sensitivity of our assay, we performed a population study in which ClpP was labeled with donor fluorophore (Alexa488) and substrate peptides were labeled with either Cy3 or Cy5 as acceptor fluorophore (**Figure 3.2d-e**). We mixed Cy3- and Cy5-labeled substrates in varying proportions (1:99, 10:90, 25:75, 50:50, 95:5) and quantified the number of translocation events. We observed a linear trend between the percentage of Cy3-labeled substrates we detected versus the expectation with an offset of 0.4 % and a slope of 0.99 (R^2 = 0.9875) (**Figure 3.2f**). We conclude that both FRET pairs are detected with equal sensitivity.

3.2.5 FRET scanner functions processively and at a constant speed

The computational analysis we performed in previous work indicated that the precision of our fingerprinting method would be enhanced if the distance between labeled cysteine and lysine residues could additionally be determined as well as the order (11). A uniform speed of the scanner, represented by ClpX, is crucial to extract distance information. To test whether the observed speed of ClpX, in complex with donor-labeled ClpP, is independent of substrate length, we determined processing times (the dwell time of fluorescence signals emitted by Cy5 labels on substrates, upon direction excitation) for three peptides (29-mer, 40-mer, 51-mer; see **Supplemental table S3.1**) and monomeric and

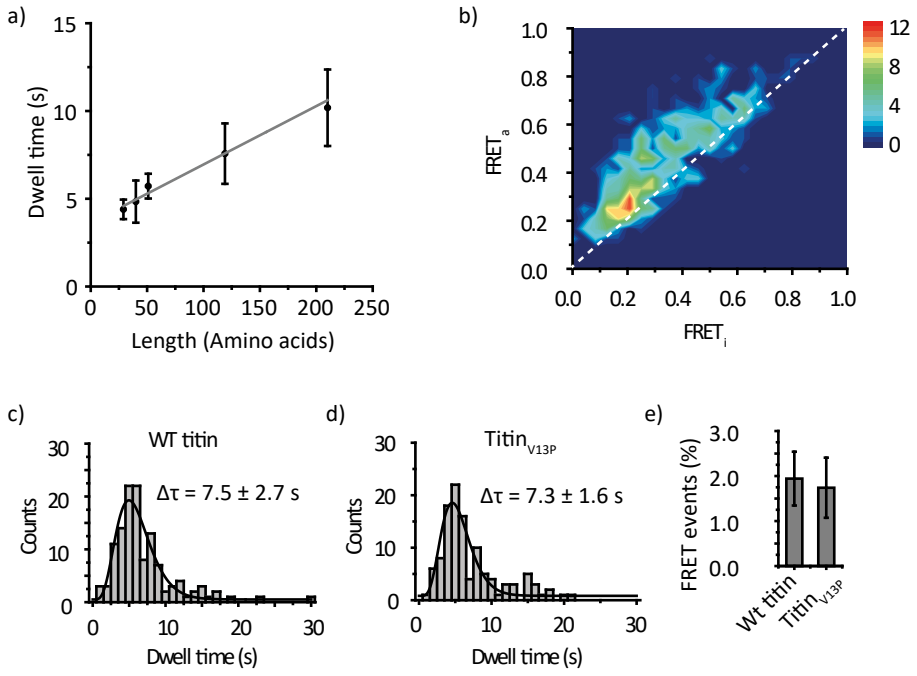


Figure 3.3: (**a**) Dwell times vs. length of the substrate. Total dwell times were obtained from fitting data in Supplemental Figure S3 with gamma distributions. Five different substrates were used: K-16-C-ssrA (n = 227), K-16-C-11-ssrA (n = 131), K-16-C-22-ssrA (n = 290), titin monomer (n = 85), titin dimer (n = 81). Error bars were obtained by bootstrapping with 1000 resamples. Fitting the data with a linear curve results in an offset of 3.6 ± 0.3 s and a speed of 30.1 ± 4.0 AA/s. (**b**) Transition density plot. FRET change was analyzed by measuring $FRET_{initial}$ at t = τ and $FRET_{after}$ at t = τ + δτ, with δτ = 0.4 s, for every time point at stage (iii) (see **Figure 3.1b** for the schematic) and was deposited in a 2D distribution plot. The dotted line represents $FRET_{initial}$ = $FRET_{after}$. A peptide substrate (K-38-C-ssrA) was used. (**c**) and (**d**) Total dwell times (stages ii + iii) (see **Figure 3.1c** for the schematics) for wild-type titin and $titin_{V13P}$. For wild-type titin and $titin_{V13P}$, Δτ = 7.5 ± 2.7 s and 7.3 ± 1.6 s were obtained respectively by fitting with a gamma distribution. Errors were obtained by bootstrapping with 1000 resamples. Wild-type titin, n = 123. $titin_{V13P}$, n = 112. (**e**) The percentage of traces showing FRET events for both wild-type titin and $titin_{V13P}$. Error bars are standard deviations based on 15 measurements.

dimeric versions of titin (all labeled at Cys, see **Supplemental table S3.1**). Plotting the total time a substrate was bound and processed by ClpXP versus the length of the substrates showed a linear increase with a processing velocity of 30.1 ± 4.0 AA/s (**Figure 3.3a and Supplemental Figure S3.3**), which is in agreement with previous results obtained from both bulk (25) and single-molecule

assays (12,13,26). The offset of 3.6 ± 0.3 s reports on the initial docking phase and the eventual retention within ClpP. Our data indicate that the ClpX fingerprinter has a potential to determine both the order and spacing distance of labeled residues. The performance of our fingerprinter would be further enhanced if the translocation speed could be controlled, which would enable us to improve the signal-to-noise ratio of FRET signals. In **Chapter 4** of this thesis, we demonstrate that we can slow down the scanning speed of ClpXP by partly replacing ATP with ATPγS, which ClpXP hydrolyzes up to 90 times more slowly than ATP (14), an important step towards increasing our scanning resolution.

Apart from a constant speed, uni-directional translocation is also of utmost importance. Backtracking of ClpX would result in insertion errors in the observed sequence and thus reduce the detection precision. To evaluate the occurrence of backtracking, we determined the change in FRET over time during processing of peptide substrates using 75 traces. We created a 2D heat map by plotting $FRET_{t=\tau+\delta\tau}$ versus $FRET_{t=\tau}$ for every time point along a time trace reporting on translocation (**Figure 3.3b**). We set $\delta\tau$ = 0.4 s, a time scale longer than our time resolution (0.2 sec) but shorter than the average translocation time (~1 sec). This value resulted in the best sensitivity in analyzing the gradual increase of FRET. Backtracking of ClpX along the substrate would result in momentary FRET decrease during translocation which would appear as $FRET_{t=\tau+\delta\tau}$ values lower than $FRET_{t=\tau}$ (lower diagonal population). We observed $FRET_{t=\tau+\delta\tau} \geq FRET_{t=\tau}$ (upper diagonal population) for a major fraction (92.5 %) of the data points. Therefore backtracking occurs at negligible levels that will not interfere with extracting length information.

A single-molecule protein sequencer should operate against any structural element of a protein. Single-molecule force spectroscopy studies of ClpXP showed that ClpX stalls on substrates with rigid secondary structures (5,27), which would inhibit the extraction of sequence information. Perturbation of cysteine residues of titin has been shown to interfere with the secondary struc-

ture of the protein, making it behave as an unstructured polypeptide chain (28,29). To confirm this, we purified wild-type I27 domain of titin, known to make ClpX stall (27), and V13P titin, a variant that is still folded but is degraded at a rate close to denatured titin (29), and labeled their cysteine residues. We obtained similar total dwell times for processing stable wild-type titin ($\Delta\tau = 7.5 \pm 2.7$ s, **Figure 3.3c**) and titin$_{V13P}$ ($\Delta\tau = 7.3 \pm 1.6$ s, **Figure 3.3d**). A similar percentage of both substrates was processed by ClpXP (**Figure 3.3e**), indicating that ClpX can process wild-type and V13P substrates with identical efficiencies. This result suggests that preparing substrates for sequencing by labeling their cysteine (and likely lysine as well) residues might sufficiently destabilize their protein structures to allow fingerprinting.

3.3 Discussion and conclusions

We have demonstrated a FRET-based detection platform utilizing an AAA+ protease as a scanner of peptide and protein sequences. Our method will be capable of scanning full-length proteins from end to end without the need for fragmentation. Sequencing substrates are processed at a constant speed, allowing for more accurate protein identification (11). In this proof-of-concept study we show our capability to detect populations of differentially labeled substrates as well as our capacity to detect distinct acceptor fluorophores on a single substrate in a sequential manner. The platform we present here has the capability to transform proteomics from a basic research tool to an invaluable asset in clinical diagnostics.

In our approach, we conjugate fluorophores to cysteine and lysine residues because these residues can be labeled with high efficiency and specificity. Our platform, however, is not limited to these two residues. With appropriate chemistry, one could target other residues including tyrosine, arginine or methionine (30). Other options include targeting the N-terminal site or the C-terminal site (31) and also post-translational modifications such as phosphorylation (32) or

glycosylation sites (33). Detection of these moieties could easily be implemented by extending our current three-color FRET scheme to four-color FRET (24).

For accurate proteomics analysis, our sequencing technique has to be able to identify cellular proteins without sequence bias. However, ClpX, the core of our platform, only recognizes substrates displaying specific sequence tags including ssrA. The substrate selectivity of ClpX might be broadened by targeted mutations in the substrate-recognition loops of the ClpX channel (34), or development of non-specific adaptor proteins (35) to deliver protein substrates to ClpXP. An additional challenge for cellular protein analysis is the detection of low-abundance proteins within a complex sample such as a clinical tissue sample. In this case, the depth of sequencing coverage might be increased by pre-separating the proteins in the sample to create multiple sub-samples, removing the influence of relatively over-abundant housekeeping proteins (36).

3.4 Materials and methods

3.4.1 $ClpX_6$ purification and biotinylation

To ensure proper immobilization and hexamer formation of $ClpX_6$ at low concentrations, $ClpX_6(\Delta N)$, a covalently linked hexamer containing a single biotinylation site, was used throughout the experiments. $ClpX_6(\Delta N)$ was overexpressed and purified as described (22). In brief, ClpX protein expression was induced from a *E. coli* BLR(DE3) strain at OD_{600}~0.6 by adding 1.0 mM IPTG and incubated overnight at 18°C. Simultaneously, 100 μM of biotin was added to increase BirA-mediated biotinylation efficiency. Cells were pelleted and resuspended in lysis buffer (20 mM HEPES pH 7.6, 400 mM NaCl, 100 mM KCl, 10% glycerol, 10 mM β-mercaptoethanol, 10 mM imidazole) in the presence of 1mM PMSF and lysed by French press twice at 20 psi. $ClpX_6$ was purified from the supernatant first with Ni^{2+}-NTA affinity resin, followed by size exclusion chromatography with a Prep Sephacryl S-300 16/60 High Resolution column (GE Healthcare).

3.4.2 ClpP mutations, purification and labeling

Point mutations were constructed in ClpP by overlap extension PCR to produce the cysteine-free mutant $ClpP_{C91S-C113S}$, and the subsequent mutants $ClpP_{A139C}$, $ClpP_{F31C}$ and $ClpP_{Q48C}$. Wild-type ClpP and ClpP mutants were overexpressed from *E. coli* BL21(DE3)pLysS at OD_{600}~0.6 by adding 0.5 mM IPTG and incubated for 3 h at 30°C.

Cells were pelleted and resuspended in lysis buffer (50 mM sodium phosphate pH 8.0, 1 M NaCl, 10% glycerol, 5 mM imidazole) in the presence of Set III protease inhibitors (Calbiochem) and lysed by French press twice at 20 psi. ClpP was purified from the supernatant first with Ni^{2+}-NTA affinity resin, followed by size exclusion chromatography with a Prep Sephacryl S-300 16/60 High Resolution column (GE Healthcare). ClpP was dialyzed overnight against PBS (pH 7.4) before labeling for 4 h at 4°C with monoreactive maleimide donor dye (Cy3, GE Healthcare, for two color experiments, and Alexa488, Invitrogen, for three-color experiments). 10x molar dye excess was used in PBS pH 7.4 under nitrogen. Free dye was removed using PD Minitrap G-25 size exclusion columns (GE Healthcare). Labeling efficiencies ranging from 0.1 dye per tetradecamer for wild-type ClpP to 5.7 dyes per tetradecamer for $ClpP_{Q48C}$ were measured by spectrophotometry (DeNovix DS-11 FX).

3.4.3 Substrate preparation

Titin-I27 (wild-type, V13P and dimer) with a C-terminal ssrA tag was expressed from *E. coli* BL21AI at OD_{600}~0.6 by adding 0.2% arabinose and incubated for 4 h at 37°C. Cells were pelleted and resuspended in lysis buffer (50 mM sodium phosphate pH 8.0, 500 mM NaCl, 10 mM imidazole), then lysed by sonication. Titin was purified from the supernatant with Ni^{2+}-NTA affinity resin. Titin was dialyzed overnight against PBS (pH 7.4) before labeling for 4 h at 4°C with 10x molar excess of monoreactive maleimide acceptor dye (Cy5, GE Healthcare) in the presence of 4M GdnCl in PBS pH 7.4 under nitrogen. Custom designed polypeptides were obtained from Biomatik. Cysteine residues of the polypeptides were labeled with monoreactive maleimide-functionalized Cy5 as an acceptor for two-color measurements and with Cy3 and Cy5 as an acceptor for three-color measurements. Polypeptides were labeled in the presence of a 10x molar excess of dye overnight at 4°C in PBS under nitrogen. For labeling with additional acceptors at the N-terminus, monoreactive NHS-ester functionalized dyes (Cy3 or Cy5, GE Healthcare) were added to the reaction mixture described above, also in 10x molar excess. Free dye was removed using PD Minitrap G-25 size exclusion columns (GE Healthcare). Labeling efficiencies as high as 95% were measured by spectrophotometry (DeNovix DS-11 FX). (See **Supplemental table S3.1** for the full list of substrates.)

3.4.4 Single-molecule sample preparation

To reduce the nonspecific binding of proteins, piranha-etched quartz slides (G. Finkenbeiner) were passivated with two rounds of polyethylene glycol (mPEG-Succinimidyl Valerate, MW 5000, Laysan) as described previously (37). After assembly of a microfluidic flow chamber, slides were incubated with 5% Tween-20 for 10 min, and excess of Tween-20 was washed with T50 buffer (10 mM Tris-HCl pH 8.0, 50

mM NaCl), followed by 1 minute incubation with streptavidin (0.1 mg/ml, Sigma). Unbound streptavidin was washed with 100 μL of T50 buffer, followed by 100 μL of PD buffer (25 mM HEPES pH 8.0, 5 mM $MgCl_2$, 40 mM KCl, 0.148% NP-40, 10% glycerol). A $ClpX_6$: $ClpP_{14}$ = 1:3 molar ratio was used to ensure ClpXP complex formation with a 1:1 molar ratio (38). 30 nM ClpX and 90 nM ClpP (either wild-type or mutant) were preincubated for 2 min at room temperature in the presence of 10 mM ATP in PD buffer. After preincubation, the sample was diluted 10 times in PD buffer to reach an expected final ClpXP complex concentration of 3 nM. The diluted sample was applied to the flow chamber and incubated for 1 min. Unbound ClpXP complexes were washed with 100 μL PD buffer containing 1 mM ATP. 10-20 nM of acceptor-labeled substrate was introduced to the flow chamber in the presence of an imaging buffer (0.8% dextrose (Sigma), 1 mg/mL glucose oxidase (Sigma), 170 mg/mL catalase (Merck), and 1 mM Trolox ((±)-6-Hydroxy-2,5,7,8-tetramethylchromane-2-carboxylic acid, 238813), Sigma)). All experiments were performed at room temperature (23 ± 2°C).

3.4.5 Single-molecule fluorescence

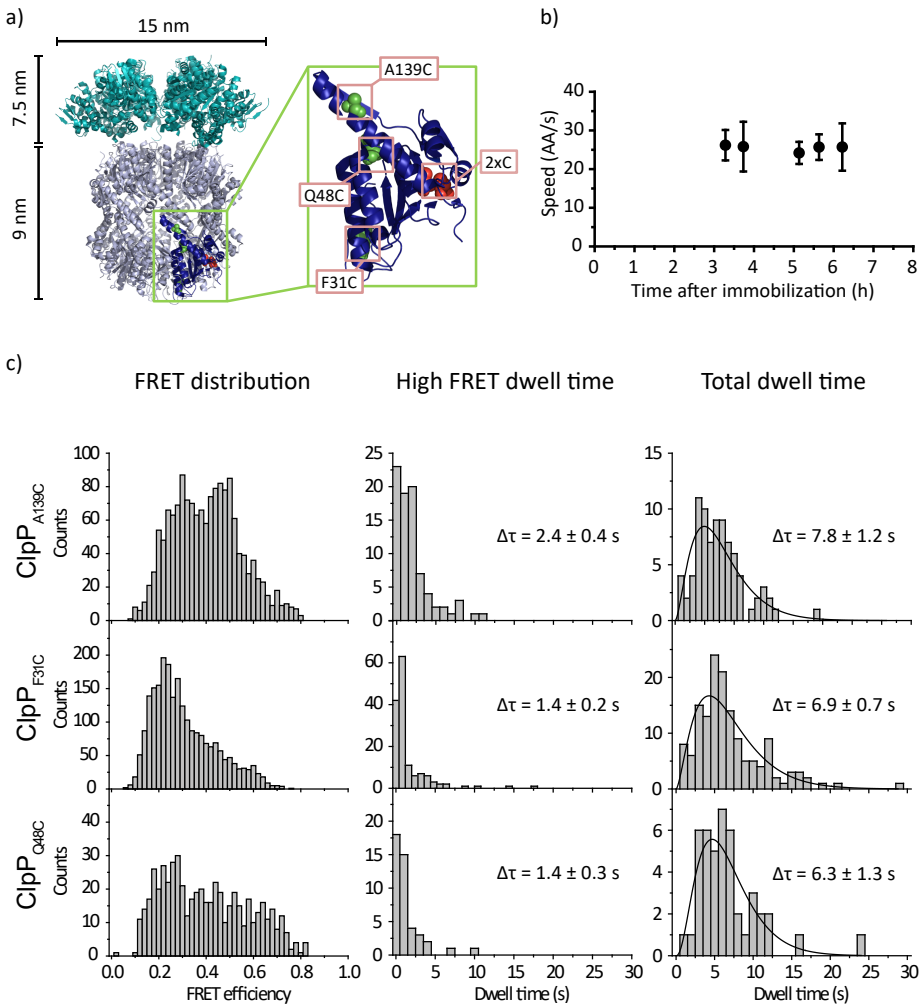
Single-molecule fluorescence measurements were performed with a prism-type total internal reflection fluorescence microscope. For two-color measurements, Cy3 molecules were excited using a 532 nm laser (Compass 215M-50, Coherent), and Cy5 molecules were excited using a 633 nm laser (25 LHP 928, CVI Melles Griot). Fluorescence signals from single molecules were collected through a 60x water immersion objective (UplanSApo, Olympus) with an inverted microscope (IX71, Olympus). Scattered light from the 532 nm and 633 nm laser beams was blocked by a triple notch filter (NF01-488/532/635, Semrock). The Cy3 and Cy5 signals were separated with a dichroic mirror (635 dcxr, Chroma) and imaged using an EM-CCD camera (Andor iXon 897 Classic, Andor Technology).

For three-color measurements, Alexa488 molecules were excited using a 473 nm laser (OBIS 473 nm LX 75 mW, Coherent), Cy3 molecules were excited using a 532 nm laser (Sapphire 532nm-100 CW, Coherent), and Cy5 molecules were excited using a 637 nm laser (OBIS 637 nm LX 140 mW, Coherent). Fluorescence signals from single molecules were collected through a 60x water immersion objective (UplanSApo, Olympus) with an inverted microscope (IX73, Olympus). The 473 nm laser beam was blocked by a 473 nm long pass filter (BLP01-473R-25, Semrock), the 532 nm laser beam was blocked by a 532 nm notch filter (NF03-532E-25, Semrock), and the 637 nm laser beam was blocked by a 633 nm notch filter (NF03-633E-25, Semrock). The Alexa488, Cy3 and Cy5 signals were separated by dichroic mirrors (540dcxr and 635 dcxr, Chroma) and imaged using an EM-CCD camera (Andor iXon 897 Classic, Andor Technology).

3.4.6 Data acquisition

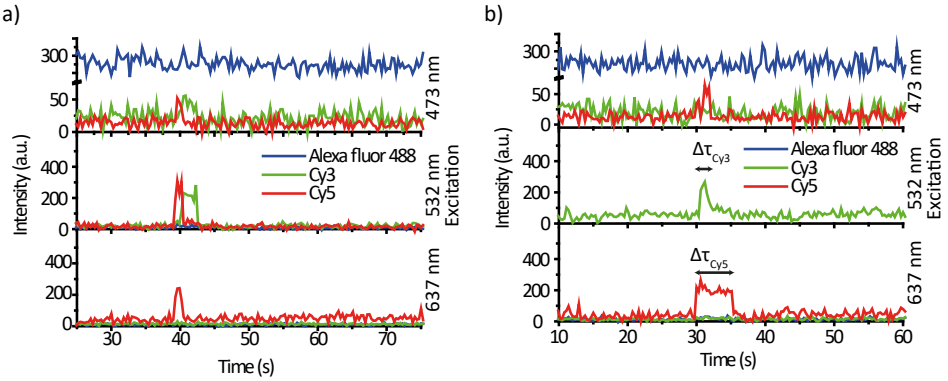
Samples were excited alternatingly with different colors and using a custom-made program written in Visual C++ (Microsoft). A series of CCD images of time resolution 0.1 s was recorded. The time traces were extracted from the CCD image series using IDL (ITT Visual Information Solution) employing an algorithm that identifies fluorescence spots with a defined Gaussian profile and with signals above the average of the background signals. Colocalization between Alexa488, Cy3 and Cy5 signals was carried out with a custom-made mapping algorithm written in IDL. The extracted time traces were processed using Matlab (MathWorks) and Origin (Origin Lab).

3.5 Supplementary data

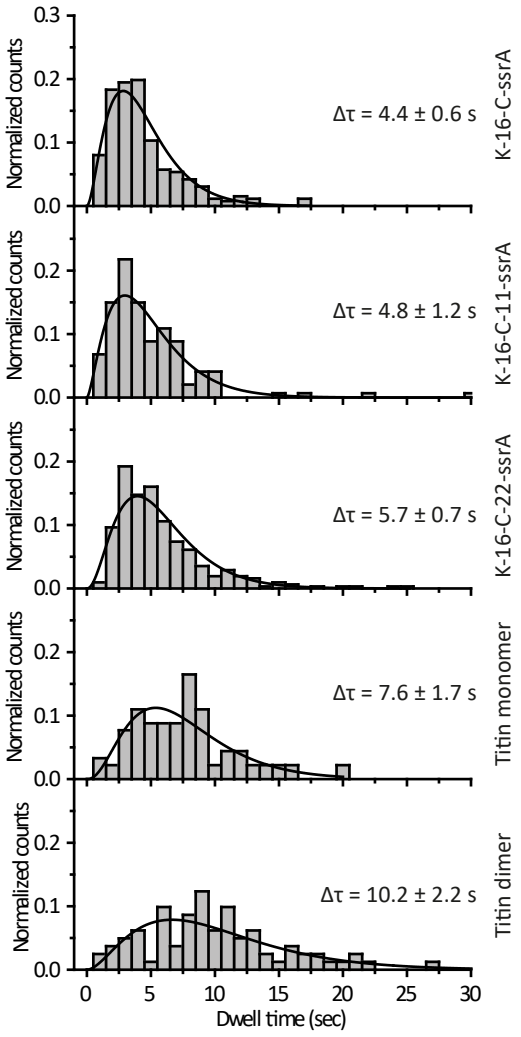


Supplemental figure S3.1: (**a**) Crystal structure of ClpXP. It was obtained by manually combining crystal structures from PDB:1YG6 and PDB:3RY2. Highlighted in red are the two cysteine residues present in wildtype ClpP. Highlighted in green are cysteines introduced in three mutants. (**b**) The translocations speed versus the time of the experiment after ClpXP immobilization. Data obtained from a single flow chamber. ClpXP translocation speeds remained the same up to 6 hours after immobilization. (**c**) FRET and dwelltime analysis. For our sequencing scheme to work, FRET events we observe should be distinct from background signals. Time trajectories showing successful translocation of acceptor (Cy5) labeled titin were selected. (Left) FRET distributions of stage iii (see **Figure 3.1b** for schematic) were

plotted for three ClpP mutants. $ClpXP_{Q48C}$ and $ClpXP_{A139C}$ showed high FRET efficiency. With $ClpXP_{F31C}$ we observe a population showing mainly low FRET values. (Middle) The average dwell time of the high FRET state for $ClpXP_{Q48C}$ was 1.4 ± 0.3 s, shorter than $ClpXP_{A139C}$. From the FRET efficiency and the dwell-time, we chose $ClpXP_{Q48C}$ for developing a single molecule protein analyzer. Labeling efficiency was also taken into consideration (see Material and Methods). For dwell times of high FRET between substrates and $ClpXP_{A139C}$, $ClpXP_{F31C}$ and $ClpXP_{Q48C}$, errors represent SEM. (Right) Total dwell times of substrate binding as reported by direct acceptor excitation. Errors were obtained by bootstrapping with 1000 resamples. $ClpXP_{A139C}$, n = 82; $ClpXP_{F31C}$, n = 136; $ClpXP_{Q48C}$, n = 44.



Supplemental figure S3.2: (**a**) The original time trace used in Figure 2b to present a three-color FRET event. Note the original, not summed up, levels of Cy3 and Cy5 signals in the middle panel. (**b**) A substrate was labeled with Cy5 fluorophore at the N-terminal end and with Cy3 internally. A typical time trace obtained from a three-color ALEX measurement showed a colocalized signal from Cy3 (middle trace) and Cy5 (bottom trace) excitation. For presentation of multiple traces with clarity, we introduced an arbitrary offset of 200 A.U. to the Alexa488 trace (top) and plotted the sum of the signals from Cy3 and Cy5 upon 532 nm excitation (middle).



Supplemental figure S3.3: Total dwell times (stages ii + iii; see **Figure 3.1b** for schematic) were determined for ssrA-tagged polypeptides of increasing lengths and monomeric and dimeric titin. Total dwell times for all substrates showed gamma-like distributions. Errors were obtained by bootstrapping with 1000 resamples.

Name	Sequence	Length (AA)	MW (kDa)	Supplier
A-16-C-ssrA	ASGERDNFAPHMALVPVCAAN-DENYALAA	29	3,018	Biomatik
K-16-C-ssrA	KSGERDNFAPHMALVPVCAAN-DENYALAA	29	3,075	Biomatik
K-16-C-11-ssrA	KSGERDNFAPHMALVPVCAAN-DENYALAAAAANDENYALAA	40	4,180	Biomatik
K-16-C-22-ssrA	KSGERDNFAPHMALVPV-CAANDENYALAAAAANDENYA-LAAAANDENYALAA	51	5,284	Biomatik
K-38-C-ssrA	KSGERDNFAPHMALVPVAANDE-NYALAAAANDENYALAACAAN-DENYALAA	51	5.284	Biomatik
Titin-ssrA	MRGSHHHHHHGLVPRGSLIEVEK-PLYGVEVFVGETAHFEIELSEP-DVHGQWKLKGQPLAASPD-CEIIEDGKKHILILHNCQLGMT-GEVSFQAANTKSAANLKVKELR-SAANDENYALAA	119	13,050	
$Titin_{V13P}$-ssrA	MRGSHHHHHHGLVPRGSLIEVEK-PLYGVEPFVGETAHFEIELSEPDVH-GQWKLKGQPLAASPDCEIIEDGK-KHILILHNCQLGMTGEVSFQA-ANTKSAANLKVKELRSAANDEN-YALAA	119	13,048	
Titin-Titin-ssrA	MRGSHHHHHHGLVPRGSLIEVEK-PLYGVEVFVGETAHFEIELSEP-DVHGQWKLKGQPLAASPD-CEIIEDGKKHILILHNCQLGMT-GEVSFQAANTKSAANLKVKELRS-LIEVEKPLYGVEVFVGETAHFEIEL-SEPDVHGQWKLKGQPLAASPD-CEIIEDGKKHILILHNCQLGMT-GEVSFQAANTKSAANLKVKELR-SAANDENYALAA	210	23,056	

Supplemental table S3.1: Substrates used in this study

3.6 References

1. Muñoz, J. & Heck, A. J. R. From the Human Genome to the Human Proteome. *Angew. Chem. Int. Ed. Engl.* 2–5 (2014).
2. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348,** (2015).
3. Rosen, C. B., Rodriguez-Larrea, D. & Bayley, H. Single-molecule site-specific detection of protein phosphorylation with a nanopore. *Nat. Biotechnol.* 1–3 (2014).
4. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* 1–5 (2013).
5. Nivala, J., Mulroney, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among Protein Variants Using an {Unfoldase-Coupled} Nanopore. *{ACS} Nano* **8,** 12365–12375 (2014).
6. Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* **9,** 466–73 (2014).
7. Ohshiro, T. *et al.* Detection of post-translational modifications in single peptides using electron tunnelling currents. *Nat. Nanotechnol.* **9,** 835–840 (2014).
8. Kolmogorov, M., Kennedy, E., Dong, Z. & Timp, G. Single-Molecule Protein Identification by Sub-Nanopore Sensors. 1–10 (2016).
9. Kennedy, E., Dong, Z., Tennant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm3 resolution using a subnanometre-diameter pore. *Nat. Nanotechnol.* (2016).
10. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **11,** (2015).
11. Yao, Y., Docter, M., van Ginkel, J., de Ridder, D. & Joo, C. Single-molecule protein sequencing through fingerprinting: computational assessment. *Phys. Biol.* **12,** 55003 (2015).
12. Aubin-Tam, M.-E., Olivares, A. O., Sauer, R. T., Baker, T. a & Lang, M. J. Single-Molecule Protein Unfolding and Translocation by an ATP-Fueled Proteolytic Machine. *Cell* **145,** 257–67 (2011).
13. Maillard, R. A. *et al.* ClpX(P) Generates Mechanical Force to Unfold and Translocate Its Protein Substrates. *Cell* **145,** 459–69 (2011).
14. Sen, M. *et al.* The ClpXP protease unfolds substrates using a constant rate of pulling but different gears. *Cell* **155,** 636–46 (2013).
15. Thompson, M. W., Singh, S. K. & Maurizi, M. R. Processive degradation of proteins by the ATP-dependent Clp protease from Escherichia coli. Requirement for the multiple array of active sites in ClpP but not ATP hydrolysis. *J.*

Biol. Chem. **269,** 18209–15 (1994).

16. Barkow, S. R., Levchenko, I., Baker, T. a & Sauer, R. T. Polypeptide translocation by the AAA+ ClpXP protease machine. *Chem. Biol.* **16,** 605–12 (2009).
17. Burton, R. E., Siddiqui, S. M., Kim, Y. I., Baker, T. a & Sauer, R. T. Effects of protein stability and structure on substrate processing by the ClpXP unfolding and degradation machine. *EMBO J.* **20,** 3092–100 (2001).
18. Kolygo, K. *et al.* Studying chaperone–proteases using a real-time approach based on FRET. *J. Struct. Biol.* **168,** 267–277 (2009).
19. Baker, T. a & Sauer, R. T. ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochim. Biophys. Acta* **1823,** 15–28 (2012).
20. Flanagan, J. M., Wall, J. S., Capel, M. S., Schneider, D. K. & Shanklin, J. Scanning transmission electron microscopy and small-angle scattering provide evidence that native Escherichia coli ClpP is a tetradecamer with an axial pore. *Biochemistry* **34,** 10910–7 (1995).
21. Kim, D. Y. & Kim, K. K. Crystal structure of ClpX molecular chaperone from Helicobacter pylori. *J. Biol. Chem.* **278,** 50664–70 (2003).
22. Martin, A., Baker, T. a & Sauer, R. T. Rebuilt AAA + motors reveal operating principles for ATP-fuelled machines. *Nature* **437,** 1115–20 (2005).
23. Kapanidis, A. N. *et al.* Fluorescence-aided molecule sorting: analysis of structure and interactions by alternating-laser excitation of single molecules. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 8936–41 (2004).
24. Lee , J. *et al.* Single-Molecule Four-Color FRET. *Angew. Chemie* **122,** 10118–10121 (2010).
25. 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,** 139–45 (2008).
26. Shin, Y. *et al.* Single-molecule denaturation and degradation of proteins by the AAA+ ClpXP protease. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 19340–5 (2009).
27. Cordova, J. C. *et al.* Stochastic but Highly Coordinated Protein Unfolding and Translocation by the ClpXP Proteolytic Machine. *Cell* **158,** 647–658 (2014).
28. Iosefson, O., Nager, A. R., Baker, T. a & Sauer, R. T. Coordinated gripping of substrate by subunits of an AAA+ proteolytic machine. *Nat. Chem. Biol.* (2015).
29. 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,** 511–20 (2003).
30. McKay, C. S. & Finn, M. G. Click chemistry in complex mixtures: bioorthogonal bioconjugation. *Chem. Biol.* **21,** 1075–101 (2014).

31. Stephanopoulos, N. & Francis, M. B. Choosing an effective protein bioconjugation strategy. *Nat. Chem. Biol.* **7,** 876–84 (2011).

32. Steinberg, T. H. *et al.* Global quantitative phosphoprotein analysis using multiplexed proteomics technology. *Proteomics* **3,** 1128–1144 (2003).

33. Steinberg, T. H. *et al.* Rapid and simple single nanogram detection of glycoproteins in polyacrylamide gels and on electroblots. *Proteomics* **1,** 841–855 (2001).

34. Farrell, C. M., Baker, T. a & Sauer, R. T. Altered specificity of a AAA+ protease. *Mol. Cell* **25,** 161–6 (2007).

35. Davis, J. H., Baker, T. a & Sauer, R. T. Engineering synthetic adaptors and substrates for controlled ClpXP degradation. *J. Biol. Chem.* **284,** 21848–55 (2009).

36. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12,** 483–90 (2008).

37. Chandradoss, S. D. *et al.* Surface passivation for single-molecule protein studies. *J. Vis. Exp.* (2014).

38. Singh, S. K. *et al.* Functional domains of the ClpA and ClpX molecular chaperones identified by limited proteolysis and deletion analysis. *J. Biol. Chem.* **276,** 29420–9 (2001).

3

Chapter 4

Single-Molecule Observation of ClpXP Substrate Recognition

4.1 Introduction

The proteome of a cell is highly dynamic but precisely orchestrated by molecular machineries. AAA+ (ATPases Associated with diverse cellular Activities) enzymes carry out the degradation of damaged, aggregated, and misfolded proteins to ensure strict quality control. They also coordinate the timing of cellular processes through the degradation and remodeling of regulatory proteins (1–3). ClpXP is a bacterial AAA+ protease that can degrade incomplete protein products (4), among other classes of substrates (5). When ribosomes stall, partial protein products are functionalized with a ssrA peptide. This tag is a marker for degradation (6) and is recognized by ClpX (7,8). ClpX is a homohexameric enzyme that utilizes iterative cycles of ATP binding and hydrolysis to unfold ssrA-tagged proteins by pulling them into the axial channel of the hexameric structure. It then transfers these linearized protein substrates into the central chamber of ClpP (9,10), a barrel-shaped protease formed by two heptameric rings. The degradation efficiency of ClpXP depends both on nucleotide cofactors (11,12) and degradation tag sequence (5,13).

Recently, mechanistic insights into the translocation of ClpX along a protein substrate have been obtained from single-molecule optical trapping studies (14–16), which are high spatio-temporal resolution tools for studying protein-protein interactions (2,17–19). However, the process of ClpX substrate recognition and initial engagement into the central pore before translocation occurs has not been explored at the molecular level, leaving the mechanism of initiation not fully comprehended. Here we utilize single-molecule FRET (Förster Resonance Energy Transfer) to understand how ATP cofactors and recognition tags influence ClpX substrate recognition and initial translocation.

4.2 Results

4.2.1 Single-molecule FRET assay to probe substrate binding and processing by ClpXP

We developed a FRET assay to probe substrate recognition and initial translocation by ClpXP. $ClpP_{14}$ was labeled with donor fluorophores (Cy3), while model protein substrates were labeled with acceptor fluorophores (Cy5) (**Chapter 3** of this thesis). Dye-labeled $ClpP_{14}$ was preincubated with a covalently linked version of biotinylated $ClpX^{\Delta N}$ hexamer (20) at a $ClpP_{14}$:$ClpX_6$ = 3:1 molar ratio to ensure ClpXP complex formation at a 1:1 molar ratio (21). The resultant ClpXP complexes were immobilized on a polymer-coated surface through biotin-streptavidin conjugation (**Figure 4.1a**). Acceptor-labeled substrates were introduced to the sample chamber, and individual binding and translocation events were observed using total internal reflection microscopy. The presence of Cy3 donor signal reported on presence of donor labeled ClpP through complex formation with immobilized ClpX (**Figure 4.1**). The initial binding of a substrate to ClpXP was detected by a sudden increase of the Cy5 acceptor signal. These events ended either in dissociation of the substrate, accompanied by total loss of Cy5 signal (**Figure 4.1b**), or in translocation of the substrate into the ClpP chamber, which caused a gradual increase in the FRET signal towards a peak prior to loss of the Cy5 signal (**Figure 4.1c**).

4.2.2 The effect of nucleotide cofactors on ClpXP activity

ClpX uses ATP chemical energy to unfold and translocate its substrates (11). To unravel the role of ATP hydrolysis, ATP analogues such as AMP-PNP and ATPγS have been widely used in ClpXP studies. AMP-PNP is a non-hydrolyzable analogue of ATP that inhibits ClpXP translocation activity (22,23). ATPγS is a poorly hydrolysable ATP analogue that has been shown to permit translocation and degradation by ClpXP (11,16,24), although degradation rates are considerably reduced. Both analogues affect ClpXP processivity; however, it is unknown which steps in binding, translocation and degradation they influence.

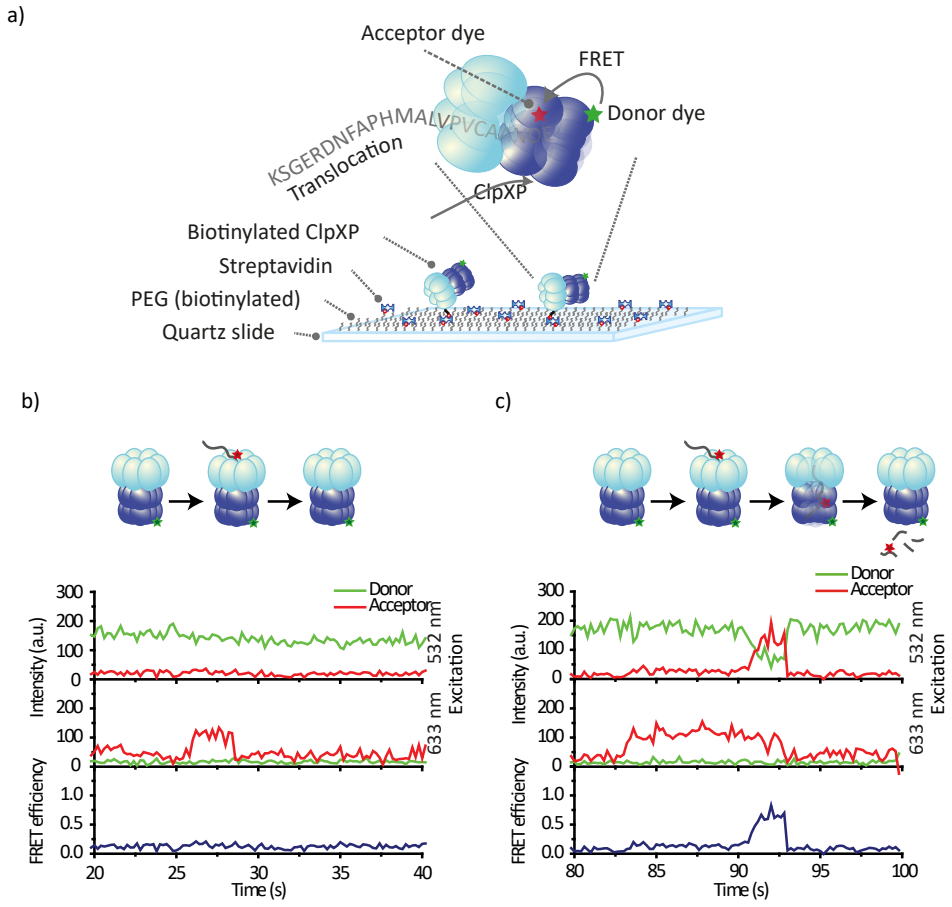


Figure 4.1: (**a**) Schematic representation of the single-molecule FRET assay. Donor labeled ClpXP is immobilized on a passivated quartz slide via biotin-streptavidin interactions. Acceptor labeled substrates with ClpX recognition tags are introduced to the fluidic chamber. ClpX will recognize and translocated the substrate into the ClpP chamber, were FRET will occur between acceptor fluorophore and donor fluorophore. (**b**) A typical time trace of a binding event obtained using ALEX. The donor signal is from Cy3-labeled ClpXP (**Figure 4.1b**, upper graph). The sudden appearance of acceptor signal during acceptor-direct excitation indicates binding of acceptor-labeled substrate to ClpXP (**Figure 4.1b**, middle graph). Loss of signal is the release of the substrate. (**c**) typical time trace for binding events resulting in translocation. Similar to **Figure 4.1b**, the sudden increase in acceptor signal after direct excitation (**Figure 4.1c**, middle graph) reports on binding of a substrate to ClpXP. Translocation into the ClpP chamber is represented by a gradual increase in FRET (**Figure 4.1c**, lower graph).

Stable complex formation between $ClpX_6$ and $ClpP_{14}$ rings requires ATP binding and loss of complex formation was observed after substitution of ATP with ADP (25). In contrary, the effect of AMP-PNP and ATPγS on ClpXP complex formation is unknown. To study the effect of nucleotide binding and hydrolysis on ClpXP complex formation, we preincubated and immobilized ClpXP under various nucleotide conditions. The sample chamber was then washed, such that only dye-labeled $ClpP_{14}$ stably associated with $ClpX_6$ rings would be detectable. A similar number of immobilized complexes was observed in the presence of either ATP, ATPγS, or AMP-PNP. In contrast, after preincubation and immobilization with ADP, or in the absence of nucleotide, ClpXP complex formation was markedly disrupted (**Figure 4.2a**). In a subsequent experiment, ClpXP was preincubated and immobilized in the presence of ATP, after which ADP or nucleotide-free buffer was flowed through to replace or deplete the ATP, respectively. Both ATP depletion and the replacement of ATP with ADP resulted in dissociation of ClpP.

To investigate how ATP analogues influence initial substrate recognition by ClpX, we determined the effect of different nucleotides on the rate of substrate binding to ClpXP. We prepared acceptor-labeled synthetic peptides (K-38-C-ssrA, **Table 4.1**), which were functionalized with an 11 amino-acid ssrA tag to promote ClpX recognition. The peptides were introduced to our immobilized ClpXP complexes, and binding events were monitored by determining colocalization of Cy3 and Cy5 signals after direct excitation. ATP and ATPγS promoted binding of the substrate with approximately equal efficiencies (**Figure 4.2b**). In contrast, AMP-PNP, which was seen to stabilize ClpXP complex formation (**Figure 4.2a**), did not facilitate the binding of substrates (**Figure 4.2b**). Similarly, little substrate binding was seen in the ADP and no-nucleotide conditions, in which poor ClpP binding to ClpX was observed.

To determine whether nucleotide hydrolysis impacts translocation efficiency, we analyzed the percentage of bound substrates that were later translocated by

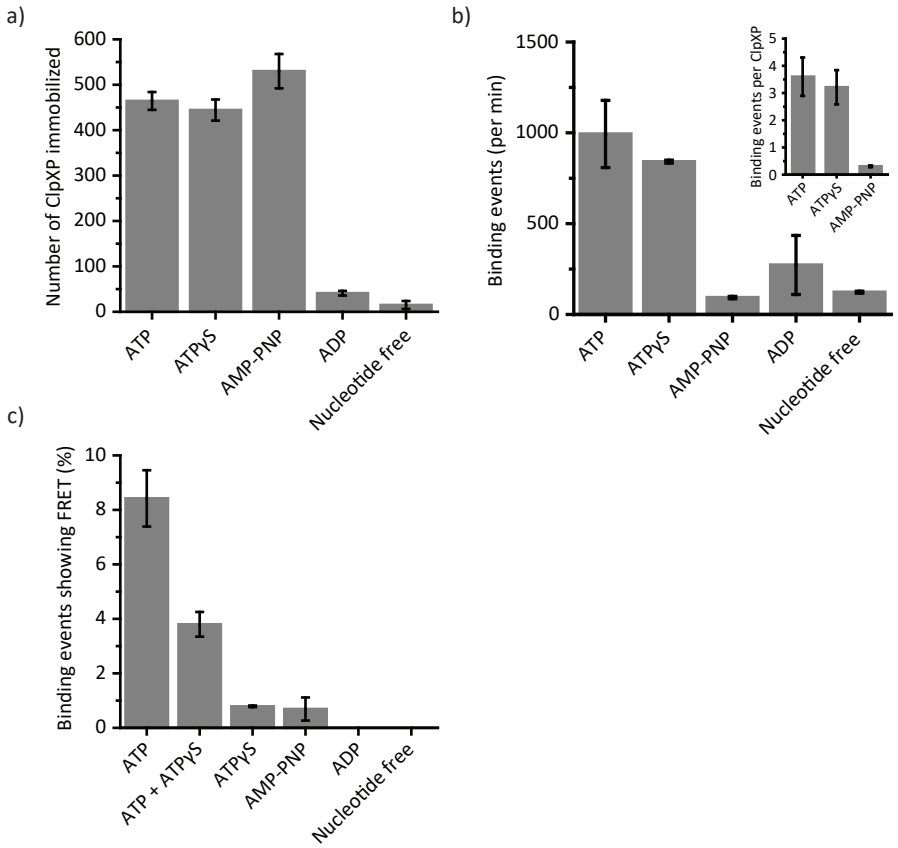


Figure 4.2: (**a**) Number of ClpXP complexes immobilized on the surface in under different nucleotide conditions. (**b**) The total number of acceptor-labeled substrates binding to donor-labeled ClpXP complexes over time. The inset shows the binding events per minute normalized to the number of immobilized ClpXP complexes. Error bars are SEM obtained from 10 measurements. (**c**) The percentage of binding events resulting in translocation of the substrate defined as traces showing a peak in FRET efficiency as shown in **Figure 4.1c.**

ClpXP in the presence of various nucleotide analogues. Binding events resulting in full translocation (**Figure 4.1c**) were distinguished from unprocessed binding events (**Figure 4.1b**) by a gradual increase in FRET prior to dissociation of the substrate. Under ATP conditions, 8.4 ± 1.0 % of all binding events were translocated by ClpXP. Replacing ATP by a mixture of 50% ATP and 50% ATPγS reduced the fraction of translocated substrates to 3.8 ± 0.5 % (**Figure 4.2c**), despite the observation that ATP and ATPγS both led to the same degree of

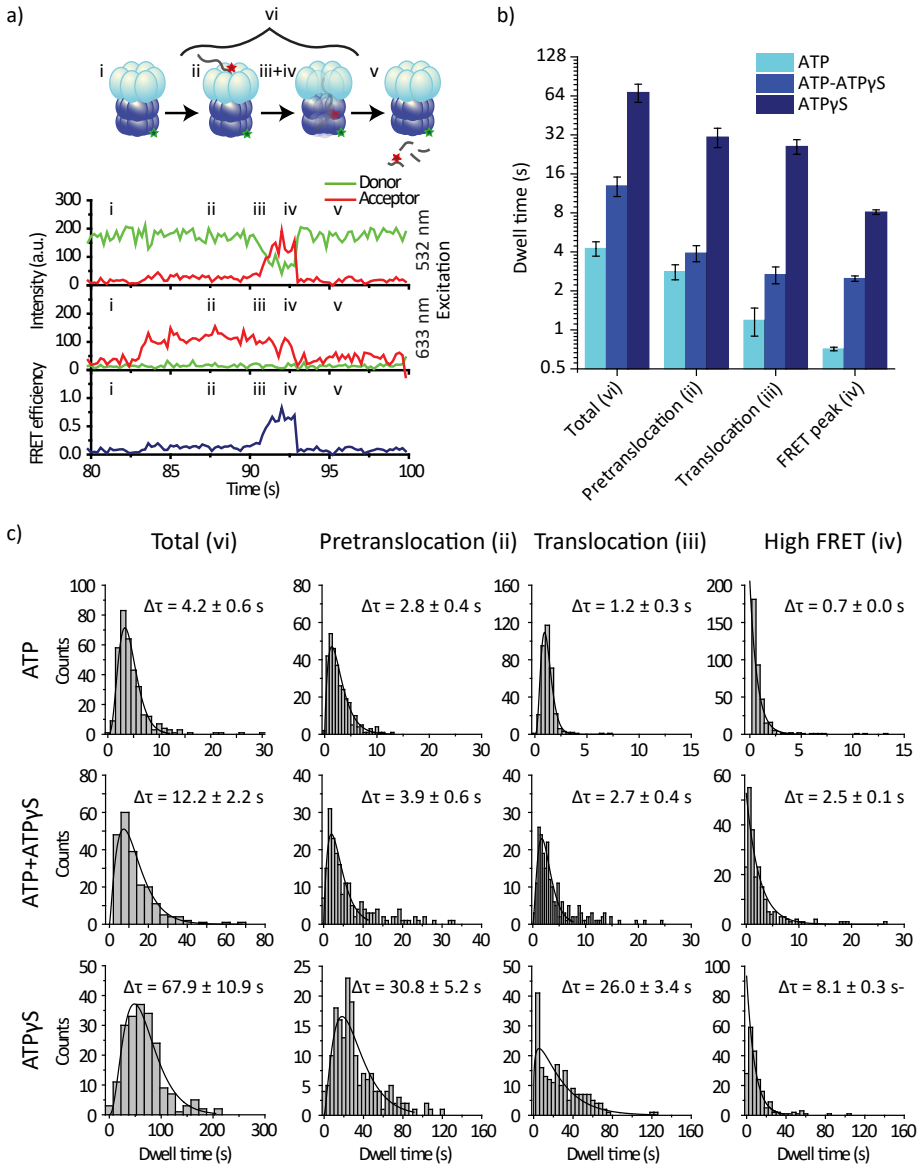


Figure 4.3: (**a**) A typical time trace representing a successful translocation event. **i**) High donor intensity reports on the presence of immobilized and labeled ClpXP. **ii**) Binding of an acceptor-labeled substrate results in a sudden increase in signal after direct acceptor excitation. **iii**) The pretranslocation phase is determined as the time a substrate remains bound to ClpXP prior to translocation. **iv**) Translocation of a substrate by ClpXP results in a gradual increase in FRET efficiency. **v**) Release of the substrate in indicated by a sudden decrease in FRET coinciding with a sudden decrease of direct acceptor excitation signal. **vi**) Represents the total time is takes for a substrate to be processed by ClpXP. (**b**) The dwell times of the

individual features in the traces were determined. (Error bars were obtained by bootstrapping with 1000 resamples.) (**c**) Dwell time histograms underlying the bar graphs in Figure 3b. (Error were obtained by bootstrapping with 1000 resamples.)

initial substrate binding efficiency (**Figure 4.2b**). Complete substitution of ATP by ATPγS further reduced the number of successful translocation events to 0.8 ± 0.03 % (**Figure 4.2c**).

Reducing the concentration of ATP in respect of ATPγS did not influence the number of initial binding efficiency, but had an influence on the number of successful translocation events. To further understand the effect of ATPγS, we identified and analyzed three features in our time traces that corresponded to (ii) the length of time a substrate was bound to ClpXP before translocation was initiated; (iii) the length of time during which the substrate was translocated by ClpX into ClpP, indicated by a gradual increase in FRET; and (iv) the length of time the substrate was retained in ClpP, indicated by the length of the final FRET peak before release of the substrate (v) (**Figure 4.3a**). In addition, the dwell of the total binding event was analyzed (vi). We observed a sixteen-fold increase in total dwell time for ATPγS compared to ATP, which confirmed that ATPγS is slowly hydrolyzed by ClpX. The average time a substrate was bound to ClpXP before translocation was initiated (ii) in the presence of ATP was 2.8 ± 0.4 s, in the presence of 50% ATP and 50% ATPγS resulted in a minor increase to 3.9 ± 0.6 s. An increase by an order of magnitude to 30.8 ± 5.2 s was observed in the presence of ATPγS (**Figure 4.3b+c**). We observed an even more substantial effect of ATPγS on the translocation speed (iii), where translocation occurred in 1.2 ± 0.3s in the presence of ATP and 2.7 ± 0.4 s in the presence of 50% ATP and 50% ATPγS, it took 26.0 ± 3.4 s in the presence of ATPγS. In addition, we observed a loss of processivity of ClpXP in suboptimal nucleotide conditions (**Figure 4.4**). Interestingly, the usage of ATPγS instead of ATP also caused the time that labeled substrates resided within the ClpP chamber (iv) to increase from 0.7 ± 0.0 s for ATP to 8.1 ± 0.3 s for ATPγS. All features we identified in our time traces were influenced by the replacement of ATP by ATPγS, although the

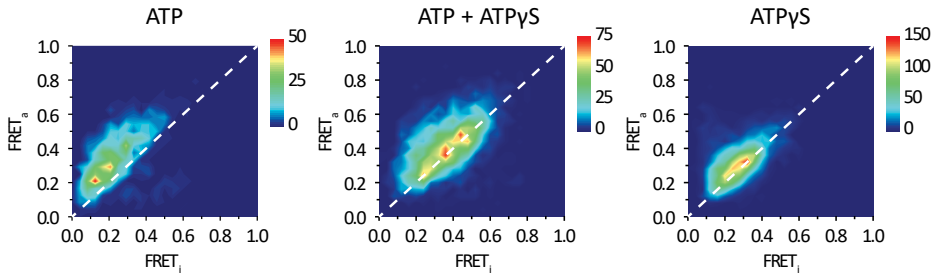


Figure 4.4: Transition density plots. FRET change was analyzed by measuring $FRET_i$ at $t = \tau$ and $FRET_a$ at $t = \tau + \delta\tau$, with $\delta\tau = 0.4$ s, for every time point at stage (iii) (see **Figure 4.3a** for the schematic) and was deposited in a 2D distribution plot. The dotted line represents $FRET_i = FRET_a$. The transition density plots were obtained for ATP, ATP+ATPγS and ATPγS. In the presence of ATP, $FRET_a \geq FRET_i$, indicating a forward movement of ClpXP. For ATP+ATPγS and ATPγS on the other hand, $FRET_a < FRET_i$ was observed, indicating backward movement of ClpXP along the substrate.

effect doesn't seem to scale with ATPγS concentration. Translocation speed (iii) was affected the most by the total substitution of ATP by ATPγS. The presence of a fraction of ATPγS had the smallest impact on pretranslocation time (ii).

4.2.3 The effect of degradation tags on ClpXP activity

ClpX can recognize proteins that contain a C-terminal ssrA-tag and target them for degradation (7). Certain residues in this 11 amino-acid sequence were seen to contribute to the degradation efficiency of ClpXP in bulk solution (13). Residues 1-4 and 7 of the ssrA-tag facilitate binding of the ClpX adaptor protein SspB to the target protein, while residues 9-11 bind ClpX. To determine which features of the ssrA tag are salient for initial recognition by ClpX, we functionalized the I27 domain of human titin with different C-terminal amino-acid sequences and studied the binding, translocation, and release of the substrates by ClpXP (**Figure 4.5a**). For titin-ssrA (wild type), we observed 3.4 ± 0.4 binding events $ClpXP^{-1}\,min^{-1}$. For titin-LAA, which has a three-amino acid tag (LAA) replacing the eleven amino-acid long ssrA tag, we observed a two-fold decrease in the frequency of binding events (1.5 ± 0.2 binding events $ClpXP^{-1}\,min^{-1}$) (**Figure 4.5b**). In the absence of an added degradation tag, 1.2 ± 0.1 binding events per min were observed. For both a variant ssrA tag ending in DAS ("8+DAS"), which is

a)

Name	Sequence
Titin-ssrA	Titin+AANDENYALAA
Titin-LAA	Titin+LAA
Titin-No tag	Titin
Titin-8+DAS	Titin+AANDENYADAS
Titin-8+LDD	Titin+AANDENYALDD

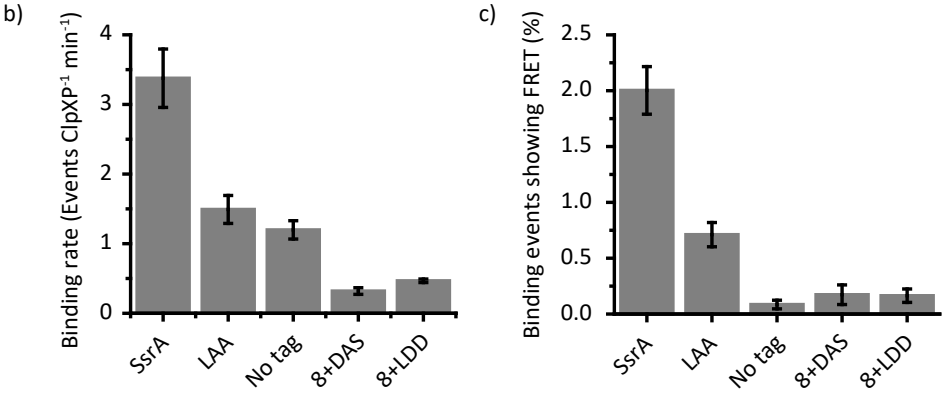


Figure 4.5: (**a**) SsrA tag mutants used in this study. The underlined amino acids are interacting with the ClpX pore to initiate binding and translocation. (**b**) The total number of acceptor-labeled substrates binding per donor-labeled ClpXP complexe per minute. Error bars are SEM obtained from 10 measurements. (**c**) The percentage of binding events resulting in translocation of the substrate defined as traces showing a peak in FRET efficiency as shown in **Figure 4.1c**.

recognized by ClpXP at exceedingly weak affinities unless SspB is present (26), and a tag ending in LDD ("8+LDD"), which inhibits recognition (13), we observed a six- to eightfold decrease in binding rate compared to titin-ssrA (0.3 ± 0.0 and 0.5 ± 0.0 binding events $ClpXP^{-1}\,min^{-1}$ respectively). These results indicate that the first step of substrate discrimination occurs at the initial binding to ClpX.

To determine whether features of the ssrA tag can additionally influence ClpXP translocation efficiency, we analyzed the number of binding events that resulted in translocation into the ClpP chamber (**Figure 4.5c**). We observed a 3-fold decrease in the translocation efficiency of bound substrates for titin-LAA compared to titin-ssrA, indicating that the ssrA-tag not only facilitates initial binding of the substrate, but also has a stabilizing effect

after engagement of the substrate to the ClpX pore. The translocation efficiency decreased ~12 times for titin-DAS and titin-LDD compared to titin-ssrA. Interestingly, the absence of a degradation tag resulted in only a three-fold reduction in binding efficiency; however, it suppressed the translocation efficiency entirely.

For titin-ssrA and titin-LAA we observed time traces with similar features as described in **Figure 4.3a**, we determined the dwell time of each feature for both degradation tags (**Figure 4.6a+b**). Total dwell times for binding, translocation and release of titin-LAA by ClpXP were on average ~3 s shorter compared to dwell times for titin-ssrA. Titin-LAA is a truncated version of titin-ssrA, translocation speed and the duration of the high FRET peak were equal for both substrates, excluding length of the substrate and speed of degradation as explanations for elongated dwell times. The time between initial binding and translocation, however, is strongly reduced for titin-LAA compared to the wild-type ssrA tag. Although binding affinity and translocation efficiency are reduced for titin-LAA, substrate binding does appear to lead to faster processing by ClpXP.

4.3 Discussion

ClpX recognizes target proteins and utilizes ATP chemical energy to translocate substrates through its axial channel into the protease ClpP (1). We developed a single-molecule FRET assay to investigate the substrate recognition by ClpX. ClpXP complex formation requires the presence of ATP or its analogues. We observed robust ClpXP complex formation in the presence of ATP, ATPγS, and AMP-PNP. Although ClpX is known to bind ADP (12,27), complex formation with ClpP, and therefore substrate degradation, were inhibited.

It has been hypothesized that natively folded substrates must make multiple transient interactions with ClpX before proceeding to unfolding and translocation (2,28). These short binding events were, however, not previously observed

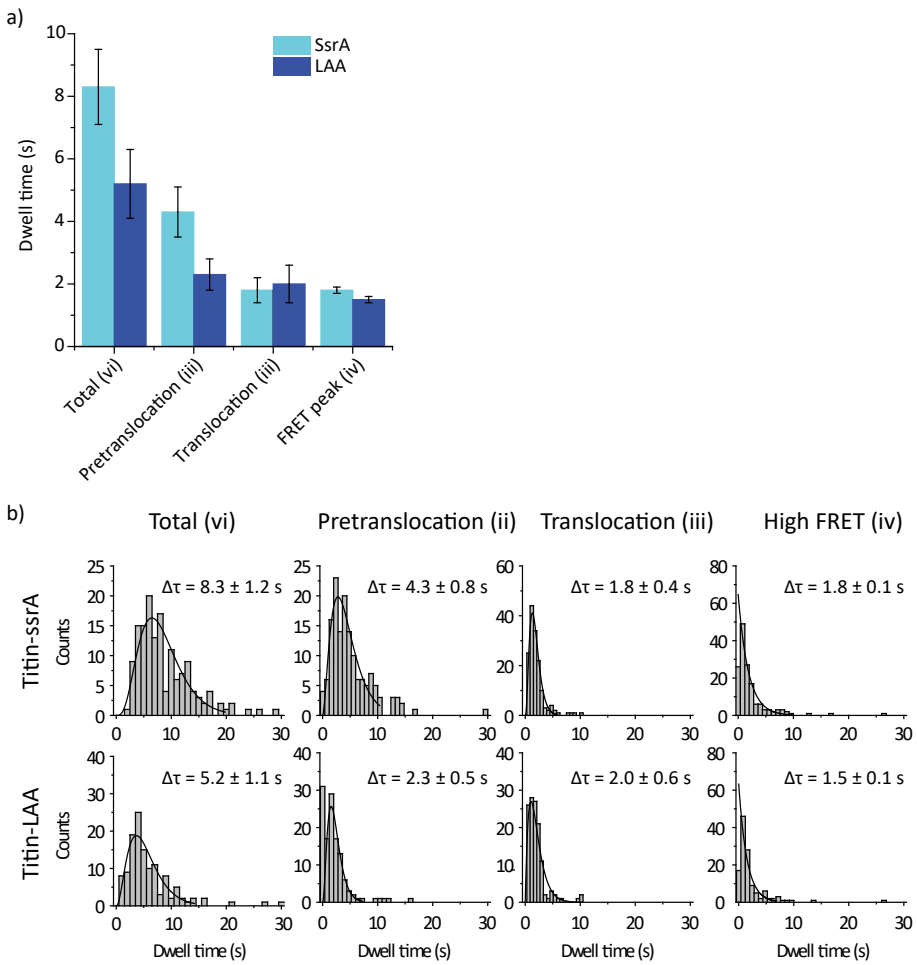


Figure 4.6: (**a**) The dwell times of the individual features in the traces as described in **Figure 4.3a** were determined for titin-ssrA and titin-LAA. (Error bars were obtained by bootstrapping with 1000 resamples.) (**b**) Dwell time histograms underlying the bar graphs in **Figure 4.6a**. (Error were obtained by bootstrapping with 1000 resamples.)

directly. Using 10 nM peptide substrates, we detected 3-6 binding events per minute per ClpXP complex in the presence of ATP. Among these, only ~8% of the binding events were translocated for synthetic polypeptide substrates (**Figure 4.2b**). This number dropped to ~2% for full-length proteins (**Figure 4.5c**). Our observations indicate a low success rate for full translocation and degradation, even though the substrates used in this study are unstructured peptides and titin with modified cysteines, which cause random coil behavior (29). In vivo,

ClpXP activity and selectivity is altered by the use of adaptor proteins (30). The low efficiency we observed in the absence of adaptor proteins potentially serves as a mechanism to prevent nonspecific degradation by ClpXP.

ClpX can bind ATPγS with similar efficiency as it does ATP (11), and binding of ATPγS and ATP both induce similar conformational changes to the ClpX pore (31). In agreement with these findings, substrate-binding rates observed here were not influenced by substitution of ATP with ATPγS. In contrast, subsequent steps of substrate unfolding and translocation are highly ATP-dependent (20). ClpX is capable of hydrolyzing ATPγS, although up to 90 times slower than hydrolysis of ATP (16). As a result, ClpXP complexed with ATPγS can degrade unfolded substrate but stalls on strongly folded substrates such as GFP (32). Here, we observed a 16-fold increase in total dwell time for successful translocation events with ATPγS compared to ATP (**Figure 4.3b**). The dwell time of translocation was most affected and increased 20 fold in the presence of ATPγS compared to ATP. Furthermore, the percentage of binding events resulting in translocation dropped from ~8% to <1%, indicating that slow hydrolysis not only reduces the processing speed but also processivity.

The interaction of ClpXP with AMP-PNP is not well characterized compared to ATPγS. Zhou et al. were able to isolate stable ClpXP-RssB-δ^S complexes in the presence of AMP-PNP (23). We observed equal numbers of immobilized ClpXP complexes in the presence of AMP-PNP compared to ATP, while substrate binding, on the other hand, was largely inhibited and observed binding events were short-lived. This discrepancy with previous studies might be explained by the absence of adaptor proteins in our study, which could potentially compensate for a low substrate affinity induced by AMP-PNP.

Our findings indicate that two different molecular mechanisms are primarily responsible for the observed decrease in ClpXP activity. In the presence of AMP-PNP, ClpXP degradation activity was potently reduced by inhibition of the initial binding of substrates, potentially due to a conformational change in

the ClpX pore, as observed previously for ADP-bound ClpX (12). In the presence of ATPγS, initial binding efficiency was preserved in comparison to the ATP condition. However, translocation efficiency and speed were severely affected. Our results demonstrate that individual ATP analogues do not share the same inhibitory mechanism of ClpXP activity.

We applied the insights gained from the study of nucleotide composition effects on ClpXP activity to study the effect of degradation tag sequence on individual steps in the degradation process. Although efficiencies varied strongly, for all degradation tags we were able to observe translocation events, even for those that are not detectably degraded in bulk solution (26,33). The degradation tag sequence had the strongest influence on binding rate (**Figure 4.5b**). Reduction in binding efficiency and translocation efficiency combined resulted in a seventy- to one-hundred-fold difference in the number of binding events detected for titin-ssrA versus titin-8+LDD, which can explain the apparent absence of degradation of LDD-tagged proteins in bulk solution.

After initial interaction with the RKH loops in ClpX (34), the ssrA tag binds to the pore-1 loops deeper within the axial channel of ClpX (24). The pore-1 loops are proposed to move synchronously during a power stroke, while the pore-2 loops, located near the interface with ClpP, hold the substrate in place in between pore strokes (35). For substrates with a truncated version of the ssrA tag, titin-LAA, we observed reduced binding rates and reduced translocation efficiencies. Interestingly, for the population of translocated titin-LAA substrates we observed shorter overall dwell times compared to titin containing the full ssrA tag. Dissecting the dwell times revealed the difference in total dwell time was caused be a decrease in time between initial engagement of the substrate with ClpX and initiation of translocation. The pore-1 loops are located inside the axial channel, during the pretranslocation phase the recognition tag has to enter inside to a certain extend and be positioned properly to interact with the pore-1 loops. Improved binding and degradation by ClpXP has been observed for elon-

gated ssrA tags in the presence of sspB (36). Potentially, the truncated ssrA tag is too short to be rearranged inside the axial channel and is either binding improperly of being released rapidly. Although we work with purified substrates, there could be heterogeneity in conformation of our substrates. Potentially, we observe a population of titin-LAA with, by chance, a more protruding C-terminus, making it easier for ClpX to grab the substrate.

Our single-molecule FRET assay is a powerful tool to study the process of binding, translocation, and release of substrates. Our study revealed that the majority of substrates binding to ClpXP are not being processed, showing the inefficiency of ClpXP in the absence of adaptor proteins. Combining our current assay with SspB adaptor proteins, or other known adapter/substrate pairs such as UmuD/D' heterodimer, MecA/ComK and RssB/σ^S (37) will give valuable insight in the tightly controlled mechanisms sculpting the proteome of the cell.

4.4 Materials and methods

4.4.1 $ClpX_6$ purification and biotinylation

To ensure proper immobilization and hexamer formation of $ClpX_6$ at low concentrations, $ClpX_6(\Delta N)$, a covalently linked hexamer with a single biotinylation site, was used throughout the experiments. $ClpX_6(\Delta N)$ was overexpressed and purified as described (20). In brief, ClpX protein expression was induced from *E. coli* BLR (DE3) cells at OD_{600}~0.6 by adding 1.0 mM IPTG and incubated overnight at 18°C. Simultaneously, 100μM of biotin was added to increase biotinylation efficiency with wild-type BirA. Cells were pelleted and resuspended in lysis buffer (20 mM of HEPES pH 7.6, 400 mM of NaCl, 100 mM of KCl, 10% of glycerol, 10 mM of β-mercaptoethanol, 10 mM of imidazole) in the presence of 1mM PMSF and lysed by French press twice at 20 psi. $ClpX_6$ was purified from the supernatant first with Ni^{2+}-NTA affinity resin, followed by size exclusion chromatography with Prep Sephacryl S-300 16/60 High Resolution column (GE Healthcare).

4.4.2 ClpP mutations, purification and labeling

Point-mutations in ClpP to produce the cysteine free mutant ClpP(C91S;C113S), and the subsequent ClpP(Q48C) were constructed by overlap extension PCR. Wild-type ClpP and ClpP mutants were overexpressed from *E. coli* BL21(DE3)pLysS cells at OD_{600}~0.6 by adding 0.5 mM IPTG and incubated for 3h at 30°C. Cells were pelleted

and resuspended in lysis buffer (50 mM of sodium phosphate pH 8.0, 1 M of NaCl, 10% of glycerol, 5 mM of imidazole) in the presence of Set III protease inhibitors (Calbiochem) and lysed by French press twice at 20 psi. ClpP was purified from the supernatant first with Ni^{2+}-NTA affinity resin, followed by size exclusion chromatography with Prep Sephacryl S-300 16/60 High Resolution column (GE Healthcare). ClpP was dialyzed overnight against PBS (pH 7.4) before labeling for 4 h at 4°C with monoreactive maleimide donor dye (Cy3, GE Healthcare), 10x molar dye excess was used in PBS pH 7.4 under nitrogen. Free dye was removed using PD Minitrap G-25 size exclusion columns (GE Healthcare). Labeling efficiencies of 5.7 dyes per tetradecamer for ClpP(Q48C) were measured by spectrophotometry (DeNovix DS-11 FX).

4.4.3 Substrate preparation

Mutations in titin-I27 to produce titin-ssrA, titin-LAA, titin-8+DAS and titin-8+LAA were constructed by overlap extension PCR. Titin and titin mutants were expressed from *E. coli* BL21-AI at OD_{600} ~0.6 by adding 0.2% arabinose and incubated for 4h at 37°C. Cells were pelleted and resuspended in lysis buffer (50 mM of sodium phosphate pH 8.0, 500 mM of NaCl, 10 mM of imidazole) lysed by sonication. Titin was purified from the supernatant with Ni^{2+}-NTA affinity resin. Titin was dialyzed overnight against PBS (pH 7.4) before labeling for 4 h at 4°C with 10x molar excess of monoreactive maleimide acceptor dye (Cy5, GE Healthcare) in the presence of 4M GdnCl in PBS pH 7.4 under nitrogen. Custom designed polypeptides were obtained from Biomatik and with monoreactive maleimide functionalized Cy5 as an acceptor. Polypeptides were labeled in the presence of a 10x molar excess of dye overnight at 4°C in PBS under nitrogen. For specific labeling with two acceptors on one substrate, monoreactive NHS-ester functionalized dyes (Cy3 or Cy5, GE Healthcare) were added to the reaction mixture described above, also in 10x molar excess. Free dye was removed using PD Minitrap G-25 size exclusion columns (GE Healthcare). (See **Table 4.1** for the full list substrates.)

4.4.4 Single-molecule sample preparation

To reduce the nonspecific binding of proteins, piranha-etched quartz slides (G. Finkenbeiner) were passivated with two rounds of polyethylene glycol (mPEG-Succinimidyl Valerate, MW 5000, Laysan) as describe previously (38). After assemble of a microfluidic flow chamber, slides were incubated with 5% Tween-20 for 10 min, access of Tween-20 was washed with T50 buffer (10 mM Tris-HCl pH 8.0, 50 mM NaCl) followed by 1 minute incubation with Streptavidin (0.1 mg/ml, Sigma). Unbound Streptavidin was washed with 100 µl of T50 buffer, followed by 100µL of PD buffer (25 mM HEPES pH 8.0, 5 mM $MgCl_2$, 40 mM KCl, 0.148% NP-40, 10% glycerol). We used a $ClpX_6$: $ClpP_{14}$ = 1:3 molar ratio to ensure ClpXP complex forma-

tion with a 1:1 molar ratio (Singh et al., 2001), 30 nM ClpX and 90 nM ClpP (either wild-type or mutant) were preincubated for 2 min at room temperature in the presence of 10 mM ATP in PD buffer. After preincubation, the sample was diluted 10 times in PD buffer to reach an expected final concentration of ClpXP complexes of 3 nM. The diluted sample was applied to the flow chamber incubated for 1 min. Unbound ClpXP complexes were washed with 100 μL PD buffer containing 1 mM ATP. 10-20 nM of acceptor labeled substrate was introduced to the flow chamber in the presence of an imaging buffer that consisted of 0.8% dextrose (Sigma), 1 mg/ml glucose oxidase (Sigma), 170 mg/ml catalase (Merck), and 1 mM Trolox ((±)-6-Hydroxy-2,5,7,8-tetramethylchro- mane-2-carboxylic acid, 238813, Sigma). All experiments were performed at room temperature (23 ± 2°C).

4.4.5 Single-molecule fluorescence

Single-molecule fluorescence measurements were performed with a prism-type total internal reflection fluorescence microscope. Cy3 molecules were excited using a 532 nm laser (Compass 215M-50, Coherent), and Cy5 molecules were excited using a 633 nm laser (25 LHP 928, CVI Melles Griot). Fluorescence signals from single molecules were collected through a 60x water immersion objective (UplanSApo, Olympus) with an inverted microscope (IX71, Olympus). The 532 nm and 633 nm laser beams were blocked by a triple notch filter (NF01-488/532/635, Semrock). The Cy3 and Cy5 signals were separated with a dichroic mirror (635 dcxr, Chroma) and imaged using an EM-CCD camera (Andor iXon 897 Classic, Andor Technology).

4.4.6 Data acquisition

Samples were excited alternatingly with two or three colors and using a custom-made program written in Visual C++ (Microsoft), a series of CCD images of time resolution 0.1 s was recorded. The time traces were extracted from the CCD image series using IDL (ITT Visual Information Solution) employing an algorithm that looked for fluorescence spots with a defined Gaussian profile and with signals above the average of the background signals. Colocalization between Cy3 and Cy5 signals was carried out with a custom-made mapping algorithm written in IDL. The extracted time traces were processed using Matlab (MathWorks) and Origin (Origin Lab).

Name	Sequence	Length (AA)	MW (kDa)	Supplier
K-38-C-ssrA	KSGERDNFAPH-MALVPVAANDENYA-LAAAANDENYALAACAAN-DENYALAA	51	5.284	Biomatik
Titin-ssrA	MRGSHHHHHHGLVPRG-SLIEVEKPLYGVEVFVGETAH-FEIELSEPDVHGQWKLKGQ-PLAASPDCEIIEDGKKHIL-ILHNCQLGMTGEVSFQA-ANTKSAANLKVKELRSAAN-DENYALAA	119	13.050	
Titin-LAA	MRGSHHHHHHGLVPRG-SLIEVEKPLYGVEVFVGETAH-FEIELSEPDVHGQWKLKGQ-PLAASPDCEIIEDGKKHIL-ILHNCQLGMTGEVSFQA-ANTKSAANLKVKELRSLAA	111	12.200	
Titin	MRGSHHHHHHGLVPRG-SLIEVEKPLYGVEVFVGETAH-FEIELSEPDVHGQWKLKGQ-PLAASPDCEIIEDGKKHIL-ILHNCQLGMTGEVSFQA-ANTKSAANLKVKELRS	108	11.950	
Titin-8+DAS	MRGSHHHHHHGLVPRG-SLIEVEKPLYGVEVFVGETAH-FEIELSEPDVHGQWKLKGQ-PLAASPDCEIIEDGKKHIL-ILHNCQLGMTGEVSFQA-ANTKSAANLKVKELRSAAN-DENYADAS	119	13.070	
Titin-8+LDD	MRGSHHHHHHGLVPRG-SLIEVEKPLYGVEVFVGETAH-FEIELSEPDVHGQWKLKGQ-PLAASPDCEIIEDGKKHIL-ILHNCQLGMTGEVSFQA-ANTKSAANLKVKELRSAAN-DENYALDD	119	13.140	

Table 4.1: Substrates used in this study.

4.5 References

1. Baker, T. a & Sauer, R. T. ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochim. Biophys. Acta* **1823,** 15–28 (2012).
2. Olivares, A. O., Baker, T. A. & Sauer, R. T. Mechanistic insights into bacterial AAA+ proteases and protein-remodelling machines. *Nat. Rev. Microbiol.* 1–12 (2015).
3. Sauer, R. T. & Baker, T. a. AAA+ proteases: ATP-fueled machines of protein destruction. *Annu. Rev. Biochem.* **80,** 587–612 (2011).
4. Moore, S. D. & Sauer, R. T. Ribosome rescue: tmRNA tagging activity and capacity in Escherichia coli. *Mol. Microbiol.* **58,** 456–466 (2005).
5. Flynn, J. M., Neher, S. B., Kim, Y. I., Sauer, R. T. & Baker, T. a. Proteomic discovery of cellular substrates of the ClpXP protease reveals five classes of ClpX-recognition signals. *Mol. Cell* **11,** 671–83 (2003).
6. Keiler, K. C., Waller, P. R. & Sauer, R. T. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science* **271,** 990–3 (1996).
7. 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,** 1338–47 (1998).
8. Wojtkowiak, D., Georgopoulos, C. & Zylicz, M. Isolation and characterization of ClpX, a new ATP-dependent specificity component of the Clp protease of Escherichia coli. *J. Biol. Chem.* **268,** 22609–17 (1993).
9. Hwang, B. J., Woo, K. M., Goldberg, a L. & Chung, C. H. Protease Ti, a new ATP-dependent protease in Escherichia coli, contains protein-activated ATPase and proteolytic functions in distinct subunits. *J. Biol. Chem.* **263,** 8727–34 (1988).
10. Wang, J., Hartling, J. a & Flanagan, J. M. Crystal structure determination of Escherichia coli ClpP starting from an EM-derived mask. *J. Struct. Biol.* **124,** 151–63 (1998).
11. Burton, R. E., Baker, T. A. & Sauer, R. T. Energy-dependent degradation: Linkage between ClpX-catalyzed nucleotide hydrolysis and protein-substrate processing. *Protein Sci.* **12,** 893–902 (2003).
12. Hersch, G. L., Burton, R. E., Bolon, D. N., Baker, T. a & Sauer, R. T. Asymmetric interactions of ATP with the AAA+ ClpX6 unfoldase: allosteric control of a protein machine. *Cell* **121,** 1017–27 (2005).
13. Flynn, J. M. *et al.* Overlapping recognition determinants within the ssrA degradation tag allow modulation of proteolysis. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10584–10589 (2001).

14. Aubin-Tam, M.-E., Olivares, A. O., Sauer, R. T., Baker, T. a & Lang, M. J. Single-Molecule Protein Unfolding and Translocation by an ATP-Fueled Proteolytic Machine. *Cell* **145,** 257–67 (2011).

15. Maillard, R. A. *et al.* ClpX(P) Generates Mechanical Force to Unfold and Translocate Its Protein Substrates. *Cell* **145,** 459–69 (2011).

16. Sen, M. *et al.* The ClpXP protease unfolds substrates using a constant rate of pulling but different gears. *Cell* **155,** 636–46 (2013).

17. Ha, T. Single-molecule methods leap ahead. *Nat. Methods* **11,** 1015–1018 (2014).

18. Joo, C., Fareh, M. & Narry Kim, V. Bringing single-molecule spectroscopy to macromolecular protein complexes. *Trends Biochem. Sci.* 1–8 (2012).

19. Schuler, B. Single-molecule FRET of protein structure and dynamics - a primer. *J. Nanobiotechnology* **11,** S2 (2013).

20. Martin, A., Baker, T. a & Sauer, R. T. Rebuilt AAA + motors reveal operating principles for ATP-fuelled machines. *Nature* **437,** 1115–20 (2005).

21. Ortega, J., Singh, S. K., Ishikawa, T., Maurizi, M. R. & Steven, a C. Visualization of substrate binding and translocation by the ATP-dependent protease, ClpXP. *Mol. Cell* **6,** 1515–21 (2000).

22. Gribun, A. *et al.* The ClpP double ring tetradecameric protease exhibits plastic ring-ring interactions, and the N termini of its subunits form flexible loops that are essential for ClpXP and ClpAP complex formation. *J. Biol. Chem.* **280,** 16185–96 (2005).

23. Zhou, Y., Gottesman, S., Hoskins, J. R., Maurizi, M. R. & Wickner, S. The RssB response regulator directly targets sigma(S) for degradation by ClpXP. *Genes Dev.* **15,** 627–37 (2001).

24. 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,** 139–45 (2008).

25. Amor, A. J., Schmitz, K. R., Sello, J. K., Baker, T. A. & Sauer, R. T. Highly dynamic interactions maintain kinetic stability of the ClpXP protease during the ATP-fueled mechanical cycle. *ACS Chem. Biol.* acschembio.6b00083 (2016).

26. McGinness, K. E., Baker, T. a. & Sauer, R. T. Engineering Controllable Protein Degradation. *Mol. Cell* **22,** 701–707 (2006).

27. Stinson, B. M. *et al.* Nucleotide Binding and Conformational Switching in the Hexameric Ring of a AAA+ Machine. *Cell* **153,** 628–39 (2013).

28. Kenniston, J. a, Baker, T. a & Sauer, R. T. Partitioning between unfolding and release of native domains during ClpXP degradation determines substrate selectivity and partial processing. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 1390–5 (2005).

29. 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,** 511–20 (2003).

30. Kirstein, J. *et al.* The antibiotic ADEP reprogrammes ClpP, switching it from a regulated to an uncontrolled protease. *EMBO Mol. Med.* **1,** 37–49 (2009).
31. Bolon, D. N., Grant, R. A., Baker, T. A. & Sauer, R. T. Nucleotide-dependent substrate handoff from the SspB adaptor to the AAA+ ClpXP protease. *Mol. Cell* **16,** 343–350 (2004).
32. Martin, A., Baker, T. a & Sauer, R. T. Diverse pore loops of the AAA+ ClpX machine mediate unassisted and adaptor-dependent recognition of ssrA-tagged substrates. *Mol. Cell* **29,** 441–50 (2008).
33. Davis, J. H., Baker, T. a & Sauer, R. T. Small-molecule control of protein degradation using split adaptors. *ACS Chem. Biol.* **6,** 1205–13 (2011).
34. Farrell, C. M., Baker, T. a & Sauer, R. T. Altered specificity of a AAA+ protease. *Mol. Cell* **25,** 161–6 (2007).
35. Iosefson, O., Olivares, A. O., Baker, T. A. & Sauer, R. T. Dissection of Axial-Pore Loop Function during Unfolding and Translocation by a AAA+ Proteolytic Machine. *Cell Rep.* **12,** 1032–1041 (2015).
36. Hersch, G. L., Baker, T. a & Sauer, R. T. SspB delivery of substrates for ClpXP proteolysis probed by the design of improved degradation tags. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 12136–41 (2004).
37. Sauer, R. T. *et al.* Sculpting the proteome with AAA(+) proteases and disassembly machines. *Cell* **119,** 9–18 (2004).
38. Chandradoss, S. D. *et al.* Surface passivation for single-molecule protein studies. *J. Vis. Exp.* (2014).

4

Chapter 5

Engineering ClpP for Single-Molecule Protein Fingerprinting

5.1 Introduction

Single-molecule protein sequencing using FRET requires a donor fluorophore on ClpXP. Sequencing substrates initially bind to ClpX and are subsequently unfolded and translocated into ClpP. The substrate is cleaved by any of the fourteen active sites concealed inside the ClpP barrel. Due to the dimensions of ClpX and ClpP (**Figure 1a**), placing a donor fluorophore on ClpP would result in energy transfer to acceptor fluorophores on sequencing substrates only after translocation of the substrate into ClpXP. Therefore, signals from acceptor

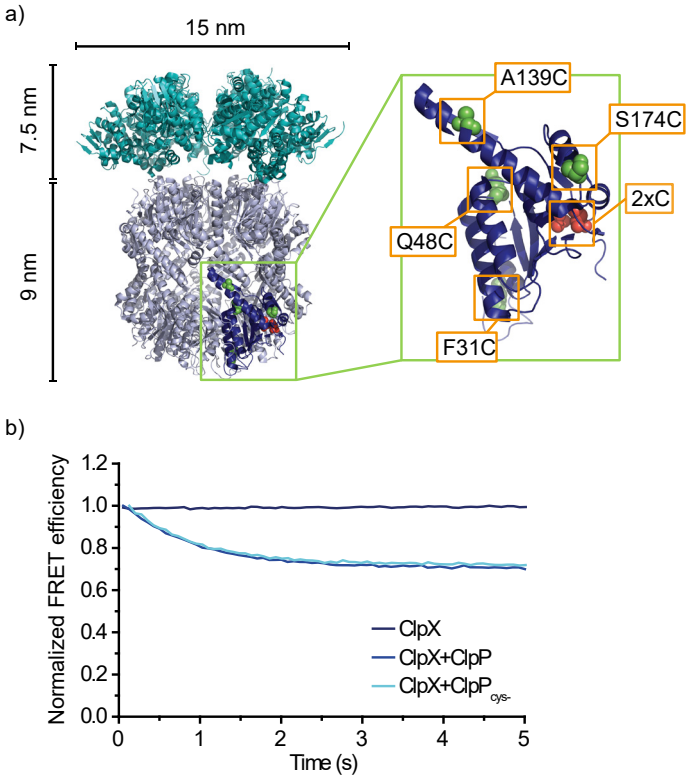


Figure 5.1: (**a**) Crystal structure of ClpXP (obtained by manually combining crystal structures obtaiened from 1YG6 and 3RY2) with highlighted in red the two cysteines present in wildtype ClpP and highlighted in green positions were cysteines were introduced in the three mutants. (**b**) Fluorescence-based activity assay to evaluate the activity of ClpXP. In the absence of ClpP, there is no degradation and the curve is flat. In the presence of ClpP, both wild-type and cys- mutant, a decrease in FRET efficiency is observed, indicating degradation of our model substrate.

fluorophores on substrates in the pretranslocation phase are not detected.

Donor conjugation to ClpP has to occur with high efficiency, high specificity and under mild chemical conditions to protect protein activity. Based on these criteria we selected mono-maleimide conjugation to cysteine residues (1,2). Cysteines have to be introduced to ClpP in specific locations by point mutations. Positioning of the point mutations requires careful consideration to prevent protein misfolding and loss of protease activity of ClpP.

5.2 Results and discussion

Based on existing literature or crystal structure we designed four ClpP mutants with cysteines in predetermined positions: $ClpP_{F31C}$ (3) at either the interface with ClpX or at the far end; $ClpP_{Q48C}$ (4,5) inside the proteolytic barrel; $ClpP_{A139C}$ near the equator of the barrel; and $ClpP_{S174C}$ on the outside of the barrel (**Figure 5.1a,** depicted in green). Each ClpP monomer contains two native cysteines that are not suitable for labeling (3) (**Figure 5.1a**, depicted in red). The two native cysteines were substituted by serine, a chemically similar amino acid unreactive with maleimide groups, prior to introduction of the above mentioned point mutations. Since ClpP is a tetradecameric homo-oligomer, mutations will occur in all 14 monomers. For visualization purposes, however, the mutations are depicted in a single monomer.

Cysteines can form disulphide bonds and are important residues for protein stability (6). Substitution of cysteines by non-disulphide bond forming residues can severely affect protein folding and activity. We designed a fluorescence-based activity assay to compare the activity of wild-type ClpP and $ClpP_{cys-}$. A synthetic peptide with a donor and an acceptor fluorophore was introduced to ClpXP. The fluorophores on the peptide are 16 amino acids apart, resulting in high FRET signal in the folded state and decreasing FRET efficiency after unfolding and degradation. The change in FRET was monitored with a spectrofluorometer. We observed a comparable decrease in FRET efficiency over time for wild-type ClpP

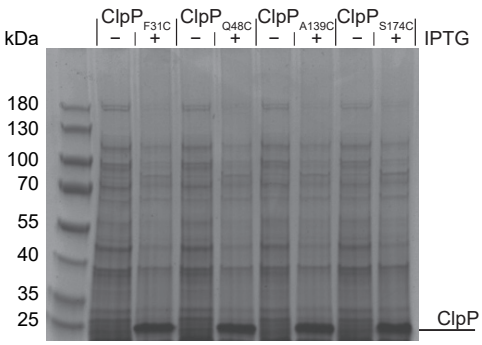


Figure 5.2: Full cell lysate on 4-20% SDS-PAGE gel after coommassie stain. Protein expression profiles were compared before and after induction with IPTG. For all four mutants a thick band appeared at the size of ClpP monomers.

and $ClpP_{cys-}$, both in complex with ClpX. These results led us to conclude the mutations did not interfere with ClpP activity (**Figure 5.1b**).

The four ClpP mutants described above were created based on $ClpP_{cys-}$. We observed overexpression of all four mutants (**Figure 5.2**). After translation, ClpP monomers

tightly self-assemble in a tetradecamer and remain assembled during purification (7). During expression and purification of the ClpP mutants, $ClpP_{S174C}$ could not be isolated successfully and remained mostly in the cell pellet. This is potentially due to improper folding, leading to aggregation and formation of inclusion bodies. As a consequence, $ClpP_{S174C}$ was not included in the rest of our study.

Besides mutations, chemical modifications such as bioconjugation of a fluorophore can interfere with protein activity (1). The fluorescence-based assay described above is incompatible with donor fluorophore conjugation to ClpP, therefore we used an electrophoresis-based assay to compare the degradation efficiency of unlabeled and labeled ClpP mutants to wild-type ClpP. We incubated ClpXP with $titin_{V13P}$-ssrA in the presence of ATP and monitored the amount of intact titin at t = 0 min and t = 30 min. For wild-type ClpP, the unlabeled mutants and most of the labeled mutants we observed time dependent loss of intensity in the band corresponding with the size of $titin_{V13P}$ (**Figure 5.3a**).

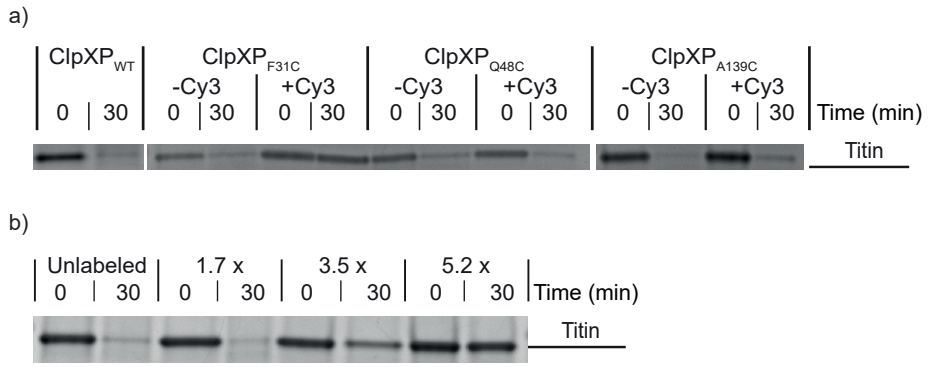


Figure 5.3: (**a**) Degradation assay of titinV13P-ssrA by ClpXP. Degradation capacity of unlabeled and labeled ClpP mutants was compared to wild-type ClpP. (**b**) Degradation assay of titinV13P-ssrA by ClpXP with increasing labeling efficiencies. ClpPF31C showed reduced degradation efficiency with increasing labeling efficiency.

For labeled $ClpP_{F31C}$, however, the loss of intensity was minimal to non-existing. By varying the incubation time of the labeling reaction, the labeling efficiency of $ClpP_{F31C}$ could be controlled. Labeling for 1 h led to 1.7 dyes per $ClpP_{14}$, 2 h lead to 3.5 dyes per $ClpP_{14}$ and 1 h lead to 5.2 dyes per $ClpP_{14}$. Using the electrophoresis-based assay described above we compared the degradation activity of $ClpP_{F31C}$ with different labeling efficiencies (**Figure 5.3b**). Interestingly, $ClpP_{F31C}$ with on average 1.7 dyes per tetradecamer showed near wild-type degradation capacity which decreased to non-existent with increasing labeling efficiency.

5.3 Conclusions

Three out of four ClpP mutants could be successfully purified. All three mutants were active, both in labeled and unlabeled condition. $ClpP_{F31C}$ protein activity showed strong correlation with labeling efficiency. To prevent redundancy throughout this thesis, please see **Chapter 3 and Figure S3.1**, for the performance of the ClpP mutants in our single-molecule assay. In short, we could obtain single-molecule FRET traces for the three mutants described in this chapter. FRET between acceptor labeled substrate and donor labeled $ClpP_{F31C}$ did not result in high energy transfer, eliminating this mutant as a candidate for our fingerprinting scheme. Although $ClpP_{Q48C}$ and $ClpP_{A139C}$ showed very

similar behavior, FRET between the substrate and ClpP resulted in a sharper and better defined high FRET peak for $ClpP_{Q48C}$. $ClpP_{Q48C}$ was selected for the single-molecule experiments throughout this thesis based on the high FRET peak profile in combination with the higher number of labels per ClpP, potentially allowing for longer imaging before photobleaching occurs.

5.4 Materials and methods

5.4.1 $ClpX_6$ purification and biotinylation

To ensure proper immobilization and hexamer formation of $ClpX_6$ at low concentrations $ClpX_6(\Delta N)$, a covalently linked hexamer with a single biotinylation site, was used throughout the experiments. $ClpX_6(\Delta N)$ was overexpressed and purified as described (8). In brief, ClpX protein expression was induced from a BLR strain at OD_{600}~0.6 by adding 1.0 mM IPTG and incubated overnight at 18°C. Simultaneously, 100μM of biotin was added to increase biotinylation efficiency with wild-type BirA. Cells were pelleted and resuspended in lysis buffer (20 mM of HEPES pH 7.6, 400 mM of NaCl, 100 mM of KCl, 10% of glycerol, 10 mM of β-mercaptoethanol, 10 mM of imidazole) in the presence of 1mM PMSF and lysed by French press twice at 20 psi. $ClpX_6$ was purified from the supernatant first with Ni^{2+}-NTA affinity resin, followed by size exclusion chromatography with Prep Sephacryl S-300 16/60 High Resolution column (GE Healthcare).

5.4.2 ClpP mutations, purification and labeling

Point-mutations in ClpP to produce the cysteine free mutant $ClpP_{C91S;C113S}$, and the subsequent ClpP mutants were constructed by overlap extension PCR. Wild-type ClpP and ClpP mutants were overexpressed from BL21pLysS at OD_{600}~0.6 by adding 0.5 mM IPTG and incubated for 3h at 30°C. Cells were pelleted and resuspended in lysis buffer (50 mM of sodium phosphate pH 8.0, 1 M of NaCl, 10% of glycerol, 5 mM of imidazole) in the presence of Set III protease inhibitors (Calbiochem) and lysed by French press twice at 20 psi. ClpP was purified from the supernatant first with Ni^{2+}-NTA affinity resin, followed by size exclusion chromatography with Prep Sephacryl S-300 16/60 High Resolution column (GE Healthcare). ClpP was dialyzed overnight against PBS (pH 7.4) before labeling for 4 h at 4°C with monoreactive maleimide donor dye (Cy3, GE Healthcare), 10x molar dye excess was used in PBS pH 7.4 under nitrogen. Free dye was removed using PD Minitrap G-25 size exclusion columns (GE Healthcare). Labeling efficiencies of 1.7 dyes per tetradecamer for $ClpP_{F31C}$ 5.7 dyes per tetradecamer for $ClpP_{Q48C}$ and 1.1 dyes per tetradecamer for $ClpP_{A139C}$ were measured by spectrophotometry (DeNovix DS-11 FX).

5.4.3 Substrate preparation

Mutations in titin-I27 to produce titin$_{V13P}$-ssrA were constructed by overlap extension PCR. Titin mutants were expressed from BL21-AI at OD_{600} ~0.6 by adding 0.2% arabinose and incubated for 4h at 37°C. Cells were pelleted and resuspended in lysis buffer (50 mM of sodium phosphate pH 8.0, 500 mM of NaCl, 10 mM of imidazole) lysed by sonication. Titin was purified from the supernatant with Ni^{2+}-NTA affinity resin. Titin was dialyzed overnight against PBS (pH 7.4) before labeling for 4 h at 4°C with 10x molar excess of monoreactive maleimide acceptor dye (Cy5, GE Healthcare) in the presence of 4M GdnCl in PBS pH 7.4 under nitrogen. Polypeptides were labeled in the presence of a 10x molar excess of monoreactive NHS-ester functionalized dyes (Cy3, GE Healthcare) and monoreactive maleimide acceptor dye (Cy5, GE Healthcare) overnight at 4°C in PBS under nitrogen. Free dye was removed using PD Minitrap G-25 size exclusion columns (GE Healthcare). Labeling efficiencies as high as 95% were measured by spectrophotometry (DeNovix DS-11 FX) and verified by mass spectrometry. (See **Table 5.1** for the full list substrates.)

Name	Sequence	Length (AA)	MW (kDa)	Supplier
K-16-C-ssrA	KSGERDNFAPH-MALVPVCAANDENY-ALAA	29	3,075	Biomatik
Titin$_{V13P}$-ssrA	MRGSHHHHHHGLVPRG-SLIEVEKPLYGVEPFVG-ETAHFEIELSEPDVH-GQWKLKGQPLAASPD-CEIIEDGKKHILILHNC-QLGMTGEVSFQAANTK-SAANLKVKELRSAAN-DENYALAA	119	13,048	

Table 5.1: Substrates used throughout this chapter

5.4.4 Fluorescence-based ClpXP activity assay

To test ClpXP activity, 10 μM of ClpX and 30 μM of ClpP were preincubated in PD buffer (25 mM HEPES pH 8.0, 5 mM $MgCl_2$, 40 mM KCl, 0.148% NP-40, 10% glycerol) at room temperature in the presence of 10 mM ATP for 2 min. Preincubated ClpXP was diluted 10x in PD buffer. 1 μM of labeled polypeptide (Kcy5-16-Ccy3-ssra) was added to preincubated ClpXP by rapid mixing. Fluorescent signals

from donor (550 nm excitation, 575 nm emission) and acceptor after FRET (550 nm excitation, 665 nm emission) were monitored by a spectrofluorometer (Cary Eclips) for 5-60 min. We normalized the FRET efficiency to correct for interexperimental variation in maximum FRET value.

5.4.5 Electrophoresis-based ClpXP activity assay

The fluorescence based ClpXP activity assay described above is incompatible with fluorophore labeled ClpP. To assess the activity of donor labeled ClpXP, 0.9 μM $ClpX_6$ and 2.9 μM of $ClpP_{14}$ (WT or mutants) in PD buffer were incubated at 30°C in the presence of 10 μM $titin_{V13P}$-ssrA and 5mM ATP. Samples were taken at t = 0 min and t = 30 min and evaluated on 4-20% precast SDS-PAGE gels (Thermo Scientific) using coomassie stain.

5.5 References

1. Stephanopoulos, N. & Francis, M. B. Choosing an effective protein bioconjugation strategy. *Nat. Chem. Biol.* **7,** 876–84 (2011).
2. McKay, C. S. & Finn, M. G. Click chemistry in complex mixtures: bioorthogonal bioconjugation. *Chem. Biol.* **21,** 1075–101 (2014).
3. Reid, B. G., Fenton, W. a, Horwich, a L. & Weber-Ban, E. U. ClpA mediates directional translocation of substrate proteins into the ClpP protease. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 3768–72 (2001).
4. Maglica, Z., Kolygo, K. & Weber-Ban, E. Optimal efficiency of ClpAP and ClpXP chaperone-proteases is achieved by architectural symmetry. *Structure* **17,** 508–16 (2009).
5. Kolygo, K. *et al.* Studying chaperone–proteases using a real-time approach based on FRET. *J. Struct. Biol.* **168,** 267–277 (2009).
6. Sevier, C. S. & Kaiser, C. a. Formation and transfer of disulphide bonds in living cells. *Nat. Rev. Mol. Cell Biol.* **3,** 836–47 (2002).
7. Alexopoulos, J. a, Guarné, A. & Ortega, J. ClpP: A Structurally Dynamic Protease Regulated by AAA+ Proteins. *J. Struct. Biol.* **179,** 202–10 (2012).
8. Martin, A., Baker, T. a & Sauer, R. T. Rebuilt AAA + motors reveal operating principles for ATP-fuelled machines. *Nature* **437,** 1115–20 (2005).

Chapter 6

Tools to Define a Technology Strategy for Single-Molecule Protein Sequencing

6.1 Introduction

This chapter is based on a report written for the TU Delft course Turning Technology into Business. The main objective of this course was to evaluate the cutting edge advantage of single-molecule protein sequencing technology over current technologies and identify the business opportunities for our patent (1). Detailed analysis of the technology, using the tools taught during the course, has been carried out and a small-scale market study was conducted based on reports available online. This chapter offers insight in strategic decision making for inventions originating from science.

6.2 Valorization

Education, research and valorization form the three core tasks of Dutch universities (2). The transfer of knowledge is what these core tasks have in common: transfer of knowledge to students through education, transfer of knowledge to other scientists through publication and transfer of knowledge to society through valorization (3). Valorization is often confused with gaining economic benefit and viewed as a distraction from practicing science. Protection of intellectual property by universities and other non-profit organizations can create additional income from licensing deals and may be helpful in securing grants. Often, protection of intellectual property is even necessary for society to benefit from scientific findings. It rarely happens that universities commercializean invention in-house. More commonly, universities partner with industry, which might not be willing to take the risk of commercializing an invention that is not properly protected.

Monetary value is not the only value that can be created through valorization. Well-known ways of valorization are patenting, starting spin-off companies and licensing, less well-known means of valorization are consultancy work, courses, post-academic teaching, exhibitions, demonstrations, media appearances and

the availability of websites, books, and software (4). Which valorization strategy to choose is strongly project dependent.

6.3 Patenting

One of the most secure ways to protect intellectual property is through patenting. The patentability of a technical invention is determined by three criteria: novelty, obviousness and industrial application. An invention should be novel, meaning it has not been disclosed in any type of publication, written or oral, prior to the filing date of a patent. The obviousness of an invention is more susceptible to interpretation. An invention is considered obvious if a person skilled in the art, in other words an expert in the field, could arrive to the same finding based on existing sources. The final criterion states that an invention should have a clear application. As a result, patenting fundamental research can be challenging, because the industrial application can be indistinct. Not all findings that meet patenting criteria will be protected by universities. Ownership of the patent should provide monetary, societal or strategic benefits. Therefore, it is important to perform a market analysis to identify potential customers and estimate the market demand.

6.4 The market need

The first step is to determine the market need. What problems are potential customers facing and how can the invention fulfull these needs.

Proteins belong to the most important molecules in life. They function as messengers, transporters and catalysts, and provide cells and tissues with structure. The expression profile of proteins is rich in information, which can be used, for example, in diagnosing diseases. Recent advances in mass spectrometry, the most common protein sequencing technique, have led to a draft of approximately 90% of the human proteome, the remaining 10%, however,

cannot yet be detected (5,6). This commonly used technique has fundamental limitations in sensitivity and sample size, making it ineffective for clinical use.

Throughout this thesis, we describe our efforts to develop a novel technology to sequence proteins on a single-molecule level. Current protein analysis technologies have several limitations. The number of different proteins that can be analyzed in parallel is limited, protein levels are measured indirectly by analyzing mRNA expression or there is a need for very large quantities of sample. The patented technology described in this thesis can directly analyze all protein levels, at the same time, in a sample as small as a single cell.

The single-molecule protein sequencing technique is currently still in the embryonic development phase and commercialization is expected within 10 years. The patent broadly covers the unique technique of sequencing proteins and offers good protection. Furthermore, most of the technological know-how and experience in assembly is tacit knowledge and resides within the scientific team that developed the technique, making it hard to reproduce by another scientific team.

6.5 Existing techniques

An overview of existing techniques gives further insight into the market need and provides insight in both potential competitors as well as customers.

6.5.1 Immunoassays, protein characterization and fusion proteins

Commonly used methods vary from directly detecting a protein type with antibodies (ELISA, western blots, etc.) or identifying proteins by measuring characteristics as size or charge (gel electrophoresis, HPLC, FPLC, etc.) (7). These methods are generally straightforward and inexpensive, but are incapable of detecting a large number of different protein types in parallel.

6.5.2 Edman degradation

Edman degradation is a method where the sequence of a protein is deter-

mined with very high accuracy (8). There are commercial companies such as BioSynthesis, Abingdon Health and Cambridge Peptides that use Edman degradation and can provide sequencing results in 5-10 days. The method they use, however, is limited in the number of amino acids that can be read (maximum of 30 amino acids) and it can only sequence very pure samples.

6.5.3 mRNA sequencing

Other techniques often indirectly determine protein content by assessing mRNA expression (next-generation sequencing) (9). Benefits of these techniques are the number of mRNAs they can assess at the same time. The major downside of these techniques, however, is their inability to report on protein levels directly. Although mRNA levels and protein levels tend to correlate, this is not always the case.

6.5.4 Mass spectrometry

The most widely used method to analyze all protein content of a sample for research purposes, and therefore the main competitor for our technology, is mass spectrometry (10). Although analysis is very complex, mass spectrometry is a suitable technology to identify the most abundant part of the protein population. There are companies and institutes offering services up to

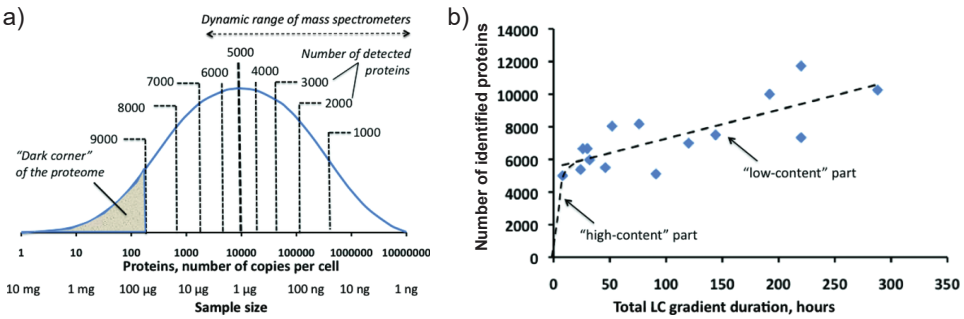


Figure 6.1: (**a**) Conventional proteomics analysis detects highly abundant proteins and stretches for about four orders of magnitude. Deeper proteome analysis requires much larger sample size. The "dark corner" represents part of the proteome that is most challenging for detection. (**b**) Analysis of the first approximately 5000 proteins is very fast, while deeper analysis yields on average <20 additional proteins per hour of LC gradient. (Figures adapted from (11)).

approximately a thousand proteins in a single sample (Alphalyse, GCB). However, the total number of over 20 thousand protein species cannot be reached. If one would try to identify the "dark corner" of the proteome, the analysis time will increase enormously (**Figure 6.1 a**), the amount of sample needed might increase more than a thousand times and the least abundant proteins have up to today not been detected (**Figure 6.1 b**) (11). An additional downside of mass spectrometry is the lack of quantitative results. Proteins have to be fragmented and ionized and the efficiency of both steps is sequence-dependent. Therefore, internal standards need to be used in every step to ensure quantitatively accurate outcomes.

6.5.5 Competitive position

To this day, a technique to analyze every single protein in a sample does not exist. The techniques described above can identify proteins, but none of them is, or ever might be, capable of qualitatively and quantitatively determining the full protein expression profile. Increasing evidence shows that in tissues that were previousbefore presumed homogeneous, there is in fact a very large cell-to-cell variation in protein expression. Analyzing these cell to cell variations will give vast insights in cellular processes involved in diseases such as cancer. Because our technology reads every single protein in a sample, we can sequence the content of a sample as small as a single cell. In comparison, the most sensitive techniques mentioned above need 5-10 thousand cells to identify only one type of protein. So even if the number of proteins of interest is limited, our technology can provide a clear advantage. The technology described in this thesis has a potential performance beyond the fundamental limitations of competing techniques, making it a potentially disruptive technology.

6.6 Market analysis

To estimate the revenues that can be expected, we performed a small-scale market analysis. The main competing technique will be mass spectrometry.

Strategic Direction International estimated the 2011 mass spectrometry sales at $3.9 billion and expected the market to grow to $4.8 billion by 2014 (12). Because the price per unit is estimated to be around $500,000, this translates into approximately 10,000 devices per year. According to the study, the top markets for mass spectrometry are the United States and Canada (38.1%), Europe (31.1%), and Japan (13.3%), followed by China and the Pacific Rim (11.5%). Academia accounts for 12.6% of the market. The highest sales activities were found in pharmaceuticals and biotech (20.4%) followed by government (18.5%). Hospitals and clinics (4.9%) were at the bottom of the list. Although academia only accounts for 12.6%, the total percentage used for medical research applications might be close to 50%. The price of our custom build device is currently $150,000, but we expect to bring the price down to $50,000 by scaling up production and reducing the complexity of the setup. Because our devices are considerably cheaper, we should be able to capture 20% of academia, meaning we would sell roughly 250 devices per year resulting in $12.5 million in revenues.

6.7 Technology assessment

Although technology is all around us, constructing a definition broad enough to cover all forms of technology is challenging. Chris Floyd (13) defined technology as know-how, distinguishing it from science as the pursuit of knowledge, and products or equipment which are physical and not intellectual assets. The technology behind an invention can be determined by completing a simple statement: “We know how to…”

A technology assessment was performed to formulate a solid technology strategy. To get an overview of the technologies underlying the invention described in this thesis, the invention was unbundled in a technology tree (**Figure 6.2**). The individual technologies were identified by completing the statement above. Furthermore, by answering “We should know how to …”, the technology tree

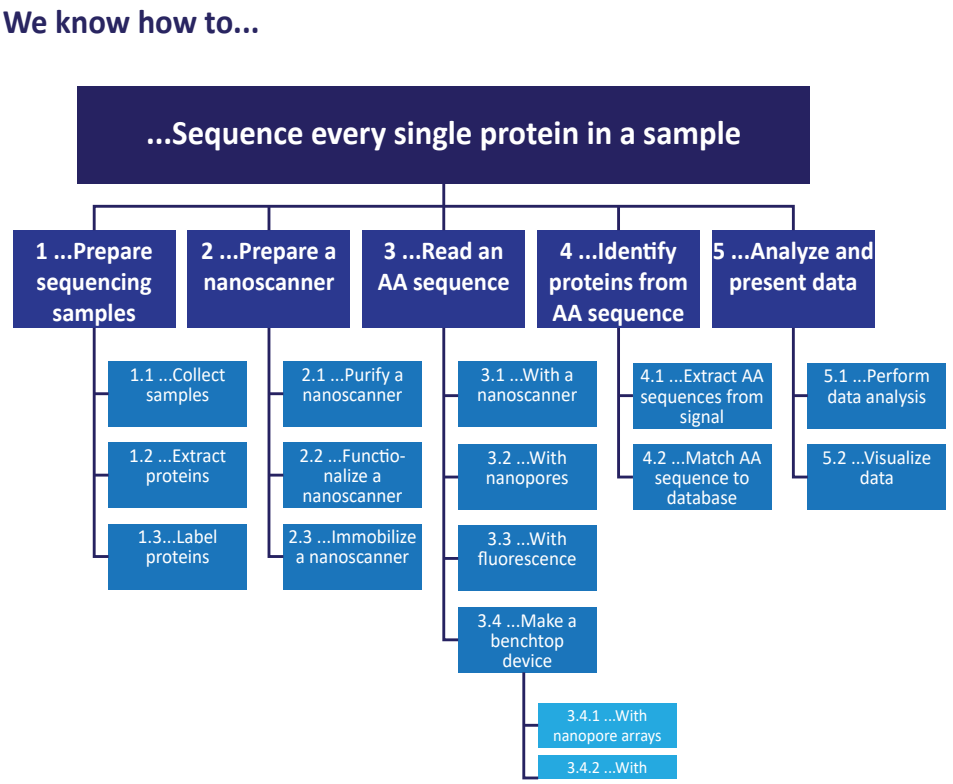


Figure 6.2: The technology tree unbundles the underlying technologies necessary for Single-Molecule Protein Sequencing.

	Embryonic	Growth	Mature	Aging
Technological certainty	Poor	Fair	High	Very high
Predictability of growth potential	Low	Fair	High	Very high
Predictability development efforts	Poor	Moderate	High	Very high
Number of competitors	Limited	Maximal	Reduced	Low
Accessibility / availability	Low	Moderate	High	Very high
Commercial advantage	High	Moderate	Fair	High

Table 6.1: Criteria for technology maturity (14).

resulting from this technology unbundling exercise will give an overview of both present and future technological developments needed to deliver a marketable product.

Technologies mature over time (14) and technologies can be classified as embryonic, growth, mature or aging based on the criteria in **Table 6.1**. Technologies can also be classified based on their level of strategic impact (15). Base technologies are essential for business, however, they offer low competitive advantage because they are widely available, these technologies have low strategic impact. Key technologies are well-integrated technologies that offer high strategic impact. Pacing technologies are still under development, but have the potential to replace key or base technologies and it is therefore likely

	We know how to...	Maturity	Strategic impact	Competitive position
1	Prepare sequencing samples			
1.1	Collect samples	Aging	Base	Weak
1.2	Extract proteins	Mature	Key	Tenable
1.3	Label proteins	Growth	Pacing	Favorable
2	Prepare a nanoscanner			
2.1	Purify a nanoscanner	Growth	Pacing	Strong
2.2	Functionalize a nanoscanner	Embryonic	Emerging	Dominant
2.3	Immobilize a nanoscanner	Growth	Pacing	Strong
3	Read an AA sequence			
3.1	With a nanoscanner	Embryonic	Emerging	Dominant
3.2	With nanopores	Embryonic	Pacing	Dominant
3.3	With fluorescence	Growth	Pacing	Dominant
3.4	Make a benchtop device	Growth	Pacing	Favorable
3.4.1	With nanopore arrays	Embryonic	Emerging	Favorable
3.4.2	With fluorescence arrays	Growth	Pacing	Favorable
4	Identify proteins from AA sequence			
4.1	Extract AA sequence from signal	Embryonic	Emerging	Dominant
4.2	Match AA sequence to database	Growth	Pacing	Favorable
5	Analyze and present data			
5.1	Perform data analysis	Mature	Key	Tenable
5.2	Visualize data	Mature	Key	Tenable

Table 6.2: Assessment of maturity, strategic impact and competitive position of the technologies in the technology tree.

Strategic impact	Technology maturity: Embryonic	Growth	Mature	Aging
Base				1.1 Collect sequencing samples
Key			1.1 Extract proteins 5.1 Perform data analysis 5.2 Visualize data	
Pacing	3.2 Scan with nanopores	1.3 Label proteins 2.1 Purify a nanoscanner 2.3 Immobilize a nanoscanner 3.3 Scan with fluorescence 3.4 Make a benchtop device 3.4.2 Benchtop device with fluorescence 4.2 Match AA sequence with database		
Emerging	2.2 Functionalize a nanoscanner 3.1 Scan with a nanoscanner 3.4.1 Benchtop device with nanopores 4.1 Extract AA sequence from signal			

Buy | Make | Co-develop with suppliers | Develop

Table 6.3: The strategic impact matrix gives a clear overview of technologies to develop in-house and technologies to outsource.

they have high impact. Emerging technologies are in early development and promising, the strategic impact, however, is unknown. A third dimension to assess technologies is by competitive position. The competitive position of technologies can be either: weak, tenable, favorable, strong or dominant (15).

For each technology identified in the technology tree from **Figure 6.2** the maturity, strategic impact and competitive position were analyzed (**Table 6.2**). Plotting the strategic impact versus the maturity reveals the areas to focus on and the areas that would benefit from outsourcing (**Table 6.3**). The technology

assessment described above can give valuable information to researchers on which technologies are beneficial to develop further and which technologies should be abandoned. Simultaneously, the technology assessment gives insight in strategic advantages gained by owning a certain technology, supporting universities in building strong patent portfolios. Many universities struggle in finding proper licensing partners for their inventions and it is estimated that 95% of university patents remain unlicensed, leaving an enormous financial burden (16).

As the technology of the patent is still under development and single-molecule protein sequencing has not been achieved before, most technologies from the technology tree are in the embryonic or growth phase. At the same time, the strategic impact of these technologies is pacing or emerging. However, this does not apply to all technologies in the technology tree. There are some technologies, such as the collection of biological samples and the extraction of proteins that have been around for decades. It might turn out that we have to customize some of these techniques to fit our device, but these are not the technologies to spend many resources on from the start. Other technologies from the technology tree are currently implemented or explored by other companies. This is the case for making a bench top device, for either nanopores (17) or fluorescence (18). In our laboratories, we use single nanopores. To get sufficiently fast read out, arrays of nanopores are needed. Oxford Nanopore Technologies, a spinoff company from the University of Oxford, produces nanopore-based DNA sequencing devices. Currently, they are testing the MinION, a device containing an array of 500 nanopores and the PromethION, which contains 75,000 nanopores in one array ,and they have plans to go to even larger numbers. In this area they are the clear leaders and our position will be favorable at best. The same holds for arrays for fluorescence measurements. We can image ~1,000 molecules in parallel. The clear leader in this market, Pacific Biosciences, sells a device for DNA sequencing that can monitor 150,000 individual wells in parallel.

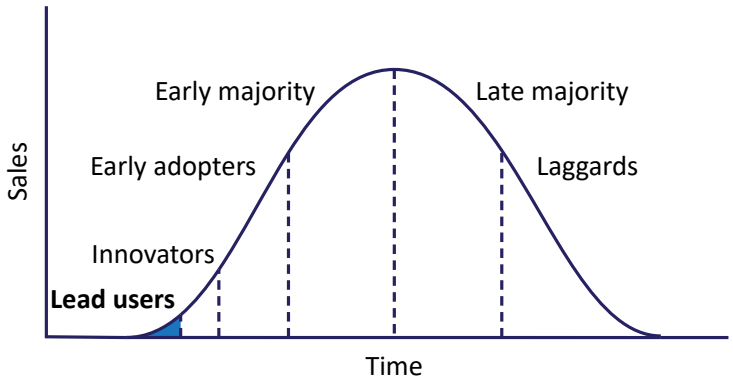


Figure 6.3: A graphic representation of Rogers theory on diffusion of innovation accomanied by the lead user concept as described by Von Hippel (19-21).

6.8 Lead users

In 1988, Eric von Hippel introduced the concept of lead users (19,20). The type of customer buying your product differs with the amount of time it has been on the market. Innovators and early adopters are more willing to try new products, early and late majority customers wait for a product to prove its value and laggards are in general the last to buy a product. As an extension on Rogers' theory on diffusion of innovation (21), Von Hippel proposed a group of lead users that comes even before the innovators (**Figure 6.3**). These costumers have most to gain from a new technology, therefore they are willing to contribute valuable insights and are more willing to invest in a product, because they also gain a larger benefit. By identifying the lead users of a new product, their needs and opinions can be considered early-on in the development process to prevent creating a product that doesn't fit the market need.

The lead users for a single-molecule protein sequencer will most likely be research groups in the field of proteomics. Efforts in this field are made to map the full human proteome, including the "dark corners" mentioned above. In addition, proteomics are used to find new biomarkers for diseases and to detect the presence of bacteria. Since users are confined in their research by the

limitations of present day techniques, they will benefit significantly from our technology. We could tailor the first prototypes to their needs. In return, we can benefit from implementing our device in a research environment; researchers are used to working with new techniques and therefore, are more willing to test prototypes. In addition, the field of proteomics is extensive and growing. Once we convince the group of lead users of the benefits of our technology, we can move on to researchers in life sciences that use other competing techniques.

6.9 Commercialization strategies

In most markets, commercialization of a product can be more challenging than creating the actual invention. From the academic point of view, creating a spin-off company can be too time consuming, costly and risky. Moreover, a spin-off company is not the only, or even the best, route to commercial success. To determine the proper commercialization strategy that fits your invention, Gans and Stern proposed the "market for ideas" as a counter part to the "market for products" (22). They proposed a decision matrix based on the excludability and the dependence of the spin-off on technologies owned by incumbents (established companies). Based on the matrix, they define four commercialization strategies: the attacker's advantage, reputation-based ideas trading, Greenfield competition and ideas factories (**Table 6.4**).

		Incumbent has assets essential for the value proposition	
		No	Yes
Innovation can be effectively protected against imitation	No	The Attacker's Advantage	Reputation-Based Ideas Trading
	Yes	Greenfield Competition	Ideas Factory

Table 6.4: Gans and Stern's strategy matrix (22).

The attacker's advantage describes a market where innovations are poorly or not protected and incumbents do not own assets necessary for commercialization by the spin-off. In these markets, start-ups and established industry face a level playing field, where start-ups might struggle with limited funds, while established companies might struggle finding the right knowledge. Start-ups have an opportunity here to become market leader, by developing competence destroying technologies. There are not many opportunities to collaborate with established industries, because sharing your invention can easily lead to imitation by your competitor. The best strategy will be working in complete secrecy.

In a market where innovations are difficult to protect and start-up innovators depend on assets of established companies, start-ups have a weak negotiation position. They need established industry and are at risk when sharing their knowledge, therefore it is called reputation-based ideas trading. A start-up should only consider disclosing an invention to a competitor with a reliable reputation.

Greenfield competition gives the innovator the most power to determine their commercialization strategy. The innovator can protect the inventions, for example via patenting, and is not dependent on assets of incumbents providing a start-up innovator with the upper hand in negotiations. Innovators have a choice to compete with incumbents with the advantage of keeping maximum control on development strategy, or collaborate to facilitate, for example, market entrance.

When the level of protection of the invention is high, but incumbents assets are needed to commercialize the product, Gans and Stern talk about idea's factories. In this environment, research institutes and research driven companies keep their focus on research and partner-up for commercialization. Assets needed to develop could be complementary technologies, monetary assets or even market access. For innovations in the field of biotech, like ours, this commercialization

strategy will be the most likely choice. A disadvantage of this strategy is the risk of losing control over the development of the technology. The amount of control largely depends on the negotiation skills of the innovator. Knowing the value of your invention is key. Ideally, a start-up innovator has multiple potential partners and the technology is auctioned off to the highest bidder.

6.10 Conclusions

The patented technology "Single-Molecule Protein Sequencing" not only offers a good business opportunity, but also carries a social aspect by potentially improving the health of millions of people. The major constraint for immediate conversion into business would be the embryonic stage of the technology. The technology needs further development to reach a stage where it can be commercialized. The technologies most worthwhile to further develop were identified here. The combination of patent protection of the protein sequencing technique and the tacit knowledge imbedded in the developer team offer a competitive advantage. By seeking contact with potential users early-on in the development process, key developments and requirements can be identified to make this technology a success. Since our technology is well-protected but resources from incumbents are needed to successfully commercialize the technology, the technology described here will benefit most from collaborating with existing industries over full in-house development.

6.11 Reference

1. Joo, C., Dekker, C., Ginkel, H. G. T. M. van & Meyer, A. S. Single molecule protein sequencing, WO patent 2014014347. (2014).
2. Rathenau Instituut. Valorisatie als feitelijke kerntaak van universiteiten. 1–2 (2012).
3. KNAW. Benutting van octrooien op resultaten van wetenschappelijk onderzoek. (2014).
4. De Jong, S. Engaging Scientists; Organising Valorisation in the Netherlands. (2015).

5. Kim, M.-S. et al. A draft map of the human proteome. Nature 509, 575–81 (2014).
6. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. Nature 509, 582–7 (2014).
7. Talapatra, A., Rouse, R. & Hardiman, G. Protein microarrays: challenges and promises. Pharmacogenomics 3, 527–36 (2002).
8. Edman, P. Method for determination of the amino acid sequence in peptides. Acta Chem. Scand. 4, 283–293 (1950).
9. Mardis, E. R. Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet. 9, 387–402 (2008).
10. Mann, M., Hendrickson, R. C. & Pandey, a. Analysis of proteins and proteomes by mass spectrometry. Annu. Rev. Biochem. 70, 437–73 (2001).
11. Zubarev, R. a. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 13, 723–6 (2013).
12. http://www.alphalyse.com/protein_characterization.html
13. Floyd, C. Managing Technology for Corporate Success. (Gower Publishing, 1997).
14. Hax, A. C. & Majluf, N. S. The life-cycle approach to strategic planning Download. Work. Pap. MIT (1983).
15. Adler, P. S. & Shenhar, A. Assessing the company's technological base. (1989).
16. Ledford, H. Universities struggle to make patents pay. Nature 501, 471–2 (2013).
17. Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz. Nat. Biotechnol. 30, 295–6 (2012).
18. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–8 (2009).
19. von Hippel, E. The sources of innovation. (Oxford University Press, 1988).
20. von Hippel, E. Democratizing Innovation. (MIT Press, 2005).
21. Rogers, E. M. Diffusion of innovations. (New York: Free Press of Glencoe, 1962).
22. Gans, J. S. & Stern, S. The product market and the market for 'ideas': commercialization strategies for technology entrepreneurs. Res. Policy 32, 333–350 (2003).

Summary

Proteins belong to the most important molecules in living organisms. They function as messengers, transporters and catalysts, and provide cells and tissues with structure. The expression profile of proteins is rich in information, which can be used, for example, in diagnosing diseases. Therefore proteomics, the large scale study of proteins, can give us valuable information on molecular pathways and state of health. As a result, proteomics has the potential to transform personalized medicine.

Recent advances in mass spectrometry have led to a draft of the human proteome. With current mass spectrometry based techniques, these types of large scale studies remain an enormous effort. Therefore, there is a great need for breakthrough technologies to push proteomics from fundamental research into the clinic.

Genomics has benefitted from fast and inexpensive emerging single-molecule techniques. We envision similar effects for single-molecule protein sequencing. In this thesis we present our technology that will allow us to analyze protein expression profiles of samples as small as a single cell with large dynamic range.

Back in 2011, when this project was initiated, there was hardly any literature available on this topic. However, the past years more research groups openly shifted their focus to single-molecule protein sequencing. In **Chapter 1**, we give an overview of recent efforts to establish single-molecule protein sequencing.

The foremost reason for the absence of highly sensitive and high-throughput protein sequencing techniques is the complexity of primary protein structures compared to DNA/RNA molecules. Where DNA and RNA consist of four unique building blocks, proteins are built from 20 distinctive amino acids. Independent of the read out method of choice, this requires the detection of 20

distinguishable signals. A non-trivial challenge. Fortunately, a limited number of proteins occur compared to the theoretical number that could be created using 20 unique building blocks. While the exact number of protein coding genes in the human genome is still under debate, the number is believed to be roughly 20,000, resulting in a number of protein products that is finite. This, together with protein databases such as UniProt, allows for an alternative way of identifying protein sequences. Rather than detecting every single element, as is essential for DNA sequencing, we choose to focus on detecting the sequence of a subset of elements.

In **Chapter 2**, we computationally demonstrate that identifying the sequence of only two out of 20 building blocks is sufficient to identify proteins from an existing database. We named this approach "single-molecule protein fingerprinting". To assess the predictive power of fingerprinting, we developed a search algorithm. In our analysis we considered a specific set of potential experimental errors. In an ideal situation with no experimental error, a high detection precision was reached. The detection precision decreased considerably after introducing errors in the simulated data. If, however, the distance between the detected building blocks, cysteines and lysines, could be taken into account, the detection precision vastly increased. Using cysteine and lysine fingerprints including the distance information, we were able to detect, with high detection precision, non-human HRSV (human respiratory syncytial virus) and TB (tuberculosis) proteins against the human protein database.

In addition to the challenging nature of the primary structure of proteins, potential protein sequencing platforms cannot benefit from natural copying machinery, like current DNA sequencing techniques. Therefore, protein sequencing techniques will only be commercially successful if they can detect very low protein numbers. In **Chapter 3**, the highlight of this thesis, we present our approach to scan proteins at the single-molecule level utilizing ClpXP, a naturally occurring nano-motor. This AAA+ protease can unfold and

degrade proteins with specific recognition tags that are used *in vivo* for protein degradation and remodeling. By attaching a donor molecule to ClpXP and acceptor molecules to two out of 20 types of amino acids of the proteins substrate, we can potentially read the sequence of this subset of amino acids by monitoring changes in FRET efficiency over time, using total internal reflection microscopy. The third chapter demonstrates the first proof-of-concept of a single-molecule fluorescence peptide analysis. We showed our ClpXP platform exhibits high processivity, uni-directional processing with a constant speed, and two orders of magnitude of dynamic range in sensitivity, making a promising approach to sequence full length protein substrates.

ClpXP is a very well-studied AAA+ protease in bacteria that is known to degrade incomplete protein products after ribosome stalling. During this process partial protein products are functionalized with a ssrA tag, marking them for degradation. Single-molecule optical trapping studies have provided great insight in the translocation behavior of ClpXP. However, substrate recognition and initial engagement into the central pore before translocation occurs has received little attention. In **Chapter 4** we utilize the single-molecule FRET assay described in **Chapter 3** to characterize the effects of ATP cofactors and a recognition tag on ClpX substrate recognition and initial translocation. We observed an ATP cofactor dependent effect on ClpXP complex formation. Initial substrate binding frequencies and subsequent translocation efficiencies were influence by both ATP cofactors and recognition tag sequence. In addition we observed a reduction of translocation speed under suboptimal energy conditions, a highly valuable observation for our single-molecule protein sequencing approach.

Obtaining modified and functional $ClpX_6$ and $ClpP_{14}$ proteins has turned out a non-trivial assignment. In **Chapter 5** we describe our efforts to create $ClpP_{14}$ mutants to allow for donor conjugation. We were able to obtain multiple active mutants, even after fluorophore conjugation, although a negative correlation between labeling efficiency and protein activity was observed. All

purified mutants exhibited FRET between the conjugated donor fluorophore and substrates labeled with acceptors. We selected the mutant that performed best, both in terms of high FRET peak profile and the high number of labels per $ClpP_{14}$, potentially allowing for longer imaging before photobleaching occurs, for the single-molecule experiments throughout this thesis.

Valorization is one of the three core tasks of Dutch universities. Not every research topic presents an obvious opportunity for valorization. The project described in this thesis, however, has great commercial potential and, once successful, will have considerable societal impact. **Chapter 6** provides insight in strategic decision making for inventions originating from science with a market potential. We provide an overview of aspects to consider before applying for a patent and form a commercialization strategy based on a technology assessment, market analysis and potential users.

This thesis describes the efforts that led to a first proof-of-concept of our FRET based single-molecule protein fingerprinting approach. Although there might be a long way to go for this technique to reach the clinic, feasibility of the technology was demonstrated. Hopefully, the work presented here will provide the groundwork for further development of the single-molecule protein fingerprinting platform.

Samenvatting

Eiwitten behoren tot de belangrijkste moleculen van levende organismen. Ze zijn belangrijk in communicatie en transport en eiwitten werken als katalysators. Tevens voorzien ze cellen en weefsels van structuur. Het expressieprofiel van eiwitten is rijk aan informatie die bijvoorbeeld gebruikt kan worden voor diagnose van ziekten. Proteomica, het bestuderen van eiwitten op grote schaal, kan derhalve waardevolle informatie geven over moleculaire pathways en gezondheid. Als gevolg hiervan heeft proteomica de potentie om personalized medicine te transformeren.

Recente ontwikkelingen in massaspectrometrie hebben geleid tot een draft van het menselijke proteoom. Met de huidige, op massaspectrometrie gebaseerde, technieken blijft dit echter een enorme inspanning. Om deze reden is er een grote behoefte aan baanbrekende technologieën om proteomica van fundamenteel onderzoek naar de kliniek te krijgen.

Genomica heeft kunnen profiteren van snelle en goedkope nieuwe technieken op enkel-molecuul niveau. Wij voorzien eenzelfde effect voor eiwit sequencing op enkel-molecuul niveau. In dit proefschrift presenteren wij onze technologie die ons in staat zal stellen om het eiwit expressie profiel te bepalen van een monster zo klein als een enkele cel met een groot dynamisch meetbereik.

Toen wij dit project startten in 2011 was er bijna geen literatuur beschikbaar over dit onderwerp. De afgelopen jaren zijn er echter steeds meer onderzoeksgroepen dit zich openlijk richten op eiwit sequencing op enkel-molecuul niveau. In **Hoofdstuk 1** geven wij een overzicht van de recente inspanningen om eiwit sequencing op enkel-molecuul niveau te verwezenlijken.

De belangrijkste reden voor het ontbreken van een zeer sensitieve en high-throughput eiwit sequencing technologie is de complexiteit van de primaire

eiwit structuur vergeleken met DNA/RNA moleculen. Waar DNA en RNA uit vier unieke bouwstenen bestaan, worden eiwitten gevormd uit 20 verschillende aminozuren. Onafhankelijk van de uitleesmethode vereist dit de detective van 20 te onderscheiden signalen, geen triviale uitdaging. Gelukkigerwijs komt slechts een beperkt aantal eiwitten voor, vergeleken met het theoretische aantal dat kan worden gevormd met 20 unieke bouwstenen. Hoewel het exacte aantal eiwit coderende genen in het menselijk genoom ter discussie staat, wordt het aantal verondersteld op ongeveer 20.000, resulterend in een aantal eiwitproducten dat eindig is. Dit gegeven, samen met bestaande eiwit databases zoals UniProt, zorgt ervoor dat een alternatieve manier ontstaat om eiwit sequenties te bepalen. In plaats van ieder afzonderlijk element te detecteren, zoals noodzakelijk voor DNA sequencing, hebben wij ervoor gekozen ons te richten op het detecteren van een subgroep van elementen.

In **Hoofdstuk 2** laten we via een computeranalyse zien dat het identificeren van slechts twee van de 20 bouwstenen voldoende is om eiwitten te identificeren met behulp van een bestaande database. We hebben deze aanpak "single-molecule protein fingerprinting" genoemd. Om het voorspellend vermogen van "fingerprinting" te bepalen, hebben wij een zoekalgoritme ontwikkeld. In onze analyse hebben we een specifieke set meetfouten meegenomen. In een ideale situatie, zonder enige meetfout, werd een hoog voorspellend vermogen behaald. Het voorspellend vermogen daalde aanzienlijk na de introductie van meetfouten in de gesimuleerde data. Wanneer echter de afstand tussen de gedetecteerde bouwstenen, cysteïnes en lysines, meegenomen kon worden in de analyse werd het voorspellend vermogen enorm verbeterd. Met de cysteïne en lysine fingerprint, inclusief de informatie over de afstand, waren wij instaat om met grote precisie non-humaan RSV (respiratoir syncytieel virus) en TB (Tuberculose) te detecteren tegen de humane eiwit databank.

Naast het uitdagende karakter van de primaire structuur van eiwitten kan een potentieel eiwit sequencing platform ook niet profiteren van een

natuurlijk kopieermechanisme, zoals huidige DNA sequencing technieken doen. Eiwit sequencing technieken zullen dan ook alleen commercieel succesvol zijn wanneer kleine hoeveelheden eiwitten kunnen worden gedetecteerd. In **Hoofdstuk 3**, het hoogtepunt van dit proefschrift, presenteren wij onze aanpak om eiwitten te scannen op enkel-molecuul niveau door gebruik te maken van ClpXP, een natuurlijk voorkomende nanomotor. Deze AAA+ protease kan eiwitten die voorzien zijn van een tag, dat *in vivo* gebruikt wordt voor eiwit degradatie en herstructurering, ontvouwen en afbreken. Door een donormolecuul aan ClpXP en acceptormoleculen aan twee van de 20 typen aminozuren te bevestigen, kunnen we potentieel de sequentie van deze subgroep uitlezen door met de tijd de veranderingen in FRET efficiëntie te monitoren met totale interne reflectie microscopie. In dit hoofdstuk hebben we een eerste proof-of-concept voor de analyse van fluorescente peptiden op enkel-molecuul niveau gedemonstreerd. We hebben laten zien dat ons ClpXP platform functioneert met hoge processiviteit, in één richting, met constante snelheid en twee orden van grootte in dynamisch meetbereik, wat het een veelbelovende aanpak maakt om de sequentie te bepalen over de volle lengte van eiwitsubstraten.

ClpXP is een zeer goed bestudeerde AAA+ protease in bacteriën dat bekend staat om degradatie van incomplete eiwitproducten na het vastlopen van ribosomen. Tijdens dit proces worden gedeeltelijk gevormde eiwitten gemarkeerd voor degradatie door ze van een ssrA tag te voorzien. Het gebruik van een optische val op enkel-molecuul niveau heeft belangrijk inzicht geleverd in het translocatie gedrag van ClpXP. Substraat herkenning en initiële binding met de centrale opening heeft echter nog weinig aandacht gekregen. In **Hoofdstuk 4** gebruiken wij de FRET analyse op enkel-molecuul niveau beschreven in **Hoofdstuk 3** om de effecten van ATP cofactoren en herkenningstags op de herkenning en initiële translocatie van een substraat door ClpX te bestuderen. Wij observeerden een ATP-cofactor afhankelijk effect op ClpXP-complex formatie. Tevens werden de frequentie van initiële binding van het substraat en daaropvolgende translocatie efficiëntie beïnvloed door zowel

ATP-cofactoren als de sequentie van de herkenningstag. Daarnaast observeerden we een afname van de translocatiesnelheid onder suboptimale energie-condities, een zeer waardevolle observatie voor onze aanpak van eiwitsequencing op enkel molecuul niveau.

Het verkrijgen van gemodificeerde en functionele $ClpX_6$ en $ClpP_{14}$ eiwitten bleek geen triviale opdracht. In **Hoofdstuk 5** beschrijven we onze inspanningen om $ClpP_{14}$ mutanten te creëren die geschikt zijn voor donor conjugatie. We waren in staat om meerdere actieve mutanten te verkrijgen, zelfs na de conjugatie van een fluorofoor, hoewel een negatieve correlatie werd geobserveerd tussen de efficiëntie van de conjugatie en de activiteit van het eiwit. Alle opgezuiverde mutanten vertoonden FRET tussen het geconjugeerde donor-fluorofoor en substraten met een acceptor label. We selecteerden de mutant die het best presteerde op de hoogte van de FRET piek en het aantal labels per $ClpP_{14}$, wat mogelijk leidt tot langere observatietijd voordat photobleaching optreedt. En gebruiken deze mutant voor de experimenten beschreven in dit proefschrift.

Valorisatie is een van de drie kerntaken van Nederlandse universiteiten. Niet elk onderzoek geeft een duidelijke mogelijkheid tot valorisatie. Het project beschreven in dit proefschrift, echter, heeft een grote commerciële potentie en zal, wanneer succesvol, een aanzienlijke impact hebben op de maatschappij. **Hoofdstuk 6** geeft inzicht in de strategische besluitvorming voor vindingen met een marktpotentieel die hun oorsprong vinden in de wetenschap. Wij bieden een overzicht van aspecten die in ogenschouw genomen moeten worden voor een patent wordt aangevraagd en wij vormden een commercialisatie strategie gebaseerd op een technologie assessment, markt analyse en potentiële gebruikers.

Dit proefschrift beschrijft onze inspanningen die geleid hebben tot een eerste proof-of-concept van onze op FRET gebaseerde eiwit fingerprinting methode op enkel-molecuul niveau. Hoewel er wellicht nog een lange weg te gaan is voor deze techniek de kliniek bereikt, is de haalbaarheid van de technologie

aangetoond. Hopelijk biedt het hier gepresenteerde werk een basis voor verdere ontwikkeling van het platform voor eiwit fingerprinting op enkel-molecuul niveau.

Acknowledgements

Many of our efforts to make this project a success are written in the pages you (might have) just read. However, even more of the blood, sweat and tears involved in this work never reached this thesis. This final chapter is the most exciting and at the same time sad chapter to write. It is sad because it marks the end of my PhD, both literally and metaphorically. It is the most exciting, because there are so many people that, all in their own way, helped me get where I am today. The following pages are devoted to those who affected me most and if you stay around as long as I did, it becomes quite list!

I would like to start with expressing my gratitude to Cees, not only as my promotor, but also for introducing me to the BN department and bringing me in contact with Chirlmin. Cees, I enjoyed our interesting, (scientific) discussions over the years. The way you approach science and your out-of-the-box thinking capabilities are admirable.

Chirlmin, most of my appreciations go to you. Before I decided to join your lab, I did my homework and made some inquiries and the consensus was you are a very intelligent, skilled and pragmatic scientist. These qualities appealed to me and, at least for me, it seemed to be a very good match. Over the years, I have enjoyed working with you a lot. I have seen our lab go through enormous growth and I have a lot of respect for how you handled the changing interactions and dynamics. I mostly appreciated the freedom you provided in my project and the space you gave me to explore options outside of research, leading to where I am today. It was an honor to be your first PhD student and I hope many will follow. After years of benefitting from your advice, it is now time for me to give you some advice. You are doing extremely well in building and leading your lab. Be proud of your achievements and celebrate your successes! Like a wise woman ones said: “The more you praise and celebrate your life, the more there is in life to celebrate.” (Oprah Winfrey).

Of course, I would like to thank my committee for spending time and energy to read and evaluate my work. Manfred Wuhrer, your input from a mass spectrometrist’s perspective is especially valuable for the future developments of our technology. I hope this will be the start of a fruitful collaboration. Giovanni Maglia, I am certain you biological pores will boost the nanopore approach of our project. Andreas Engel, we haven’t interacted much in the past, but I have really

enjoyed our conversation recently. It seems like Max shared my opinion. David Dulin, I am so honored to have you in my defense committee. I want to wish you the best of luck in your career as a PI. Marileen Dogterom, it was a pleasure to work in your department. Last, but definitely not least, Anne Meyer, thank you for all your advice over the years. Without you we wouldn't have been able to take the project to the level we are today. I enjoyed our conversations on research, but also on culture, family life and careers.

Dick de Ridder, we have you and Yao, your excellent student, to thank for the expertise in bioinformatics you brought to this project. Your help, and of course the help of Margreet, demonstrated the feasibility of our project, making it that much stronger.

As a first member of the Joo-lab I have felt lost many times. Ilja, Serge, Susan, Jaap and Jacob, many, many, many thanks to you for taking such good care of me from the start. Even with my sometimes endless flood of, most likely stupid, questions. Inge, Jan, Anke, Roland, Eve, Sacha, Marek, Erwin, Esengül, Anne, Jaco and Theo I sincerely believe the BN department will colapse without you! Thanks to all of you for always being there and the fun times we had. Anna, special thanks go to you. The Joo-lab would have exploded without you (not a joke!). I have enjoyed working with you a lot over the years and I will miss your "Friese nuchterheid". Dimitri and Jelle, thank you for your help in making the most amazing solutions for our setups. Jelle, I will miss your delicious birthday cakes! Dijana, Emmylou, Amanda, Chantal and Jolijn, also many thanks to you for always arranging everything to perfection, both for the Joo-lab and for the BN department. Esther, Liset, Angela and others, the department can't run without you, let alone expand as much as it has.

With my protein sequencing project I sometimes felt like an isolated entity in between the "RNA people". This changed after the team expanded. Laura and Mike, I am very happy to have you by my side as my paranymphs. I wouldn't know two persons more qualified to back me up, now let's hope it will not be necessary. Laura, I admire your positive and can-do attitude. I am convinced you will take nanopore protein sequencing a long way! Mike, I enjoyed supervising you a lot and I am very delighted (and a little bit proud) we could make you so excited about our single-molecule protein sequencing project that you will continue your PhD in Chirlmin's lab. I want to wish you the very best of luck and you know where to find me if you need any advice. Many of the struggles that, unfortunately, didn't make it into these pages where in substrate design and labeling. Pawel and Malwina, thank you so

much for your never ending efforts to try to tame titin, who would have thought this protein could be such a pain in the *#$%&. What started as a small side project in collaboration with Yamuna Krishnan from the University of Chicago, turned into a multi-person effort. Giacomo, Jeroen and Ivo, although your results in our i-Switch project did not make it into my thesis, your work was very much appreciated.

"RNA people" of the Joo-lab, thank you for tolerating my "dirty" research in your RNAse free lab. We had many fun parties, drinks, dinners, trips, etc that I will miss very much. I might even miss the many meetings we had… Kay, special gratitude goes to you for teaching me your excellent molecular biology skills. I wish you and your family a lot of love and prosperity. Stanley, it was impressive to see you develop your skills over the years. Good luck in finding your next challenge! Mohamed, although you might appear a bit disorganized on the outside, you have an amazing mind. I sincerely hope you can start your own lab, maybe even in a bit warmer climate. Viktorija and Thijs, please take good care of the lab after all the grandpa's and grandma's have left. Luuk, good luck with finishing up your PhD. I am secretly very glad you did not surpass me in graduating, that would have made me feel really bad. We all expect great things from you, you know that right. Tim, Sung Hyun, Sungchul, Seung Hwan and Iasonas, thank you for the fun time and nice conversations in the lab and the office, and outside the BN department of course.

When I joined the department, the number of women was limited, making it even more special I had the privilege to share my office with two amazing girls. Michela and Mathia, you made my life as a PhD student so much more cheerful, meaningful, bearable and enjoyable. I could always share my frustrations, little victories or personal issues. Unfortunately, it were not only happy times we shared, but due to the unhappier times, I got to know two very strong and positive women. Too bad, time comes we have to move on! I will miss you!

I would also like to thank the BN members that made me feel at home from the beginning, Jan, Greg, Felix, Fabai, Zohreh, Regis, Stefan, Christophe, Iwijn, Marijn, David, Bertus, Calin, Tim and Meghan, Charl, Elio, Aviva, Sriram, Maarten, Francesco, Magnus and others, thank you for all the fun at parties, borrels, conferences and of course in the BN coffee corner! Also many thanks for the fun times and many memories I share with those who joined the department after I arrived: Pauline, Mahipal, Adi, Natalia, Yaron, Gautam, Bojk, Tessa, Victor, Greg, Vanessa, Sam, Anthony, Jakub, Helena, Daniel, Yoones, Afke, Dominik, Jorine, Johannes, Stephanie, Misha, Alicia, Jonás, Benjamin, Federico, Richard, Fabrizio,

Yoones, Mariana and so many, many others. Louis, please stay away from the canals. Florian, Roy and Sumit, you could never compete with my previous office mates. Nevertheless, it was a pleasure sharing an office with you. Orkide, special thanks to you for having me and my family at your wedding. All the best for you and Paul.

During part of my time at BN I had the pleasure of being a member of the Casimir Research School PhD Platform. Maddy, Simone, Jan Ruitenbeek, Marije and the members of the platform, I enjoyed our meetings and interaction very much. I have seen valuable developments in the Casimir curriculum and I hope the spring school will be a success for many years to come.

I want to acknowledge Jennifer, Ard Ellens, Dap Hartmann and Olga for all their help towards commercialization of my project. Over the years you have sparked my interest in valorization, entrepreneurship, innovation and IP management, leading to where I am today. Also a word of thanks to my colleagues at Erasmus MC TTO for giving me the opportunity to pursue a career in this direction. I am grateful to have landed in such a warm nest. Kaarmuhilan, Vidhvath, Prem and Valery thank you for your new insights to my project during the Turning Technology into Business course, it was motivating and fun to see what applications you came up with.

Naast de vele collega's en anderen waarmee mijn wegen op professioneel vlak kruisten, zijn er ook een heel aantal personen waarmee ik het juist af en toe niet over werk hoefde te hebben. Bart, Theun, Maarten en oud-Yupjummers, helaas wonen we niet meer zo dichtbij elkaar als voorheen in in Enschede. Desalniettemin, lijkt het of wij gewoon verder gaan waar we gebleven waren wanneer wij elkaar weer zien en weten we elkaar bij grote levensgebeurtenissen toch weer te vinden. Bedankt voor alle gezellige momenten de afgelopen jaren. Bart, Berdien, Femke, Janine, Jobke, Julie, Luuk, Manon, Marjolein, Mathieu, Redmar, Renske, Rob, Saskia, Sonja, Sovianne, Sven en alle aanhang, bedankt voor alle kerstdiners, weekendjes, heerlijk avondjes en andere evenementen die voor de nodige afleiding zorgende de afgelopen jaren. De meeste van jullie ken ik al (bijna) twintig jaar en ik vind het bijzonder en fijn dat wij, als Le Metropolitain, nog steeds met zoveel plezier elkaars gezelschap opzoeken. Ook is het bijzonder hoeveel van jullie mij al zijn voorgegaan en hun doctorstitel inmiddels op zak hebben. Bart, Erwin, Jobke, Janine, Bas en Marjolein, het was fijn om over de jaren heen mijn ervaringen met jullie te kunnen delen. Jobke en Erwin, ook was het fijn om jullie als voorbeeld te hebben en te zien dat een gezinnetje en promoveren best goed samen gaat. Rens, bedankt dat ik bij jou altijd mijn ei kwijt kan!

Saskia, Marijn, Paul, Hidde & Robin, ik had mij geen leukere schoonfamilie kunnen wensen. Bedankt dat jullie altijd interesse hebben in wat ik doe. Ik geniet keer op keer van de ontspannen en gezellige avonden, etentjes, weekendjes, etc. Sas en Marijn, extra dank dat jullie je graag over Max ontfermen, zodat Bart en ik ons werk, en ons sociale leven, af en toe wat extra aandacht konden geven.

Pap, Mam, Karin, Niels en Nanneke, bedankt dat jullie mij altijd gesteund hebben. Ka en Nan, hoewel wij alle drie een compleet ander pad zijn ingeslagen, herken ik veel van mijzelf in wat jullie drijft. Ik bewonder jullie doorzettingsvermogen, ook als dingen even niet gaan zoals gepland (en wat houden wij van plannen!). Pap en Mam, ik denk dat jullie ons bovenal geleerd hebben om die dingen na te streven waar we gelukkig van worden en wat ons een beter mens maakt. Ik ben jullie hier heel dankbaar voor.

Bart, lief, je wilde in dit dankwoord een eigen pagina en die heb je ook verdiend. Zonder jou was dit boekje er zeker niet gekomen. Bedankt voor het aanhoren van mijn gemopper wanneer het me even te veel was, het klaar zitten met een heerlijke maaltijd wanneer ik veel later thuis was dan gepland, je liefdevol ontfermen over Max wanneer ik weer eens geen tijd of energie had, me achter mijn computer vandaan trekken wanneer het wel weer genoeg was geweest voor die dag. Bedankt voor al die dingen die ik veel te weinig tegen je zeg, bedankt dat je er altijd voor me bent! Ik zou niet weten wat ik zonder je zou moeten!

En lief, wie had gedacht dat wij samen zoiets moois konden maken. Max, je bent nog geen drie jaar oud en ik heb al zoveel van je geleerd. Jij hebt mij laten zien wat echt belangrijk is in het leven. Je hebt mij geholpen om mijn werk in het juiste perspectief te zetten en geleerd om afstand te nemen. En wanneer ik dat even vergeet, dan help je mij dat graag herinneren: “Mama je moet minder werken en meer spelen met Max!”. Je bent een wijs mannetje! Jouw zusje zit nu nog in mijn buik, maar ik weet nu al dat ze de meest geweldige grote broer krijgt die ze zich kan wensen.

Jetty van Ginkel
Delft, November 2016

Curriculum Vitæ

Hendrika Geertruida Theodora Maria van Ginkel

22-05-1984	Born in Gouda, The Netherlands
1996-2002	Secondary Education Isendoorn College Warnsveld
2002-2008	B.Sc. in Biomedical Engineering University of Twente Enschede
2005-2010	M.Sc. in Biomedical Engineering University of Twente Enschede
2011-2016	Ph.D. research Department of Bionanoscience, Technical University of Delft *Promotor:* Prof. dr. C. Dekker *Copromotor:* Dr. C. Joo

List of Publications

6. **J. van Ginkel**, M. Filius, M. Szczepaniak, P. Tulinski, A.S. Meyer & C. Joo, Single-molecule peptide fingerprinting, (*In preparation*)

5. **J. van Ginkel**, M. Filius, P. Tulinski, C. Joo & A.S. Meyer, Single-molecule observation of ClpXP substrate recognition, (*In preparation*)

4. **J. van Ginkel***, L. Restrepo-Pérez*, C. Dekker & C. Joo, The road to single-molecule protein sequencing, (*In preparation*).

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

2. C. Joo, C. Dekker, **H.G.T.M. van Ginkel** & A. S. Meyer, Single molecule protein sequencing, WO patent 2014014347. (2014).

1. H. Alves*, **J. van Ginkel***, N. Groen, M. Hulsman, A. Mentink, M. Reinders, C. van Blitterswijk, & J. de Boer, A mesenchymal stromal cell gene signature for donor age, *PLoS One* **7**, e42908, (2012).

* *These authors contributed equally*

Casimir PhD Series, Delft - Leiden 2016-37
ISBN 978-90-8593-281-9