TU Delft

Delft University of Technology

Next-Generation Protein Identification

Advancing Single-Molecule Fluorescence Approaches

Filius, M.

DOI
10.4233/uuid:29bd3863-7008-4825-9856-34ee7beafb56

Publication date
2022

Document Version
Final published version

Citation (APA)
Filius, M. (2022). *Next-Generation Protein Identification: Advancing Single-Molecule Fluorescence Approaches*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:29bd3863-7008-4825-9856-34ee7beafb56

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Next-Generation Protein Identification

Advancing Single-Molecule Fluorescence Approaches

Next-Generation Protein Identification

Advancing Single-Molecule Fluorescence Approaches

Proefschrift

ter verkrijging van de graad van doctor aan
de Technische Universiteit Delft,
op gezag van de Rector Magnificus
Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
donderdag 27 januari 2022 om 12:30.

door

Mike FILIUS

Master of Science in de Biomoleculaire Wetenschappen
Vrije Universiteit Amsterdam, Nederland
geboren te Vlaardingen, Nederland

Dit proefschrift is goedgekeurd door

Promotor:	Prof.dr. C. Joo &
Promotor:	Prof.dr. C. Dekker

Samenstelling van het promotiecommissie:

Rector Magnificus		Voorzitter
Prof.dr. C. Joo	Technische Universiteit Delft	Promotor
Prof.dr. C. Dekker	Technische Universiteit Delft	Promotor
Onafhankelijke leden:		
Prof.dr. P.G. Steeneken	Technische Universiteit Delft	
Prof.dr. G. Maglia	Rijks Universiteit Groningen	
Prof.dr. M.T.C. Walvoort	Rijks Universiteit Groningen	
Prof.dr. Tae-Young Yoon	Seoul National University	
Reserve lid:		
Prof. dr. M. Dogterom	Technische Universiteit Delft	

Geprint door : Gildeprint
Omslag ontwerp: Ella Marushchenko

Copyright © 2022 door Mike Filius
Casimir PhD series: 2021-47
ISBN: 978-90-8593-510-0

Een electronische versie van dit proefschrift is te vinden op:
http://repository.tudelft.nl/

To my parents,

who taught me that with hard work, dedication and passion, anything is possible.

Table of Contents

Chapter 1. The Emerging Landscape of Single-Molecule Protein Sequencing Technologies 1

1.1 Abstract 2
1.2 Introduction 3
1.3 A renaissance of classical techniques 5
1.3.1 Massively parallel Edman degradation 6
1.3.2 Single-molecule mass spectrometry 7
1.3.3 Tunneling conductance measurements 10
1.4 DNA nanotechnologies for protein sequencing 11
1.4.1 Fingerprinting via DNA PAINT 11
1.4.2 DNA proximity recording 13
1.4.3 Protein fingerprinting using FRET 13
1.5 Biological and solid-state anopores 14
1.5.1 Reading the amino acid sequence of linearized peptides 16
1.5.2 Fingerprinting linearized proteins 17
1.5.3 Characterization and identification of folded proteins 17
1.6 Chemistry for next-generation proteomics technologies 18
1.7 Discussion: a spectrum of opportunities 21
1.7.1 Challenges for next-generation protein sequencing 23
1.8 Thesis Outline 24
1.9 References 25

Chapter 2. Single-Molecule Peptide Fingerprinting 31

2.1 Abstract 32
2.2 Introduction 33
2.3 Results 34
2.4 Discussion 39
2.5 Materials And Methods 41
2.5.1 $ClpX_6$ purification and biotinylation 41
2.5.2 ClpP mutations, purification and labeling 41
2.5.3 ClpP inactivation 41
2.5.4 ClpXP cleavage reaction 42
2.5.5 Substrate preparation 42
2.5.6 Single-molecule sample preparation 42
2.5.7 Single-molecule fluorescence 43
2.5.8 Data Acquisition 43
2.6 Supporting Information 44
2.6.1 Supporting Figures 44
2.6.2 Supplementary Table 48
2.7 References 50

Chapter 3. High-Resolution Single-Molecule FRET via DNA eXchange (FRET X) 53

3.1 Abstract 54
3.2 Introduction 55
3.3 Results 55
3.4 Discussion 64
3.5 Materials and Methods 65
3.5.1 Single-Molecule Setup 65
3.5.2 Single-Molecule Data Acquisition 65
3.5.3 Data Analysis 66
3.6 Supporting Information 67
3.6.1 Supporting Figures 67
3.7 References 79

Chapter 4. Evaluation of FRET X for Single-Molecule Protein Fingerprinting 81

4.1 Abstract 82
4.2 Introduction 83
4.3 Approach 84
4.3.1 FRET X for protein fingerprinting 84
4.3.2 Fingerprinting simulations 86
4.4 Results 86
4.4.1 Experimental FRET X fingerprinting of model peptides 86
4.4.2 Fingerprinting simulation of protein spliceoforms 87
4.4.3 Analysis of simulated protein mixtures 90
4.4.4 Robustness against suboptimal experimental conditions 90
4.5 Discussion 92
4.6 Materials and Methods 94
4.6.1 Peptide Labeling 94
4.6.2 Single-Molecule Setup 94
4.6.3 Single-Molecule Data Acquisition 94
4.6.4 Data analysis 95
4.6.5 Simulations 95
4.6.6 Lattice structure 96
4.6.7 Tag implementation 96
4.6.8 Simulated labeling scenarios 96
4.6.9 Structure collection 97
4.6.10 Folding simulation 97
4.6.11 Fingerprint extraction 98
4.6.12 Classification 98
4.7 Supporting Information 99
4.7.1 Supporting Figures 99
4.7.2 Supporting Tables 106
4.8 References 108

Chapter 5. Single-Molecule Protein Identification Using FRET X 111

5.1 Abstract 112
5.2 Introduction 113
5.3 Results 114
5.3.1 Single-molecule protein fingerprinting using FRET X 114
5.3.2 FRET X fingerprinting of globular proteins 117
5.4 Discussion 118
5.5 Materials and Methods 120
5.5.1 Protein expression and purification 120
5.5.2 Biomarker labeling 120
5.5.3 Single-molecule setup 121
5.5.4 Single-molecule data acquisition 121
5.5.5 Data analysis 121
5.5.6 Protein fingerprinting simulation 122
5.6 Supporting Information 123
5.6.1 Supporting Figures 123
5.6.2 Supporting Tables 127
5.7 References 128

Chapter 6. High-Speed Super-Resolution Imaging Using Protein-Assisted DNA-PAINT 131

6.1 Abstract 132
6.2 Introduction 133
6.3 Results 134
6.4 Discussion 139
6.5 Materials And Methods 140
6.5.1 Expression and purification of CbAgo 140
6.5.2 Single-molecule setup 140
6.5.3 Single-molecule data acquisition 141
6.5.4 Assembly of DNA oligo plate 141
6.5.5 Super-resolution data analysis 142
6.6 Supporting Information 143
6.6.1 Supporting Figures 143
6.6.2 Supporting Tables 149
6.7 References 151

Chapter 7. Towards Single-Molecule Protein Sequencing: Outlook and Concluding Remarks **153**

7.1 Proteins are the workhorses of the cell 154
7.2 Amino acid analysis of proteoforms 155
7.3 Deciphering proteoforms at the PTM level 156
7.4 Affinity-based approaches for single-molecule proteoform analysis. 158
7.5 Towards single cell proteomics. 158
7.6 Industry opportunities for single-molecule protein sequencing 161
7.7 Concluding remarks 162
7.8 References 164

Summary **167**

Samenvatting **171**

Acknowledgments **175**

Curriculum vitae **181**

List of Publications **183**

1

The Emerging Landscape of Single-Molecule Protein Sequencing Technologies

Javier Alfaro*, Peggy Bohländer*, Mingjie Dai*, **Mike Filius***, Cecil J. Howard*, Xander F. van Kooten*, Shilo Ohayon*, Adam Pomorski*, Sonja Schmid*, Aleksei Aksimentiev, Eric V. Anslyn, Georges Bedran, Cao Chan, Mauro Chinappi, Etienne Coyaud, Cees Dekker, Gunnar Dittmar, Nicholas Drachman, Rienk Eelkema, David Goodlett, Sebastien Hentz, Umesh Kalathiya, Neil L. Kelleher, Ryan T. Kelly, Zvi Kelman, Sung Hyun Kim, Bernhard Kuster, David Rodriguez--Larrea, Stuart Lindsey, Giovanni Maglia, Edward M. Marcotte, John P. Marino, Christophe Masselon, Michael Mayer, Patroklos Samaras, Kumar Sarthak, Lusia Sepiashvili, Derek Stein, Meni Wanunu, Mathias Wilhelm, Peng Yin, Amit Meller, and Chirlmin Joo.

* These Authors have contributed equally to this work

Nature Methods
Nature Methods. Volume 18, pages 604–617 (2021)
DOI: 10.1073/pnas.1707207115



1

1.1 Abstract

Proteins perform most functions crucial in all living systems and are involved in all structures and biochemical reactions of living cells. To understand the fundamental molecular processes governing all biological systems, as well as to monitor and combat emerging diseases, it is imperative that there exist techniques that permit sequence determination, identification and classification of proteins down to the single-cell and single-molecule levels. Currently, it is challenging to identify and sequence proteins due to the insufficient instrument sensitivity and difficulties in analyzing peptides within mixtures, hampering efforts to comprehensively understand proteomes. A diversity of new single-molecule protein-sequencing and protein-identification technologies, along with innovations in the currently available strategies, will eventually enable high-coverage single-cell profiling. These emerging methods make use of single-molecule fluorescence, nanopores, and other nanotechnologies to sequence or identify individual proteins. The ultimate precision and sensitivity of proteomes promised by these single-molecule-resolution technologies is expected to create many new directions in research and biomedical applications, from enabling global proteomics of single cells and bodily fluids to sensing and classifying low-abundance protein biomarkers for disease screening and precision diagnostics.

1.2 Introduction

The emergence of next-generation sequencing and single-molecule DNA sequencing technologies has revolutionized genomics and, consequently, has profoundly altered precision medicine diagnostics. Proteomics awaits similar transformative waves of protein sequencing techniques that will allow for the examination of proteins at the single-cell and ultimately single-molecule level, even with low-abundance proteins. The proteome is not a direct reflection of the transcriptome, and the way that RNA abundance relates to protein abundance varies from transcript to transcript. Further, the post-translationally modified proteome is inaccessible from the transcriptome. Therefore, whole-proteome sequencing and profiling of the vast repertoire of cell types is expected to fundamentally enhance understanding of all living systems. This necessitates analysis of the proteome with ultra-high resolution, complementing today's single-cell RNA sequencing studies.

DNA sequencing technologies are routinely used for whole-genome and whole--transcriptome profiling with extensive read depths and high sequence coverage. In the absence of an amplification method similar to those available with DNA, conventional bottom–up mass spectrometry (MS)-based proteomics assays fall short of providing the same breadth of view for proteins (**Box 1.1**). Analysis of complex protein mixtures is particularly challenging because the more than 20,000 genes in the human genome[1] are translated into a diversity of proteoforms that may include millions of variants as a result of post-translational modifications, alternative splicing and germline variants.[2] In cancer, for example, the proteoform landscape can be aberrant with many new protein variants resulting from non-canonical splicing, mutations, fusions and post-translational modifications. Characterization of such proteoforms is likely to benefit from improvements in current protein sequencing techniques and the emergence of new methods.

MS remains a staple of protein identification and continues to develop toward single-cell methods (**Box 1.2**). In addition, a diverse range of protein sequencing and identification techniques have emerged that aim to increase the sensitivity of proteomics to the single-molecule level. Many of these techniques rely on fluorescence and nanopores for single-molecule sensing as an alternative means to sequence or identify proteins (**Figure 1.1**). The landscape of emerging proteomics technologies is already vast, with different approaches at various stages of development, some of which have already secured industry investment[3,4], an important step toward broad dissemination to the research community. Other technologies have shown great promise and gained popularity among the single-molecule biophysics community, while others are available as proofs of concept at just one or a few laboratories.

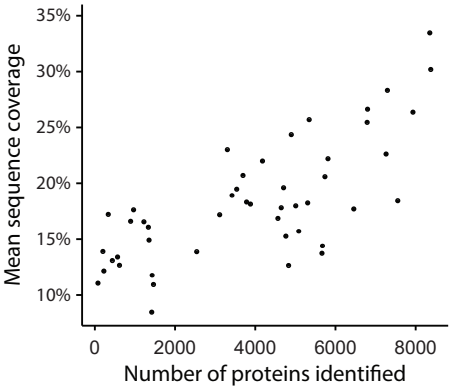
Here we describe prominent emerging protein sequencing and fingerprinting techniques in the context of mature methods such as MS-based proteomics, discuss challenges for their real-world application and assess their transformative potential.

Box 1.1: Mass spectrometry-based global proteomics

The last decade saw the maturation of MS use in global proteomics. The typical proteomics workflow is 'bottom–up' in nature and involves digesting a protein sample using a protease and characterizing the resulting peptides by MS[114]. Two types of measurements are typically made in succession: (1) MS^1 spectra survey the masses of a set of peptides present in the mass spectrometer at a given moment and (2) MS^2 spectra probe the structures of peptide ion species identified in the MS^1 survey by isolating, fragmenting and measuring the fragment masses of one or a few of them. Peptides identified from the MS^2 spectra are then mapped back to proteins to infer overall protein abundance.

Current mass spectrometers have drawbacks in terms of their dynamic range, the read length (peptide length) of 'sequenced' peptides and biases in detectability arising from the ionization mechanism, transmission and the mass analyzer used. Consequently, although 'top–down' proteomics methods capable of analyzing intact proteins exist[115], most state-of-the-art proteomics approaches characterize the proteome with high numbers of proteins but on average characterize proteins with low sequence coverage and low sequencing depth. Different sample preparation strategies, instruments and elution profiles can improve the numbers and average sequence coverage of the proteins identified in an experiment. Summarizing the best single-sample run from 47 experiments (a summary of over 1,000 distinct samples) in ProteomicsDB[116] revealed that, even with complex sample preparation, the mean sequence coverage (the average percentage of amino acids covered in an identified protein) for a single sample reaches just 33%. The resulting challenge in proteoform inference is demonstrated in studies evaluating the sensitivity of detection for various cancer aberrations in proteomics datasets. For example, in a study of over 30 sample process replicates, only about 10% of germline and somatic single-nucleotide variants detected at both the DNA and RNA level were detectable as peptides, and an even smaller proportion of peptides corresponding to novel splice junctions were detected that had been observed with RNA sequencing.[117]

Figure B1.1: Sequence coverage in global proteomics studies.
MS-based global proteomics studies identify and quantify the proteins with variable sequence coverage. The single best run from 47 publications present in proteomicsDB shows how sample-specific protein sequence coverage improves with sample preparation methods. Sequence coverage generally decreases with sample complexity and increases with time (cost) dedicated to studying the sample.



1.3 A renaissance of classical techniques

Edman degradation, MS and enzyme-linked immunosorbent assay (ELISA) have been broadly used for protein/peptide sequencing and identification for several decades; therefore, it is no surprise that further enhancements of these classical technologies are being sought. The biophysics community has been developing methods to increase the throughput[5] and sensitivity[6] of single-molecule ELISA, Edman degradation, single-particle MS, neutral-particle nanomechanical MS and single-particle electrospray. Even established tools commonly used in materials science, such as electric tunneling and direct current measurements, can be repurposed for protein sequencing.

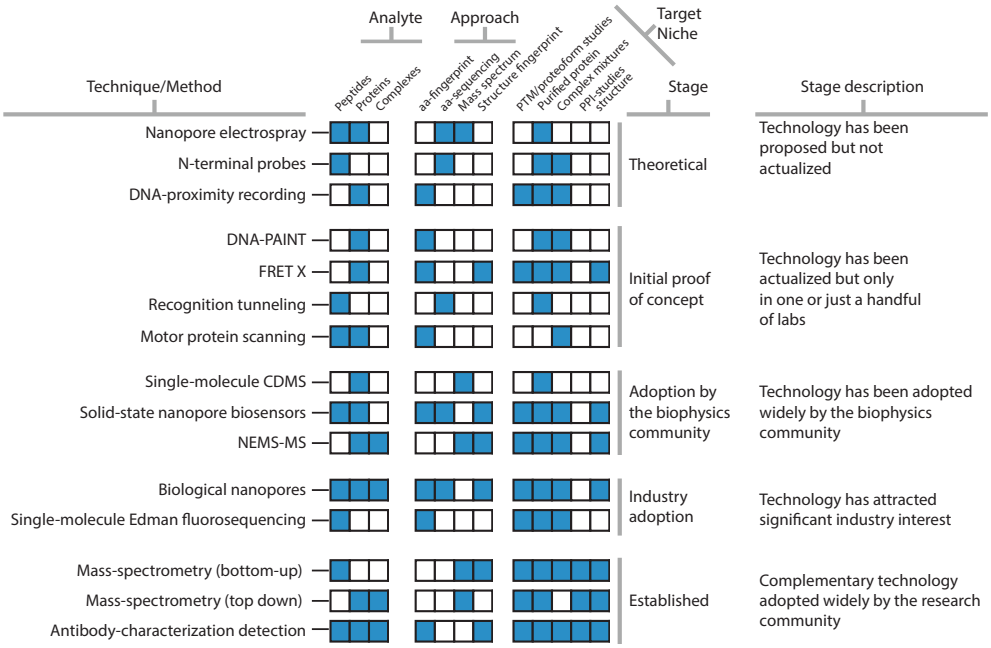


Figure 1.1: The emerging landscape of single-molecule protein sequencing and fingerprinting technologies. The new technologies address a range of analytes, methods of protein identification and target niches. Various techniques, particularly those involving complex readout signals, are suitable for characterizing short peptide sequences, while others are primed to characterize full-length proteins or larger complexes. The method of protein identification may fingerprint certain classes of amino acids (AA fingerprint) or reveal each amino acid down to its physiochemical class or better (AA sequencing). Technologies might characterize proteins by their mass or the mass of their fragments (mass spectrum). Other methods aim to characterize the properties of folded proteins (structure fingerprint). PTM, post--translational modification; PPI, protein–protein interaction; NEMS-MS, nanoelectromechanical systems MS.

1.3.1 Massively parallel Edman degradation

Edman degradation[7] was the first method to determine the amino acid sequence of a purified peptide. The method entails chemical modification of the N-terminal amino acid, cleavage of this amino acid from the peptide and determination of the identity of the cleaved labeled amino acid using high-performance liquid chromatography. Until recently, conducting sequencing of this sort in a massively parallel fashion was not feasible because the method requires highly purified peptides. However, recent multiplex strategies that use peptide arrays and either sequence chemically labeled peptides ('fluorosequencing') or successively detect the N-terminal amino acid are making breakthroughs.

Fluorosequencing combines Edman chemistry, single-molecule microscopy and stable synthetic fluorophore chemistry (**Figure 1.2**a). Proteins are digested to shorter peptides and immobilized on a glass surface using the C terminus.[8] Millions of individual fluorescently labeled peptides can be visualized in parallel, and changing fluorescence intensities are monitored as N-terminal amino acids are sequentially removed through multiple rounds of Edman degradation. The resulting fluorescence signatures serve to uniquely identify individual peptides.[8] This method allows for millions of distinct peptide molecules to be sequenced in parallel, identified and digitally quantified on a zeptomole scale.[9] Specific amino acids are covalently labeled with spectrally distinguishable fluorophores, and the peptide fingerprint comes from measuring the decrease in fluorescence of peptides following Edman degradation.[9] Much as in MS, the partial sequence is mapped back to a reference proteome within a probabilistic framework.

The technology is not without challenges, as the reagents used for Edman degradation chemistry lead to increased rates of fluorescent dye destruction, which in turn limits the read length. These reagents include slightly basic structures such as pyridine, strong acids such as trifluoroacetic acid and the electrophile phenyl isothiocyanate. Furthermore, the reliance on chemical labeling leads to partial sequencing of the peptide, with the unidentified remainder inferred by comparison to a reference proteome. In addition, inefficient labeling can lead to errors that must be modeled into the reference proteome comparison, spurring the development of new protocols to increase yields.[10] Exciting new proposals could add the dimension of protonation-based sequencing. The pKa of the N-terminal amino acid could be used for identification by observing and interpreting the protonation–deprotonation signal of the peptide at fixed pH through the Edman degradation process.[11] Much like fluorosequencing, the signal observed would be for the whole peptide and the decay pattern would be interpreted to derive a pKa for each N-terminal amino acid.

Several natural proteins and RNA molecules recognize specific amino acids either as free amino acids or as a part of a polypeptide chain.[12] These proteins and nucleic acids provide different solutions for N-terminal amino acid recognition. Each N-terminal amino acid binder (NAAB) probe selectively identifies a specific N-terminal amino acid or an N-terminal amino acid derivative. With each cycle, another amino acid is revealed in the sequence of the peptide. However, further directed evolution and engineering of NAAB probes is required to meet the stringent affinity, selectivity and stability requirements for error-free sequencing applications. In addition, such

Box 1.2: MS-based single-cell proteomics

The dream of extending MS-based proteomics to the single-cell level has eluded researchers for decades. Even as the sensitivity of MS instrumentation has improved to provide single-cell-compatible detection limits, in practice, samples comprising at least thousands of cells have been required to obtain an in-depth proteome profile. Two recent advances have made single-cell proteomics a reality. Miniaturized sample processing workflows such as nanodroplet processing in one pot for trace samples (nanoPOTS)[118] have dramatically increased the efficiency of single-cell sample preparation. NanoPOTS utilizes a robotic nanopipettor to interface with a microfabricated nanowell plate. The reduced surface contact and increased protein concentrations within the nanoliter-sized droplets dramatically enhance digestion kinetics and increase sample recovery for single cells and other trace samples. Concurrently, multiplexed strategies (for example, single-cell proteomics by mass spectrometry, SCoPE-MS)[119] have been developed in which proteins from single cells are labeled with unique isobaric tags and several cells are analyzed together in the presence of a larger carrier sample. The single cells and carrier provide a combined MS signal for each protein, and unique reporter ions released upon fragmentation enable protein quantification for each cell. While nanoPOTS and SCoPE-MS originally enabled quantification of hundreds of proteins[119,120], the combination of these two techniques, as well as advances in miniaturized liquid chromatography and gas-phase separation, now enables more than 1,000 proteins to be quantified from single mammalian cells.[121]

probes would need to discriminate among all amino acids, including the same amino acid in alternative positions in the peptide sequence. Probes that bind a class of N-terminal amino acids (for example, short aliphatic residues) could also be useful but would introduce ambiguity in the sequencing process. Different probes could also be designed to recognize short N-terminal k-mers, which would increase the number of probes needed but reduce the ambiguity in the resulting sequencing information. To circumvent this limitation, it may be possible to sequence the N-terminal amino acid by selective recognition using a plurality of probes in each cycle of Edman degradation[13,14] (**Figure 1.2**b).

1.3.2 Single-molecule mass spectrometry

MS is a century-old method that measures the mass-to-charge (m/z) ratio of ions, in particular, charged peptides/proteins and their assemblies. Single-ion detection has been possible since the 1990s, for example, in Fourier-transform ion cyclotron resonance instruments.[15] Charge detection MS (CDMS) is a single-ion method where the charge assignment of each individual ion is determined directly, enabling conversion of the mass-to-charge ratio into the neutral mass domain. This approach has focused on the analysis of large biomolecular complexes, especially viruses in the range of 1–100 MDa.[16] While previously CDMS was limited to specialized instrumentation, the past year has seen breakthroughs built on early work producing mass spectra of single ions in Orbitrap mass analyzers.[17,18,19] Today, these mass analyzers can be used to directly derive the charge states of single proteins and even their fragment ions.[20]

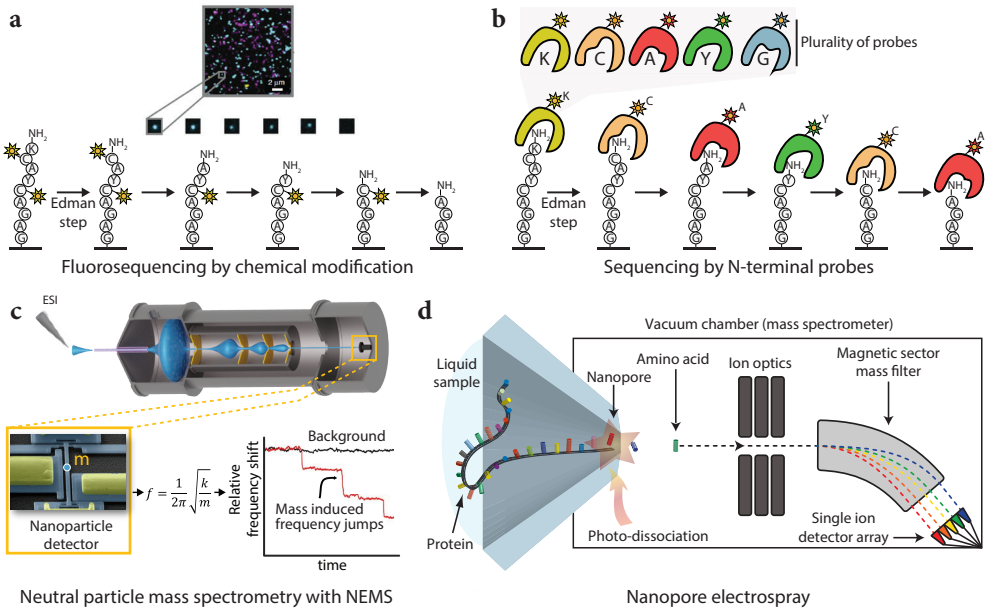


Figure 1.2: The renaissance of classic techniques.
(**a,b**)High-throughput fluorosequencing by Edman degradation featuring amino acid-specific chemical modification of peptides with fluorophores **(a)** and N-terminal amino acid recognition using a plurality of probes **(b)**. **(c)** Neutral-particle MS is a promising technique to characterize proteoforms. Currently, the technology can be used to characterize large megadalton-scale complexes using silicon-based nanosensors. Graphene nanosensors and further developments may push the technology toward smaller and smaller proteins and potentially lead to increased sequence coverage in global proteomics. ESI, electrospray ionization. **(d)** Nanopore electrospray is a marriage of nanopores, classical electrospray and single-particle detection techniques to sequence single proteins by measuring amino acids one at a time. Panel **A** adapted with permission from ref.[9]

Orbitrap instruments are particularly useful because the readout of individual ions can be multiplexed by 100- to 1,000-fold in Orbitrap-based CDMS.[20] Individual ion MS has already shown resolution of mixtures with approximately 1,000 proteoforms that provided no data using standard MS.[20,21] This has greatly expanded the top–down approach to confirm DNA-inferred sequences of whole proteins, including localization of their post-translational modifications.[20,21,22] Without extensive alteration, Orbitrap mass analyzers can therefore measure tens of thousands of proteins in a matter of minutes. With these rapidly evolving technologies, charting the full human proteoform atlas has already begun[23], making strides toward a comprehensive human proteoform project. However, ionization is a critical requirement for MS of proteins and peptides, and not all peptides are efficiently ionized and transmitted through the mass spectrometer. This might restrict some of the proteoform mapping efforts, providing a niche for the other technologies in **Figure 1.1**.

For higher-molecular-weight species, the ionization of proteins and complexes yields a mixture of macro ions with variable charge states, resulting in a net reduction of sensitivity as the signal distributes over multiple peaks in the mass-to-charge dimension. Moreover, charge state distributions may overlap above a certain mass or in the case of mixtures, creating challenges in species identification. Since their inception[24], nanomechanical mass sensors have made tremendous progress toward protein characterization.[25] Such devices, which take the shape of cantilevers or beams with lateral dimensions in the range of hundreds of nanometers, can detect individual particles accreting onto their active surface through changes in vibration frequency. Importantly, as the inertial mass of a particle is determined directly from the frequency change, these devices are insensitive to charge states.[26] This realization prompted the development of new MS instrument designs devoid of ion guides, which no longer depend on electromagnetic fields to collect and transmit analytes (**Figure 1.2**c). Such a nanomechanical resonator-based MS system has recently been shown to have the ability to characterize large protein assemblies such as individual viral capsids above 100 MDa in size.[27] Outside of proteomics, a resolution of 1 Da has been demonstrated with carbon nanotubes.[28] Moreover, recent reports suggest the possibility of determining other physical parameters such as the stiffness or shape of the analyte by monitoring multiple vibrational modes.[29,30] These previously inaccessible metrics may open new avenues to discriminate peptides, proteins and their complexes. Nonetheless, one of the challenges of the nanoresonator mass spectrometer lies in devising efficient ways to bring individual proteins onto the resonator's active surface for mass sensing.

Ionization is commonly achieved by electrospray ionization of a solution containing the compound(s) of interest. The use of ever-smaller electrospray ion source apertures has led to substantial improvements in the sensitivity of MS.[31,32] Mass spectrometers with a nanopore ion source have been developed for the purpose of sequencing single proteins[33] (**Figure 1.2**d). A nanopore electrospray can potentially deliver individual amino acid ions directly into a high-vacuum gas phase, where the ions can be efficiently detected by their mass-to-charge ratios. This opens a path to sequencing peptides one amino acid at a time. The concept makes use of nanopores to guide a protein into a linear configuration so that its monomers can be delivered into the mass spectrometer sequentially.[34] Individual amino acids must be cleaved

from the protein molecule as it transits the nanopore, which could potentially be accomplished with photodissociation[35] or chemical digestion methods. The 100-MHz bandwidth of the channeltron single-ion detectors used in this setup is also sufficient to resolve the arrival order of the ions. The high mass resolution makes this technique promising for identifying post-translational modifications, which change the masses of particular amino acids by predictable amounts. One challenge on the path for this technology will be achieving high throughput, which might require a strategy for parallelizing mass analysis.

1.3.3 Tunneling conductance measurements

The appearance of the scanning tunneling microscope in the 1980s introduced a new way to analyze molecules. Small organic molecules can be transiently trapped between two metal electrodes with sub-nanometer separation, with the tunneling currents between the electrodes reporting on the molecular signature of the analyte. Recently, several technical advances have been made to move toward single-molecule amino acid and protein analysis. Extracting insightful information from electron tunneling is complicated by the noise resulting from water and contaminants reaching the electrode surfaces. To overcome this problem, recognition tunneling has been developed in which the electrodes are covalently modified with adaptor molecules that form transient but well-defined links to the target molecule.[36] The rapidly fluctuating tunnel current signals are processed using machine learning algorithms, which makes it possible to distinguish individual amino acids and small peptides.[37] Moreover, smaller electrode gaps have been introduced to obtain distinct signals from different amino acids and post-translational modifications38. Further development of the technology will depend on a reliable source of tunnel junctions with a defined gap to replace the cumbersome scanning tunneling microscopy, but it is clear that both the sequence and post-translational modifications of small peptides can be determined.[37] Currently, tunneling conductance is a proof-of-concept technology for fully sequencing short peptides that could one day be used for the analysis of protein digests and expanded to analysis of post-translational modifications (**Figure 1.1**).

Recently, it was discovered that electrical charges can be transmitted through a protein if the electrodes are bridged by a protein via formation of chemical bonds or ligand binding.[39] Specifically, changes in protein conformation upon nucleotide addition could be followed in real time from the direct currents passing through a DNA polymerase.[40] Although the observation was preliminary, the electronic signatures were distinctive when the polymerase was associated with different DNA sequences, enabling a new approach to label-free single-molecule DNA sequencing. A similar approach could potentially be used for protein sequencing with enzymes such as proteases or glycopeptidases that process substrates sequentially.

1.4 DNA nanotechnologies for protein sequencing

DNA nanotechnologies, in which a large number of sequences with prescribed pairing interactions and dynamic properties can be custom designed, have facilitated developments in fields ranging from synthetic biology to diagnostics and drug delivery.[41] For example, programmable transient binding between short DNA strands is central to the super-resolution technique of DNA-based point accumulation for imaging in nanoscale topography (DNA-PAINT)[42,43,44] (**Box 1.3**). Here we describe the application of DNA-PAINT and DNA-based local and global pairwise distance measurement methods for single-molecule protein detection and identification.

Box 1.3: DNA-PAINT

DNA-PAINT relies on the transient binding of dye-labeled DNA strands (imagers) to their complementary target sequence (docking site) attached to a molecule of interest. The transient binding of imager strands is detected as 'blinking' in an intensity versus time trace. DNA-PAINT has a few unique advantages. First, the blinking kinetics (on and off rates) can be tuned over a wide range, by altering the length and sequence of the imager strands or buffer conditions, making the method compatible with different sample conditions. Second, repetitive binding with different imager strands makes the target 'non-bleachable', allowing for the collection of a large number of high-quality and high-precision blinking events and for high-sensitivity imaging on single-molecule targets with discrete molecular resolution (<5 nm). Finally, in combination with orthogonal sequence labels, DNA-PAINT can be multiplexed by imaging with up to dozens of molecular species (Exchange-PAINT).

1.4.1 Fingerprinting via DNA PAINT

DNA-PAINT uses repetitive binding between designed docking and imager DNA strands to allow for imaging with molecular-level resolution (**Box 1.3**). This method provides a promising way to fingerprint proteins on the level of single molecules. A simple approach to characterize proteins could involve amino acid counting using quantitative DNA-PAINT (qPAINT).[44] In this technique, the total blinking rate of a region of interest is measured, which linearly reflects the number of molecular targets in the region. It has been proposed that high-efficiency DNA labeling of specific amino acids (**Figure 1.3**a) followed by qPAINT could lead to single-molecule protein fingerprinting of intact proteins (**Figure 1.3**b).[45]

The recent development of DNA-PAINT has allowed discrete molecular imaging (DMI) of individual molecular targets with spatial resolution below 5 nm.[43] Therefore, protein identification by fingerprinting of amino acids along an extended protein backbone is a possibility. DMI was achieved by combining a systematic analysis and optimization of the DNA-PAINT super-resolution workflow and a high-accuracy (<1 nm) drift correction method. To effectively unfold and extend the protein backbone, N- and C-terminal-specific modifications should be used to attach surface and microbead anchors. The protein can then be subjected to mechanical or elec-

tromagnetic extension force (**Figure 1.3**c). Proposals to combine protein extension methods with high-resolution DMI45 indicate that, with lysine labeling alone and 5-nm effective imaging resolution, more than 50% of the human proteome could be uniquely identified, even with up to 20% amino acid imaging error. Labeling lysine and cysteine would allow coverage of the proteome to increase to more than 75%.

Protein fingerprinting using DNA-PAINT single-molecule imaging combines the ultra-high imaging resolution and quantitative capacity of this technique and the inherent throughput of wide imaging-based methods. qPAINT can produce signals linearly (with <5% deviation), based on the amino acid composition of a particular protein. The proposed methods will be particularly useful for global proteomics analysis of complex protein mixtures and post-translational modification patterns as well as combinatorial analysis of PTM patterns at the single-molecule level.

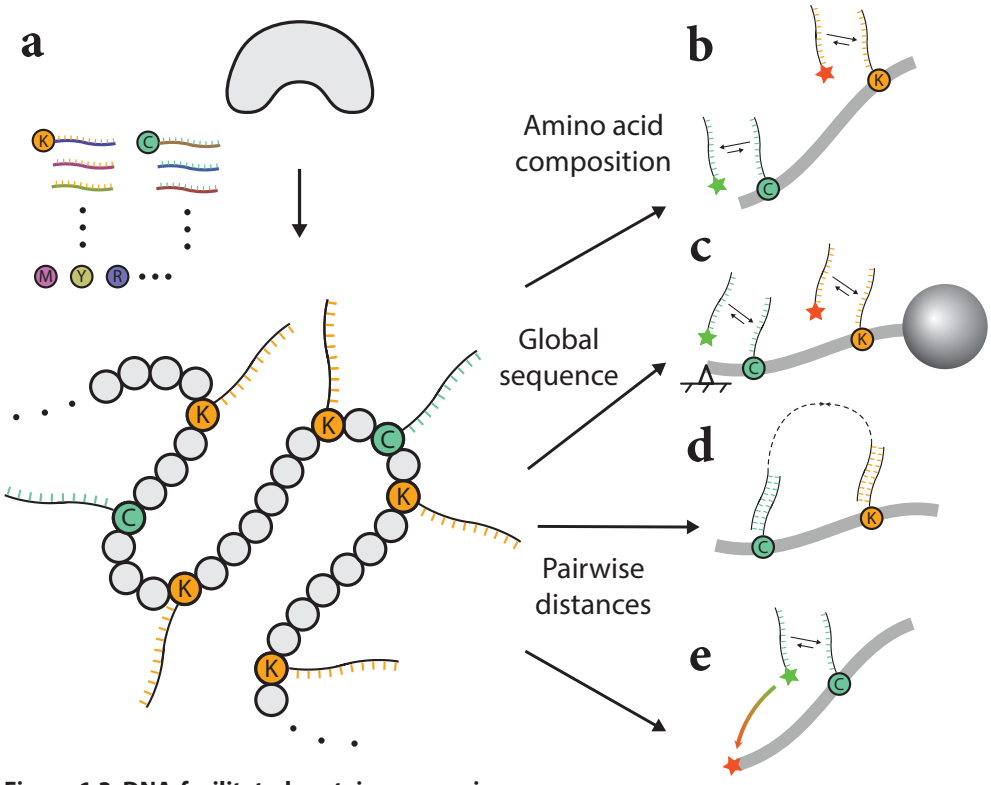


Figure 1.3: DNA-facilitated protein sequencing.
(a) Schematic of specific amino acid labeling on a denatured protein with DNA strands. Each DNA strand contains a barcode for the specific amino acid and (optionally) a UMI. **(b-e)**, Various readout strategies of DNA-labeled samples for protein identification. **(b)**, Protein kinetic fingerprinting using qPAINT. **(c)**, Protein linear barcoding using molecular-resolution DNA-PAINT. **(d)**, DNA proximity recording. **(e)**, Protein structural fingerprinting using FRET-X.

1.4.2 DNA proximity recording

An alternative method for DNA-based protein identification attaches DNA probes to specific amino acids on a protein and uses enzymatic DNA amplification between nearby probes to generate DNA 'records' that vary in length and abundance according to pairwise distances within a protein[46], as exemplified by autocycling proximity recording (APR)[47] (**Figure 1.3**d). The distribution of the lengths of these molecular records is then analyzed to decode the pairwise distance between two DNA tags. It is possible to use unique molecular identifier (UMI) barcoding and repetitive enzymatic recording, such that each lysine and cysteine residue can be studied and used to construct a pairwise distance map, allowing for single-molecule protein identification.[48,49] DNA proximity recording takes advantage of high-throughput next-generation DNA sequencing methods for efficient protein fingerprinting analysis and will be useful for the analysis of both purified proteins and complex protein mixtures.

1.4.3 Protein fingerprinting using FRET

A different approach that allows for global pairwise distance measurements combines DNA technology with single-molecule Förster resonance energy transfer (FRET).[50] The current state of the art for single-molecule FRET analysis allows only one or two FRET pairs to be probed at a time[51], and new high-resolution FRET using transient binding between DNA tags allows for one FRET pair to be probed at a time while many probes are collectively present on a single protein.[50] Similarly to the approaches described above, specific amino acids (for example, lysine, cysteine, etc.) required for fingerprinting have to be labeled with a set of different DNA docking strands. Furthermore, a fixed position on the protein (either the N or C terminus) is labeled with the acceptor fluorophore. Only a single FRET pair forms at a time using DNA strands that are complementary to only a single docking strand. Measurements are then repeated to probe the remaining docking strands and thus the amino acids. The output of this approach is a FRET histogram containing information on the position (referred to as FRET fingerprint) of each detected amino acid relative to one of the reference points. This information is compared to a database consisting of predicted FRET fingerprints, allowing for identification of the protein species (**Figure 1.3**e). The proposed high-resolution FRET approach (named FRET using DNA eXchange, or FRET X) benefits from the immobilization of protein molecules, allowing users to probe each protein multiple times to obtain fingerprints with high resolution. FRET X is a particularly promising tool for targeted proteomics or proteoform analysis as it is able to distinguish small structural changes.

1.5 Biological and solid-state anopores

Since its first demonstration as a single-biomolecule sensor[52], nanopore sensing has dramatically advanced, ultimately achieving the goal of single-molecule DNA sequencing[53]. Many of the nanopore sequencing applications thus far have materialized using an ultra-small device[54] that features vast arrays of biological nanopores, each coupled to its own current amplifier, allowing readout of hundreds of DNA strands simultaneously. Owing primarily to the long read lengths and portability capabilities of this technology, nanopore-based DNA and direct RNA sequencing have become key players in the sequencing field. Nanopore sensing involves drawing biomolecules through the nanopore in a single-file manner. During their passage, the analytes partially block the flow of the ionic current through the pore, leading to time-dependent and sequence-specific electrical signals. Over the past two decades, a variety of synthetic nanopore biosensors have shown substantial progress and are currently used in diverse applications beyond sequencing, including the detection of epigenetic variations and ultra-sensitive detection of mRNA expression[55], among many others.

Just like gel electrophoresis, nanopores may serve as a generic tool to analyze biomolecules. Therefore, as nanopore-based DNA sequencing continues to advance, this technique is poised to extend to proteins, metabolites and other analytes. But despite the remarkable advances in DNA and RNA sequencing, nanopore-based protein sensing is still in its infancy, facing challenges unique to proteins and proteomics. In particular, proteins span a large range of sizes and have a stable three-dimensional folded structure. In contrast to nucleic acids, the backbones of peptides are not naturally charged, complicating the possibility of single-file electrokinetic threading into nanopores. In addition, proteins are composed of combinations of 20 different amino acids instead of 4 nucleobases, further complicating the task of relating the ionic current signals to the amino acid sequence.

While substantial progress in nanopore-based protein sensing has already been made, the development of full-protein sequencers and single-protein identification based on nanopores remains a topic of intense focus. Here we elaborate on three of the principal directions in this field (**Figure 1.4**): (1) single-file threading and direct sensing of the sequence of a polypeptide's amino acids, analogous to the nanopore DNA sequencing principle—in this approach, translocation of either full-length proteins or shorter polypeptide digests of proteins may be targeted; (2) protein identification methods based on sensing unique fingerprints in linearized proteins, without de novo amino acid sequencing; and (3) identification of folded proteins on the basis of specific patterns in their current blockade while in the nanopore. In the following sections, we provide short overviews of the current state of these approaches and refer to additional methods.

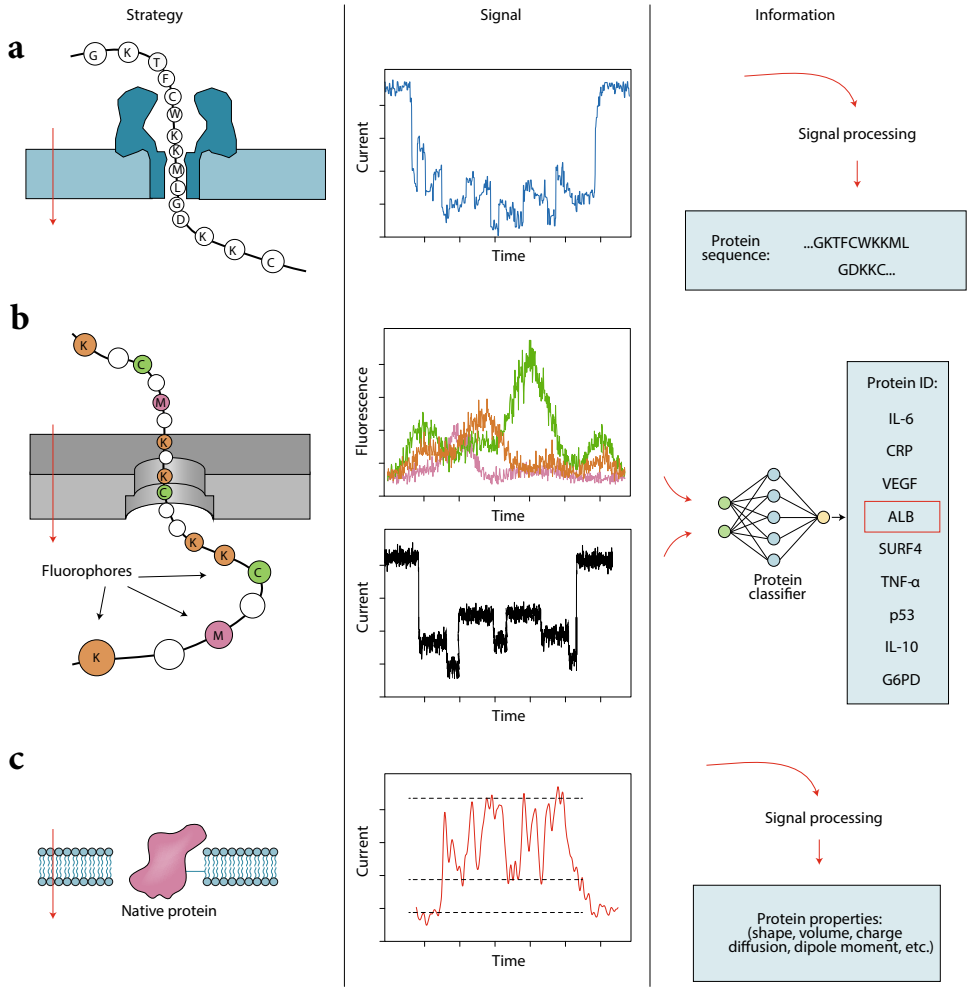


Figure 1.4: Three strategies of nanopore-based protein sequencing and sensing.
In all cases, a voltage bias is applied across an insulating membrane (left panels) and the analytes translocate through the nanopore from top to bottom (red arrows). **(a)** Reading unlabeled proteins or peptides using a biological nanopore. **(b)** Identification of whole proteins or peptides by fingerprinting with deep learning algorithms. Residue-specific fluorescent labels (for example, at lysine, cysteine and methionine) can be used to fingerprint proteins and peptides alongside electrical current sensing. **(c)** Identification of folded proteins using lipid tethering. Other possible tethers include DNA carriers, DNA origami anchors and plasmonic trapping.

1.5.1 Reading the amino acid sequence of linearized peptides

In this proposed approach, a single protein or peptide is linearized and threaded through a nanopore and the resulting ionic current is interpreted to yield an amino acid sequence (**Figure 1.4**a). All-atom molecular dynamics simulations using the α-hemolysin pores have demonstrated a global correlation between the volume of an amino acid and the current blockade in homopolymers.[56] Computationally efficient predictions using coarse-grained models have also performed well in comparison to all-atom molecular dynamics simulations for both solid-state and biological pores.[57]

Discrimination among peptides differing by one amino acid (alanine to glutamate substitution) has been demonstrated using engineered fragaceatoxin C (FraC) nanopores.[58] Moreover, single-amino acid differences within short polyarginine peptides were resolved with superb resolution, using the aerolysin protein pore in its wild-type conformation.[59] Combining molecular dynamics simulations and single-channel experiments, Cao et al. rationally introduced specific point mutations in aerolysin to fine-tune the charge and diameter of the pore, which enhanced its sensitivity and selectivity as showcased experimentally using DNA and peptides.[60] Notably, protein pore sensors were used for the analysis of bodily fluids (blood, sweat, etc.), indicating a substantial potential for applications in diagnostics.[61] As an alternative to nanopore sequencing of intact polypeptide chains, smaller digested fragments can also be analyzed, allowing for detection of minute differences in amino acid composition.[62] Even post-translational modifications can be detected, including individual phosphorylation and glycosylation modifications, using the FraC protein pore.[63]

An essential step in the development of nanopore-based DNA sequencing came with the application of an enzymatic stepping motor (for example, a helicase) that facilitated nucleotide-by-nucleotide progression of the DNA through the nanopore. A similar system is being pursued for single-molecule protein sequencing: molecular motors of the type II secretion system (SecY)[64] and the AAA family (ClpX)[65] are known to unfold and pull protein substrates through pores in an ATP-dependent manner. Nivala et al.[66,67] used ClpXP (or ClpX alone) to unfold and translocate a multidomain fusion protein through the α-hemolysin pore using energy derived from ATP hydrolysis. In this approach, the motor is at the exit of the nanopore and the step size of translocation is therefore dependent on stable structural motifs that resist translocation, rather than being controlled by the enzyme. This approach is currently being expanded by several groups who conjugated ClpXP covalently to α-hemolysin at the entrance of the nanopore to form a combination sensor and substrate delivery machine. The Maglia laboratory genetically introduced a nanopore directly into an archaeal proteasome and found that assisted transport across the nanopore was not influenced by the unfolding of the protein. These nanoscale constructs would also allow a 'chop-and-drop' approach in which single proteins are recognized by their pattern of peptide fragments as they are sequentially cleaved by the peptidase above the nanopore.[68] Knyazev et al. introduced a protein-secreting ATPase as an additional natural choice for a potential peptide-translocating motor.[69,70] Other proteins have the potential to control protein translocation through nanopores, beyond secretases and unfoldases, including chaperones (Hsp70), via processes resembling protein translocation into the mitochondrial matrix.[71] Recently, Rodriguez-Larrea's group

has discussed how protein refolding at the entry and exit compartments can oppose and promote protein translocation, respectively[72,73], and the use of deep learning networks to analyze raw ionic current signals for accurate classification of single point mutations in a translocating protein.[74] In addition, Cardozo et al. built a library of approximately 20 proteins that are orthogonally barcoded with an intrinsic peptide sequence and successfully read them with nanopore sensors.[75]

1.5.2 Fingerprinting linearized proteins

Accurate quantification of different protein species in the proteome with single--molecule resolution would in itself be an achievement of great importance. This can be realized through single-molecule fingerprinting, that is, through the identification of individual protein molecules on the basis of prior knowledge of their amino acid sequences or specific signal patterns, recognized by machine learning[8,76,77] (**Figure 1.4**b). To this end, several nanopore approaches have been pursued: Restrepo-Pérez et al.[78] established a fingerprinting approach using six chemical tags, which were placed on a dipolar peptide.[79] Additionally, Wang et al. reported the ability to distinguish individual lysine and cysteine residues in short polypeptides through specific coupling to fluorescent tags while using a solid-state nanopore with low fluorescence background.[80] In all these approaches, separating the proteins by mass before single-molecule sensing may have greatly facilitated the identification of proteins in complex samples containing many different proteins.[81]

Nanopore protein fingerprinting can make extensive use of advanced deep learning artificial intelligence (AI) strategies to identify patterns in noisy signals. Ohayon et al. recently showed computationally that more than 95% of the proteins in the human proteome can be identified with high confidence when labeling three amino acids (lysine, cysteine and methionine) and threading them linearly through a solid-state nanopore.[77] These simulations predict that even partial labeling of proteins would be sufficient to achieve a high degree of accurate whole-proteome identification, owing to the ability of AI functions to correctly recognize partial protein patterns. This identification method involves the incorporation of sub-wavelength light localization in the proximity of the nanopore using plasmonic nanostructures.[82] The work in this field benefits from recent advances in nanofabrication and nanopatterning technologies allowing for the formation of complex metallic nanostructures to localize fluorescence through plasmon resonance.[83]

1.5.3 Characterization and identification of folded proteins

Thus far, nanopores have been successfully used to detect specific sets of folded proteins and protein oligomers[84] (**Figure 1.4**c) such as large globular proteins, various cytokines and even low-molecular-weight proteins such as ubiquitin. Holding proteins in their folded state inside the nanopore for sufficiently long periods of time is a key requirement. Early studies have shown that globular proteins in the molecular weight range of roughly 5 to 50 kDa can only be detected for a few tens of microseconds or

less85, which is too short for characterization. Several approaches to overcome this challenge have been devised. A lipid bilayer coating of a solid-state nanopore can be used to tether the proteins for extended periods of time.[86] Lipid-tethered proteins[86] and, more recently, freely diffusing proteins (using a higher-bandwidth sensing system)[87] have been characterized with respect to their size, shape, charge, dipole and rotational diffusion coefficient.[88] Various strategies are being pursued to 'trap' proteins in a nanopore. One such strategy is to use plasmonics to hold a protein in a nanopore for seconds or even minutes.[89,90] More recently, single proteins have been held at the nanopore's most sensitive region for minutes to hours using the nanopore electro-osmotic trap (NEOtrap), which exploits strong electro-osmotic water flows created in situ by a charged, permeable object, such as a DNA origami structure.[91] Another approach for slowing down the translocation of proteins involves the use of nanopores smaller than those in earlier studies to increase the hydrodynamic drag, thus resulting in longer translocation dwell times that are easier to measure.[92,93] In addition, high-bandwidth measurements can resolve differential size and conformational flexibility between and within folded proteins.[92-96] Biological nanopores with a diameter of 5.5 or 10 nm[97] can also be used to measure folded proteins, including protein conformations[98] and post-translational modifications[99] such as ubiquitination. Lastly, Aramesh et al.[100] used a combination of atomic-force microscopy and nanopore technology to carry out the first steps of nanopore sensing directly inside cells. Altogether, the detection, identification and sequencing of proteins using single-nanopore approaches has become a highly active, thriving research field, with great potential to revolutionize proteomics, medical diagnostics and also the fundamental biosciences.

1.6 Chemistry for next-generation proteomics technologies

Single-molecule protein fingerprinting has underlined the need for innovative approaches to attach various functional groups to peptides, such as fluorescent moieties. A high degree of chemical specificity is required to avoid downstream misidentification of amino acids, which could lead to sequencing errors. Chemists are making headway on a suite of selective and high-yield methods for labeling specific amino acid side chains, amino acid termini and post-translational modifications with minimal cross-reactivity (**Box 1.4**).

Labeling stability and efficiency are paramount to the success of sequencing technologies but are also a challenge. First, modification of most or all individual residues of one amino acid type is desired for explicit identification of a peptide sequence, which requires selective and highly efficient reactions. Second, error-free sequence prediction requires multiple chemical labels, but the stability of the chemical labels has been an issue in some sequencing techniques. Such issues have been best characterized for fluorosequencing (**Box 1.4**).

For many of the sequencing techniques, amino acids must be labeled with a chemical tag to allow for differentiation between them. While it is theoretically possible to obtain broad coverage of the proteome with labeling for a minimal set of amino acids, specific identification of peptides and broader sequence coverage

Box 1.4: Chemistry concepts in protein sequencing

Labeling efficiency and stability. The challenges in labeling efficiency and stability are well characterized in fluorosequencing, which uses harsh conditions (including neat trifluoroacetic acid) that can lead to reversal of maleimide labeling of cysteine residues. To circumvent this reversal, fluorosequencing instead uses iodoacetamide chemistry, which generates a more stable bond. Another point of complexity is that full conversion is dictated by the solvent accessibility of the targeted amino acid side chains, which can influence labeling efficiency. However, modeling suggests that labeling efficiencies and stabilities substantially below 100% can be compensated for computationally, at least to some degree, during the reference database matching process8.

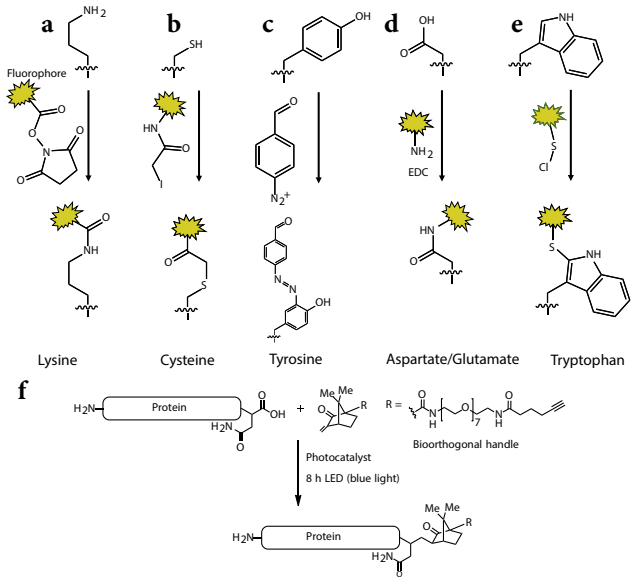


Figure B1.2: Chemistry concepts in protein sequencing.
(a) Lysine labeling with NHS esters. **(b)** Cysteine labeling with iodoacetamide reactive groups. **(c)** Strategies for labeling the phenol ring of tyrosine. **(d)** Aspartate/glutamate labeling. **(e)** Tryptophan labeling with sulfenyl chlorides. **(f)**, C-terminal derivatization through mono-alkylation of the insulin A chain (yield 41%).

Labeling side chains. The most widely accessible labels are those that target lysine using NHS esters and cysteine using maleimide and iodoacetamide reactive groups. Additionally, the phenol ring of tyrosine can be labeled using benzyl diazo groups[122]; however, the attachment of fluorescent molecules generally requires a two-step labeling procedure owing to the cross-reactivity with fluorescent molecules. Another robust bioconjugation method to selectively target tyrosine side chains is an ene-like reaction with cyclic diazodicaboxamides in aqueous buffer[123]. Carboxylic acids have also been labeled on peptides, but, owing to the similar reactivities of aspartate, glutamate and the C terminus, this has primarily been used on synthetic peptides. The method makes use of a standard technique (EDC coupling) for binding amines covalently to carboxylic acids, forming an amide bond. In a recently reported promising bioconjugation approach, light-activated 2,5-disubstituted tetrazoles have been shown to be able to convert glutamate and aspartate residues with high yield[124]. Finally, tryptophan can be labeled at the C2 position using sulfenyl

Box 1.4: Chemistry concepts in protein sequencing

chlorides. However, this comes with the limitations that the reaction is extremely water sensitive and the reactive group must be made in situ.[101] There are also promising new methods that allow for chemical modification of other amino acids. Methionine, for example, can either be elegantly labeled with hypervalent iodine reagents[125] or by the use of urea-derived oxaziridines.[126,127] Recently, a histidine-selective conjugation methodology was reported where thiophosphorodichloridates selectively form a covalent bond with the histidine residues in proteins[128].

C-terminal labeling. Labeling of the C terminus is a challenge in that the C terminus must be differentiated from aspartate and glutamate, which carry the same functionality. A photoredox reaction on the C terminus of peptides and proteins entailing decarboxylation of the C-terminal carboxylic acid followed by an alkylation step with a Michael acceptor has recently been reported[129]. Because of their higher oxidation potential, the carboxylates of internal amino acid chains are less prone to this modification, making the method highly site selective. This technique has been applied for a variety of peptide substrates as well as for C-terminus-specific alkylation of human insulin chain A.

N-terminal labeling. Several methods exist for modifying the N terminus.[130] Classic approaches such as reductive amination with aldehydes or acylation with NHS esters, which rely on pH control to increase selectivity, are not sufficiently specific. Other strategies involve the side chain of the N-terminal amino acid. Native chemical ligation[131] or condensation reactions with aldehydes[132] could be used to label N-terminal cysteine, serine, threonine or tryptophan residues. Furthermore, oxidizing N-terminal serine or threonine residues to their corresponding aldehydes allows oxime conjugation with hydrazides or hydroxylamines.[133] A more general methodology has also emerged where the N-terminal amine condenses with 2-pyridinecarboxaldehyde, forming an imine structure that further reacts via cyclization with the nearby amide nitrogen of the second amino acid to form a stable imidazolidinone product.[134] This reaction has recently been shown to be useful for single-molecule peptide sequencing as a method for the immobilization of peptides onto a solid-phase resin, multiple chemical derivatization steps without purification and subsequent traceless release before fluorosequencing.[10]

Post-translational modifications. As an example of elimination replacement chemistries, phospho-serine and phospho-threonine residues can be labeled by β-elimination followed by Michael addition (BEMA). In MS-based phosphoproteomics, this is used to introduce an additional trypsin cleavage site at the phosphorylated amino acid[135], while at the single-molecule level it can be used to site specifically attach a fluorescent label. Such an approach has been established for the Edman degradation described above.[9]

Protein glycosylation can be complex, featuring many different types of monomeric units bound in possibly branching polymer structures. Full structural characterization often requires derivatization and is done on glycans that are released from the protein. Therefore, schemes for understanding site-specific and simple glycosylation events should be the current focus. N-glycan-anchoring asparagine residues can be converted to aspartate by glycan removal with PNGase F for practically all protein sequencing approaches, reducing complexity in the detection of this modification. Another possibility to introduce site-selective labels is the incorporation of azide-tagged glycans, achieved by adding modified carbohydrates to the cell medium.[136] In other detection schemes, the location of a modification could also be inferred using glycan-specific reporter molecules such as lectins, engineered proteins or aptamers.[137]

require a larger suite of labels. Overall, there are 12 distinct side chain types in peptides, ranging from those for highly reactive amino acids such as lysine and cysteine to functional groups that are more challenging to modify, such as amides (glutamine and asparagine) and alkanes (alanine, glycine, isoleucine, leucine, proline and valine). There are a large number of methods to label amino acids; however, some chemistries do not provide sufficiently stable bonds for some single-molecule sequencing approaches. Thus far, labeling for only eight amino acids (lysine, cysteine, glutamate, aspartate, tyrosine, tryptophan, histidine and arginine) has been shown to be stable, selective and reactive enough for the single-molecule fluorosequencing approach.[9,101] Research is ongoing to test a wide variety of other labeling conditions to cover all of the proteinogenic amino acids (**Box 1.4**).

Chemical modification of protein termini is highly desired for several sequencing techniques such as the fluorosequencing, nanopore and DNA-PAINT approaches where end labeling or ligation is required (**Figure 1.2–1.4**). The terminus provides an attachment point for surface immobilization and can offer a simple way to remove excess chemical reagents during procedures that require multiple labeling steps. Two terminus-specific methods have shown great promise for single-molecule sequencing, C-terminal labeling using decarboxylative alkylation and modification of the N terminus with 2-pyridinecarboxaldehyde (**Box 1.4**).

The long-term goal of characterizing proteoforms requires methods to detect and differentiate post-translational modifications. Such modifications can be recognized by MS through the mass shifts they cause on a protein, peptide and their fragments[102,103], and databases of the expected mass shifts such as Unimod are used to support identification.[104] However, these databases show that there can be substantial overlap between post-translational modifications of the same or similar mass, suggesting that orthogonal methods are needed. Single-molecule protein sequencing methods rely on either site-specific labeling or elimination and replacement chemistries (**Box 1.4**).

1.7 Discussion: a spectrum of opportunities

An emerging landscape of single-molecule protein sequencing and fingerprinting technologies is unfolding with the promise of resolving the full proteome of single cells with single-protein resolution, opening up unprecedented opportunities in basic science and in medical diagnostics. For example, resolving the cellular and spatial heterogeneity in tissue proteomes with integration of other layers of the central dogma could open new research avenues from embryonic development to cancer research. Diagnostics could benefit from the ultimate single-molecule resolution by resolving very low amounts of protein in bodily samples. The detection of rare proteins with copy numbers as low as one or a few may uncover new molecular regulatory networks within cells. Some of the emerging technologies described here are still at early proof-of-concept stages in development, whereas others, including sequencing by Edman degradation and nanopore sequencing technologies, have already attracted industry funding. Additional single-molecule approaches are also promoted by commercial entities but are outside the scope of this Perspective.

A real-world application of a technology that is not MS or antibody based for whole-proteome characterization is yet to be achieved. Meanwhile, MS will continue to improve in its capacity to support single-ion detection[22] and ultimately single-cell proteomics.[105] Similarly, antibody-based methods such as immunoassays that rely on specific antigen–antibody interactions have served as the standard methods for protein identification and quantification for the last few decades. Specifically, antibody-based methods have enabled multiplexed protein analysis with improvements of several orders of magnitude in sensitivity over conventional immunoassays. A notable example is the Single Molecule Array technology (Simoa)[106] by Quanterix, a digital immunoassay based on single-molecule counting used for the analysis of minute biological samples with up to sub-femtomolar sensitivity.[107] The coronavirus disease 2019 (COVID-19) pandemic has accelerated the development of high-throughput serological tests of clinical samples using Simoa[108] based on ultra-small blood samples. These sensitive antibody-based methods will continue to have a main role in molecular diagnostics, in parallel with other single-molecule techniques that will permit comprehensive proteoform inference or differentiation.

The emerging landscape of alternative protein sequencing and fingerprinting technologies in **Figure 1.1** could one day help to sequence human proteoforms in a more complete way. High-throughput Edman degradation could pair with bottom–up MS strategies to alleviate current limitations on sequence coverage (**Box 1.1**). These bottom–up methods could benefit from nanopore sequencing and DNA fluorescence-based methods that aim for long-read sequencing and structural fingerprinting of whole proteins. Integration of both existing and emerging technologies promises to iteratively reveal an atlas of full-length proteoforms, which could itself assist these up-and-coming technologies to infer what cannot be directly measured in terms of protein primary sequence and structure.

An additional far-reaching goal for single-molecule proteomics lies in the analysis of protein–protein interactions. A map covering a wide range of proteoforms and their interactions is an unmet milestone needed to uncover protein networks in normal tissues and in disease. Bottom–up MS-based approaches, such as cross-linking[109,110] and affinity purification, are implemented to identify physical[111] and proximal[112] interactions. However, these techniques present either biochemical or sample processing yield limitations, as a result of challenges such as over-representation of intra-protein cross-linking, loss of protein–protein interactions upon solubilization and limitations inherent to MS analysis, hindering single-cell interactome analysis. Currently, single-molecule analysis of protein–protein interactions has not reached mainstream proteomics, which is even more true for single-cell interactomics. Achieving these goals would be of great interest in accurately defining, for example, protein organization within highly dynamic membraneless organelles[113], such as in resolving protein condensates and spatial and temporal organization at a single-organelle or single-cell scale, and would provide an unprecedented resolution for the organization of protein–protein interactions.

1.7.1 Challenges for next-generation protein sequencing

Two grand challenges await technological innovations and need to be addressed to enable the high-throughput sequencing of complex protein mixtures. First, there is no method to amplify the copy number of proteins similar to the methods used for nucleic acids. New techniques focus on characterizing individual proteins. The aim is to sequence proteomes starting from a low number of cells or minute samples that often contain just a few or single copies of specific proteins. This presents a second problem: a single eukaryotic cell contains billions of proteins. While the presented methods may enable single-molecule protein identification, they must reach an extremely high sensing throughput to profile all proteins in the cell and permit whole-cell analysis on a reasonable timescale. These two seemingly contractionary requirements (single-protein molecule sensitivity and extremely high throughput) present one of the main challenges to the field, and striking an optimal balance between them will be key for all the technologies discussed. Of the orthogonal methods presented, nanopores, fluorosequencing and protein linear barcoding using DNA-PAINT, to name a few, stand a chance to eventually measure billions of proteins within a few hours.

Emerging technologies will be evaluated in terms of their sensitivity, proteome coverage (fraction of whole proteins in the sample covered), sequence coverage (average fraction of protein sequences covered), peptide read length (mean number of amino acids in a single read), accuracy (error in calling an amino acid or in identifying a whole protein), cost and throughput. In this regard, additional research and validation will be required to demonstrate the benefits of these orthogonal technologies. The formation of a dedicated global academic/scientific community in single-protein sequencing may catalyze further development and implementation of these technologies for more widespread use. Multidisciplinary meetings that bring together experts in chemistry, physics, engineering, computer sciences and other relevant areas of expertise (for example, pathologists and clinicians) with a clear vision of the most relevant problems and unmet needs will need to be embraced.

1.8 Thesis Outline

In this thesis the advances in the emerging landscape of single-molecule fluorescence approaches to sequence proteins are described.

In **Chapter 2**, we demonstrate proof-of-concept experiments of the first single-molecule protein fingerprinting machine. We repurposed the naturally occurring nanomotor, ClpXP to detect FRET events between donor-labeled ClpP and acceptor labeled protein substrates. We could detect two different residues in a single peptide read and determine the order of these residues relative to the C-terminus.

The remainder of this thesis mainly focuses on the development of another, new high-resolution protein fingerprinting technique using single-molecule FRET. In **Chapter 3**, we introduce a novel FRET technique that utilizes DNA nanotechnology to probe multiple FRET pairs in a single nanoscopic object. We name this approach FRET X for FRET via DNA eXchange. We evaluate the precision of FRET X on several DNA model substrates and show that FRET X can resolve FRET efficiencies with sub-nanometer . The use of FRET X for single-molecule protein fingerprinting is demonstrated in **Chapters 4** and **5**.

In **Chapter 4** we use FRET X to localize a subset of amino acids within a protein structure. Our simulations demonstrate that with a FRET fingerprint for cysteines, lysines and arginines > 95% of the human proteome can be identified when probed in a complex mixture of >300 proteins. Furthermore, we provide proof-of-concept experimental data that demonstrate the ability to localize cysteines in different model peptides and thereby distinguish them.

In **Chapter 5**, we explore the capacity of FRET X to obtain fingerprints of full-length protein substrates. We use a biomedically relevant protein, alpha-synuclein, to demonstrate the ability to localize multiple cysteines within the protein structure. Furthermore, we demonstrate that also on more complex protein substrates that are folded into a globular conformation, FRET X can reliably and reproducibly obtain a fingerprint.

Our FRET X protein fingerprinting approach relies on the binding and unbinding of fluorescent labeled DNA probes. However, the binding rate of DNA is relatively slow which necessitates long acquisition times for FRET X measurements. To overcome this limitation, in **Chapter 6** we utilize the Argonaute protein and demonstrate that it can accelerate the DNA binding rate by an order of magnitude.

Finally, in **Chapter 7**, I discuss some of the most common challenges and contemplate about future directions for the field of single-molecule protein sequencing.

Happy Reading!

1.9 References

1 Breuza, L. et al. The UniProtKB guide to the human proteome. Database 2016, bav120 (2016).

2 Smith, L. M. et al. Proteoform: a single term describing protein complexity. Nat. Methods 10, 186–187 (2013)

3 Seattle Times Business Staff. Seattle biotech startup Nautilus to get $350 million, stock listing in blank-check deal. The Seattle Times https://www.seattletimes.com/business/seattle-biotech-startup--nautilus-to-get-350-million-stock-listing-in-blank-check-deal/ (8 February 2021).

4 Reuters Staff. Protein sequencing firm Quantum-Si to go public via $1.46 billion SPAC merger. Reuters https://www.reuters.com/article/us-quantum-si-m-a-highcape-capital-idUSKBN2AI1HT (18 February 2021).

5 Cohen, L. & Walt, D. R. Single-molecule arrays for protein and nucleic acid analysis. Annu. Rev. Anal. Chem. 10, 345–363 (2017).

6 Aggarwal, V. & Ha, T. Single-molecule fluorescence microscopy of native macromolecular complexes. Curr. Opin. Struct. Biol. 41, 225–232 (2016).

7 Edman, P. A method for the determination of the amino acid sequence in peptides. Arch. Biochem. 22, 475–476 (1949).

8 Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. PLoS Comput. Biol. 11, 1076–1082 (2015).

9 Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. Nat. Biotechnol. 36, 1076–1082 (2018).

10 Howard, C. J. et al. Solid-phase peptide capture and release for bulk and single-molecule proteomics. ACS Chem. Biol. 15, 1401–1407 (2020).

11 Miclotte, G., Martens, K. & Fostier, J. Computational assessment of the feasibility of protonation-based protein sequencing. PLoS ONE 15, e0238625 (2020).

12 Tullman, J., Marino, J. P. & Kelman, Z. Leveraging nature's biomolecular designs in next-generation protein sequencing reagent development. Appl. Microbiol. Biotechnol. 104, 7261–7271 (2020).

13 Rodriques, S. G., Marblestone, A. H. & Boyden, E. S. A theoretical analysis of single molecule protein sequencing via weak binding spectra. PLoS ONE 14, e0212868 (2019).

14 Tullman, J., Callahan, N., Ellington, B., Kelman, Z. & Marino, J. P. Engineering ClpS for selective and enhanced N-terminal amino acid binding. Appl. Microbiol. Biotechnol. 103, 2621–2633 (2019).

15 Smith, R. D., Cheng, X., Brace, J. E., Hofstadler, S. A. & Anderson, G. A. Trapping, detection and reaction of very large single molecular ions by mass spectrometry. Nature 369, 137–139 (1994).

16 Keifer, D. Z. & Jarrold, M. F. Single-molecule mass spectrometry. Mass Spectrom. Rev. 36, 715–733 (2017).

17 Rose, R. J., Damoc, E., Denisov, E., Makarov, A. & Heck, A. J. High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. Nat. Methods 9, 1084–1086 (2012).

18 Makarov, A. & Denisov, E. Dynamics of ions of intact proteins in the Orbitrap mass analyzer. J. Am. Soc. Mass Spectr. 20, 1486–1495 (2009).

19 Kafader, J. O. et al. Measurement of individual ions sharply increases the resolution of Orbitrap mass spectra of proteins. Anal. Chem. 91, 2776–2783 (2019).

20 Kafader, J. O. et al. Multiplexed mass spectrometry of individual ions improves measurement of proteoforms and their complexes. Nat. Methods 17, 391–394 (2020).

21 Wörner, T. P. et al. Resolving heterogeneous macromolecular assemblies by Orbitrap-based single-particle charge detection mass spectrometry. Nat. Methods 17, 395–398 (2020).

22 Kafader, J. O. et al. Individual ion mass spectrometry enhances the sensitivity and sequence coverage of top down mass spectrometry. J. Proteome Res. 19, 1346–1350 (2020).

23 Smith, L. et al. The human proteoform project: a plan to define the human proteome. Preprint at Preprints doi:10.20944/preprints202010.0368.v1 (2020).

24 Ekinci, K. L., Huang, X. M. H. & Roukes, M. L. Ultrasensitive nanoelectromechanical mass detection. Appl. Phys. Lett. 84, 4469–4471 (2004).

25 Hanay, M. S. et al. Single-protein nanomechanical mass spectrometry in real time. Nat. Nanotechnol. 7, 602–608 (2012).

26 Sage, E. et al. Neutral particle mass spectrometry with nanomechanical systems. Nat. Commun. 6, 6482 (2015).

27 Dominguez-Medina, S. et al. Neutral mass spectrometry of virus capsids above 100 megadaltons with nanomechanical resonators. Science 362, 918–922 (2018).

28 Chaste, J. et al. A nanomechanical mass sensor with yoctogram resolution. Nat. Nanotechnol. 7, 301–304 (2012).

29 Hanay, M. S. et al. Inertial imaging with nanomechanical systems. Nat. Nanotechnol. 10, 339–344 (2015).

30 Malvar, O. et al. Mass and stiffness spectrometry of nanoparticles and whole intact bacteria by multimode nanomechanical resonators. Nat. Commun. 7, 13452 (2016).

31 Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. Anal. Chem. 68, 1–8 (1996).

32 El-Faramawy, A., Siu, K. M. & Thomson, B. A. Efficiency of nano-electrospray ionization. J. Am. Soc. Mass Spectr. 16, 1702–1707 (2005).

33 Bush, J. et al. The nanopore mass spectrometer. Rev. Sci. Instrum. 88, 113307 (2017).

34 Maulbetsch, W., Wiener, B., Poole, W., Bush, J. & Stein, D. Preserving the sequence of a biopolymer's monomers as they enter an electrospray mass spectrometer. Phys. Rev. Appl. 6, 054006 (2016).

35 Brodbelt, J. S. Photodissociation mass spectrometry: new tools for characterization of biological molecules. Chem. Soc. Rev. 43, 2757–2783 (2014).

36 Chang, S. et al. Tunnelling readout of hydrogen-bonding-based recognition. Nat. Nanotechnol. 4, 297–301 (2009).

37 Zhao, Y. et al. Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. Nat. Nanotechnol. 9, 466–473 (2014).

38 Ohshiro, T. et al. Detection of post-translational modifications in single peptides using electron tunnelling currents. Nat. Nanotechnol. 9, 835–840 (2014).

39 Zhang, B. et al. Observation of giant conductance fluctuations in a protein. Nano Futures 1, 035002 (2017).

40 Zhang, B. et al. Engineering an enzyme for direct electrical monitoring of activity. ACS Nano 14, 1360–1368 (2020).

41 Seeman, N. C. & Sleiman, H. F. DNA nanotechnology. Nat. Rev. Mater. 3, 17068 (2017).

42 Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with DNA-PAINT. Nat. Protoc. 12, 1198–1228 (2017).

43 Dai, M., Jungmann, R. & Yin, P. Optical imaging of individual biomolecules in densely packed clusters. Nat. Nanotechnol. 11, 798–807 (2016).

44 Jungmann, R. et al. Quantitative super-resolution imaging with qPAINT. Nat. Methods 13, 439–442 (2016).

45 Dai, M. & Yin, P. Methods and compositions relating to super-resolution imaging and modification. US patent 10006917 (2018).

46 Woo, S. & Yin, P. Methods and compositions for protein identification. US patent 10697974 (2020).

47 Schaus, T. E., Woo, S., Xuan, F., Chen, X. & Yin, P. A DNA nanoscope via auto-cycling proximity recording. Nat. Commun. 8, 696 (2017).

48 Kishi, J. Y., Schaus, T. E., Gopalkrishnan, N., Xuan, F. & Yin, P. Programmable autonomous synthesis of single-stranded DNA. Nat. Chem. 10, 155–164 (2018).

49 Gopalkrishnan, N., Punthambaker, S., Schaus, T. E., Church, G. M. & Yin, P. A DNA nanoscope that identifies and precisely localizes over a hundred unique molecular features with nanometer accuracy. Preprint at bioRxiv doi:10.1101/2020.08.27.271072 (2020).

50 Filius, M., Kim, S. H., Severins, I. & Joo, C. High-resolution single-molecule FRET via DNA eXchange (FRET X). Nano Lett. 21, 3295–3301 (2021).

51 Lerner, E. et al. Toward dynamic structural biology: two decades of single-molecule Förster resonance energy transfer. Science 359, eaan1133 (2018)

52 Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl Acad. Sci. USA 93, 13770–13773 (1996).

53 Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. Nat. Biotechnol. 34, 518–524 (2016).

54 Loman, N. J. & Watson, M. Successful test launch for nanopore sequencing. Nat. Methods 12, 303–304 (2015).

55 Rozevsky, Y. et al. Quantification of mRNA expression using single-molecule nanopore sensing. ACS Nano 14, 13964–13974 (2020).

56 Di Muccio, G., Rossini, A. E., Di Marino, D., Zollo, G. & Chinappi, M. Insights into protein sequencing with an α-hemolysin nanopore by atomistic simulations. Sci. Rep. 9, 6440 (2019).

57 Wilson, J., Sarthak, K., Si, W., Gao, L. & Aksimentiev, A. Rapid and accurate determination of nanopore ionic current using a steric exclusion model. ACS Sens. 4, 634–644 (2019).

58 Huang, G., Voet, A. & Maglia, G. FraC nanopores with adjustable diameter identify the mass of opposite-charge peptides with 44 dalton resolution. Nat. Commun. 10, 835 (2019).

59 Piguet, F. et al. Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. Nat. Commun. 9, 966 (2018).

60 Cao, C. et al. Single-molecule sensing of peptides and nucleic acids by engineered aerolysin nanopores. Nat. Commun. 10, 4918 (2019).

61 Galenkamp, N. S., Soskine, M., Hermans, J., Wloka, C. & Maglia, G. Direct electrical quantification of glucose and asparagine from bodily fluids using nanopores. Nat. Commun. 9, 4085 (2018).

62 Ouldali, H. et al. Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. Nat. Biotechnol. 38, 176–181 (2020).

63 Restrepo-Pérez, L., Wong, C. H., Maglia, G., Dekker, C. & Joo, C. Label-free detection of post-translational modifications with a nanopore. Nano Lett. 19, 7957–7964 (2019).

64 Korotkov, K. V., Sandkvist, M. & Hol, W. G. The type II secretion system: biogenesis, molecular architecture and mechanism. Nat. Rev. Microbiol. 10, 336–351 (2012).

65 Olivares, A. O., Baker, T. A. & Sauer, R. T. Mechanical protein unfolding and degradation. Annu. Rev. Physiol. 80, 413–429 (2018).

66 Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. Nat. Biotechnol. 31, 247–250 (2013).

67 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).

68 Zhang, S. et al. Bottom–up fabrication of a multi-component nanopore sensor that unfolds, processes and recognizes single proteins. Preprint at bioRxiv doi:10.1101/2020.12.04.411884 (2020).

69 Sachelaru, I. et al. YidC and SecYEG form a heterotetrameric protein translocation channel. Sci. Rep. 7, 101 (2017).

70 Knyazev, D. G., Kuttner, R., Zimmermann, M., Sobakinskaya, E. & Pohl, P. Driving forces of translocation through bacterial translocon SecYEG. J. Membr. Biol. 251, 329–343 (2018).

71 Backes, S. & Herrmann, J. M. Protein translocation into the intermembrane space and matrix of mitochondria: mechanisms and driving forces. Front. Mol. Biosci. 4, 83 (2017).

72 Feng, J. et al. Transmembrane protein rotaxanes reveal kinetic traps in the refolding of translocated substrates. Commun. Biol. 3, 159 (2020).

73 Rosen, C. B., Bayley, H. & Rodriguez-Larrea, D. Free-energy landscapes of membrane co-translocational protein unfolding. Commun. Biol. 3, 160 (2020).

74 Rodriguez-Larrea, D. Single-aminoacid discrimination in proteins with homogeneous nanopore sensors and neural networks. Biosens. Bioelectron. 180, 113108 (2021).

75 Cardozo, N. et al. Multiplexed direct detection of barcoded protein reporters on a nanopore array. Preprint at bioRxiv doi:10.1101/837542 (2019).

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

77 Ohayon, S., Girsault, A., Nasser, M., Shen-Orr, S. & Meller, A. Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. PLoS Comput. Biol. 15, e1007067 (2019).

78 Restrepo-Pérez, L. et al. Resolving chemical modifications to a single amino acid within a peptide using a biological nanopore. ACS Nano 13, 13668–13676 (2019).

79 Asandei, A. et al. Placement of oppositely charged amino acids at a polypeptide termini determines the voltage-controlled braking of polymer transport through nanometer-scale pores. Sci. Rep. 5, 10419 (2015).

80 Wang, R. et al. Single-molecule discrimination of labeled DNAs and polypeptides using photoluminescent-free TiO2 nanopores. ACS Nano 12, 11648–11656 (2018).

81 Zrehen, A., Ohayon, S., Huttner, D. & Meller, A. On-chip protein separation with single-molecule resolution. Sci. Rep. 10, 15313 (2020).

82 Assad, O. N. et al. Light-enhancing plasmonic-nanopore biosensor for superior single-molecule detection. Adv. Mater. 29, 1605442 (2017).

83 Spitzberg, J. D., Zrehen, A., van Kooten, X. F. & Meller, A. Plasmonic-nanopore biosensors for superior single-molecule detection. Adv. Mater. 31, 1900422 (2019).

84 Houghtaling, J., List, J. & Mayer, M. Nanopore-based, rapid characterization of individual amyloid particles in solution: concepts, challenges, and prospects. Small 14, 1802412 (2018).

85 Plesa, C. et al. Fast translocation of proteins through solid state nanopores. Nano Lett. 13, 658–663 (2013).

86 Yusko, E. C. et al. Controlling protein translocation through nanopores with bio-inspired fluid walls. Nat. Nanotechnol. 6, 253–260 (2011).

87 Houghtaling, J. et al. Estimation of shape, volume, and dipole moment of individual proteins freely transiting a synthetic nanopore. ACS Nano 13, 5231–5242 (2019).

88 Yusko, E. C. et al. Real-time shape approximation and fingerprinting of single proteins using a nanopore. Nat. Nanotechnol. 12, 360–367 (2017).

89 Pang, Y. & Gordon, R. Optical trapping of a single protein. Nano Lett. 12, 402–406 (2012).

90 Verschueren, D., Shi, X. & Dekker, C. Nano-optical tweezing of single proteins in plasmonic nanopores. Small Methods 3, 1800465 (2019).

91 Schmid, S., Stömmer, P., Dietz, H. & Dekker, C. Nanopore electro-osmotic trap for the label-free study of single proteins and their conformations. Preprint at bioRxiv doi:10.1101/2021.03.09.434634 (2021).

92 Larkin, J., Henley, R. Y., Muthukumar, M., Rosenstein, J. K. & Wanunu, M. High-bandwidth protein analysis using solid-state nanopores. Biophys. J. 106, 696–704 (2014).

93 Nir, I., Huttner, D. & Meller, A. Direct sensing and discrimination among ubiquitin and ubiquitin chains using solid-state nanopores. Biophys. J. 108, 2340–2349 (2015).

94 Waduge, P. et al. Nanopore-based measurements of protein size, fluctuations, and conformational changes. ACS Nano 11, 5706–5716 (2017).

95 Varongchayakul, N., Hersey, J. S., Squires, A., Meller, A. & Grinstaff, M. W. A solid-state hard microfluidic-nanopore biosensor with multilayer fluidics and on-chip bioassay/purification chamber. Adv. Funct. Mater. 28, 1804182 (2018).

96 Hu, R. et al. Differential enzyme flexibility probed using solid-state nanopores. ACS Nano 12, 4494–4502 (2018).

97 Huang, G. et al. Electro-osmotic vortices promote the capture of folded proteins by PlyAB nanopores. Nano Lett. 20, 3819–3827 (2020).

98 Soskine, M., Biesemans, A. & Maglia, G. Single-molecule analyte recognition with ClyA nanopores equipped with internal protein adaptors. J. Am. Chem. Soc. 137, 5793–5797 (2015).

99 Wloka, C. et al. Label-free and real-time detection of protein ubiquitination with a biological nanopore. ACS Nano 11, 4387–4394 (2017).

100 Aramesh, M. et al. Localized detection of ions and biomolecules with a force-controlled scanning nanopore microscope. Nat. Nanotechnol. 14, 791–798 (2019).

101 Hernandez, E. T., Swaminathan, J., Marcotte, E. M. & Anslyn, E. V. Solution-phase and solid-phase sequential, selective modification of side chains in KDYWEC and KDYWE as models for usage in single-molecule protein sequencing. New J. Chem. 41, 462–469 (2017).

102 Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat. Methods 14, 513–520 (2017).

103 Zhong, J. et al. Proteoform characterization based on top–down mass spectrometry. Brief. Bioinform. 22, 1729–1750 (2021).

104 Creasy, D. M. & Cottrell, J. S. Unimod: protein modifications for mass spectrometry. Proteomics 4, 1534–1536 (2004).

105 Marx, V. A dream of single-cell proteomics. Nat. Methods 16, 809–812 (2019).

106 Rissin, D. M. et al. Single-molecule enzyme-linked immunosorbent assay detects serum proteins at subfemtomolar concentrations. Nat. Biotechnol. 28, 595–599 (2010).

107 Wu, C., Garden, P. M. & Walt, D. R. Ultrasensitive detection of attomolar protein concentrations by dropcast single molecule assays. J. Am. Chem. Soc. 142, 12314–12323 (2020).

108 Norman, M. et al. Ultrasensitive high-resolution profiling of early seroconversion in patients with COVID-19. Nat. Biomed. Eng. 4, 1180–1187 (2020).

109 Liu, F., Rijkers, D. T., Post, H. & Heck, A. J. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat. Methods 12, 1179–1184 (2015).

110 Iacobucci, C., Götze, M. & Sinz, A. Cross-linking/mass spectrometry to get a closer view on protein interaction networks. Curr. Opin. Biotechnol. 63, 48–53 (2020).

111 Dunham, W. H., Mullin, M. & Gingras, A.-C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. Proteomics 12, 1576–1590 (2012).

112 Gentzel, M., Pardo, M., Subramaniam, S., Stewart, A. F. & Choudhary, J. S. Proteomic navigation using proximity-labeling. Methods 164, 67–72 (2019).

113 Zhao, Y. G. & Zhang, H. Phase separation in membrane biology: the interplay between membrane-bound organelles and membraneless condensates. Dev. Cell 55, 30–44 (2020).

114 Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. 5, 699–711 (2004).

115 Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in top–down proteomics and the analysis of proteoforms. Annu. Rev. Anal. Chem. 9, 499–519 (2016).

116 Samaras, P. et al. ProteomicsDB: a multi-omics and multi-organism resource for life science research. Nucleic Acids Res. 48, D1153–D1163 (2020).

117 Ruggles, K. V. et al. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. Mol. Cell. Proteomics 15, 1060–1071 (2016).

118 Zhu, Y. et al. Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. Nat. Commun. 9, 882 (2018).

119 Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. Genome Biol. 19, 161 (2018).

120 Zhu, Y. et al. Proteomic analysis of single mammalian cells enabled by microfluidic nanodroplet sample preparation and ultrasensitive NanoLC–MS. Angew. Chem. Int. Ed. 57, 12370–12374 (2018).

121 Kelly, R. T. Single-cell proteomics: progress and prospects. Mol. Cell. Proteomics 19, 1739–1748 (2020).

122 Gavrilyuk, J., Ban, H., Nagano, M., Hakamata, W. & Barbas, C. F. Formylbenzene diazonium hexafluorophosphate reagent for tyrosine-selective modification of proteins and the introduction of a bioorthogonal aldehyde. Bioconjugate Chem. 23, 2321–2328 (2012).

123 Ban, H., Gavrilyuk, J. & Barbas, C. F. III Tyrosine bioconjugation through aqueous ene-type reactions: a click-like reaction for tyrosine. J. Am. Chem. Soc. 132, 1523–1525 (2010).

124 Bach, K., Beerkens, B. L., Zanon, P. R. & Hacker, S. M. Light-activatable, 2,5-disubstituted tetrazoles for the proteome-wide profiling of aspartates and glutamates in living bacteria. ACS Cent. Sci. 6, 546–554 (2020).

125 Taylor, M. T., Nelson, J. E., Suero, M. G. & Gaunt, M. J. A protein functionalization platform based on selective reactions at methionine residues. Nature 562, 563–568 (2018).

126 Lin, S. et al. Redox-based reagents for chemoselective methionine bioconjugation. Science 355, 597–602 (2017).

127 Christian, A. H. et al. A physical organic approach to tuning reagents for selective and stable methionine bioconjugation. J. Am. Chem. Soc. 141, 12657–12662 (2019).

128 Jia, S., He, D. & Chang, C. J. Bioinspired thiophosphorodichloridate reagents for chemoselective histidine bioconjugation. J. Am. Chem. Soc. 141, 7294–7301 (2019).

129 Bloom, S. et al. Decarboxylative alkylation for site-selective bioconjugation of native proteins via oxidation potentials. Nat. Chem. 10, 205–211 (2018).

130 Rosen, C. B. & Francis, M. B. Targeting the N terminus for site-selective protein modification. Nat. Chem. Biol. 13, 697–705 (2017).

131 Busch, G. K. et al. Specific N-terminal protein labelling: use of FMDV 3C pro protease and native chemical ligation. Chem. Commun. 29, 3369–3371 (2008).

132 Bandyopadhyay, A., Cambray, S. & Gao, J. Fast and selective labeling of N-terminal cysteines at neutral pH via thiazolidino boronate formation. Chem. Sci. 7, 4589–4593 (2016).

133 Agten, S. M., Dawson, P. E. & Hackeng, T. M. Oxime conjugation in protein chemistry: from carbonyl incorporation to nucleophilic catalysis. J. Pept. Sci. 22, 271–279 (2016).

134 MacDonald, J. I., Munch, H. K., Moore, T. & Francis, M. B. One-step site-specific modification of native proteins with 2-pyridinecarboxyaldehydes. Nat. Chem. Biol. 11, 326–331 (2015).

135 Matheron, L. et al. Improving the selectivity of the phosphoric acid β-elimination on a biotinylated phosphopeptide. J. Am. Soc. Mass Spectr. 23, 1981–1990 (2012).

136 Du, J. et al. Metabolic glycoengineering: sialic acid and beyond. Glycobiology 19, 1382–1401 (2009).

137 Tommasone, S. et al. The challenges of glycan recognition with natural and artificial receptors. Chem. Soc. Rev. 48, 5488–5505 (2019).

2

Single-Molecule Peptide Fingerprinting

Jetty van Ginkel, **Mike Filius**, Malwina Szczepaniak, Pawel Tulinski, Anne S. Meyer, and Chirlmin Joo.

Proceedings of the National Academeny of Sciences
PNAS March 27, 2018 115 (13) 3338-3343
DOI: 10.1073/pnas.1707207115



2

2.1 Abstract

Proteomic analyses provide essential information on molecularpathways of cellular systems and the state of a living organism. 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 challenging. Here, we demonstrate a proof of concept of single-molecule FRET-based protein fingerprinting. We harnessed the AAA+protease ClpXP to scan peptides. By using donor fluorophore-labeled ClpP, we sequentially read out FRET signals from acceptor-labeled amino acids of peptides. The repurposed ClpXP exhibits unidirectional processing with high processivity and has the potential to detect low-abundance proteins. Our technique is a promising approach for sequencing protein substrates using a small amount of sample.

2.2 Introduction

Proteomic analyses provide essential information on molecular pathways of cellular systems and the state of a living organism.[1] Thereby, for understanding of biological processes and their (dys)regulation, including disease, it is critical to monitor the protein composition of cells by sequencing (i.e. determination of the amino-acid sequence). Mass spectrometry is currently the first choice for protein sequencing. However, mass spectrometry analysis often fails to recognize minor species embedded among other dominant species since sequence prediction is made through analysis of complex spectral peaks.[2] As many cellular proteins exist in low abundance[3], it is difficult to obtain large-scale proteomic information. DNA sequencing presents similar challenges, but they are overcome by amplifying DNA samples until a high signal-to-noise ratio is achieved. This solution cannot be applied to protein analysis since there is no natural machinery that can amplify proteins.

Single-molecule techniques have the potential to provide radically new protein sequencing tools that can quantify cellular proteins with accuracy as high as for mass spectrometry while requiring sample amounts as small as a single cell. However, despite several recent explorations[4-8], bona fide single-molecule protein sequencing has not yet been achieved due to the complexity that arises from primary protein sequences. Whereas DNA consists of only 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, our team and another 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.[9, 10] In brief, the number of protein species in an organism is finite and predictable. Through bioinformatics-based comparison with proteomics databases, ordered detection of only two types of amino acids can still allow for protein identification. For example, ordered detection of cysteine and lysine residues, which can be modified using orthogonal chemistries, is sufficient to sequence the human proteome.[10] We named this approach "single-molecule protein fingerprinting" to distinguish it from full protein sequencing. Here we demonstrate the first proof of concept of a single-molecule fingerprinting technology that reads out fluorescently labeled amino acids of synthetic peptides and a model cellular protein.

To obtain ordered determination of fluorescently labeled amino acids, we needed a molecular probe that can scan a peptide 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 monomers form a homohexameric ring ($ClpX_6$) that can exercise a large mechanical force to unfold proteins using ATP hydrolysis.[11, 12] Through iterative rounds of force-generating power strokes, $ClpX_6$ translocates substrates through the center of its ring in a processive manner[13, 14], with extensive promiscuity towards unnatural substrate modifications including fluorescent labels.[15-17] Protein substrates are recognized by $ClpX_6$ when they display specific disordered sequences such as the 11

amino-acid C-terminal ssrA tag.[18] $ClpX_6$ targets substrates for degradation by feeding them into $ClpP_{14}$, a homotetradecameric protease that contains 14 cleavage sites and self-assembles into a barrel-shaped complex that encloses a central chamber.[19]

2.3 Results

To immobilize ClpXP ($ClpX_6P_{14}$) for single-molecule imaging, we biotinylated $ClpX_6$ and bound $ClpX_6P_{14}$ to a PEG-coated quartz surface through biotin-streptavidin conjugation (**Figure 2.1**a). A combination of total internal reflection fluorescence microscopy (TIRF) and Alternating Laser EXcitation (ALEX) imaging[20, 21] was used to monitor individual ClpXP complexes bind, translocate and degrade dye-labeled substrates in real time.

To detect the progression of fluorescently labeled amino acids through the $ClpX_6$ pore with nanometer-scale accuracy, we employed FRET (Förster Resonance Energy Transfer).[22-24] We used two different types of model substrates for fingerprinting—short synthetic peptides and a small protein (the titin I27 domain). These substrates were labeled with acceptor fluorophores and were also appended with the ssrA tag. We constructed a FRET scanner by adding a fluorophore (donor) to the $ClpP_{14}$ chamber

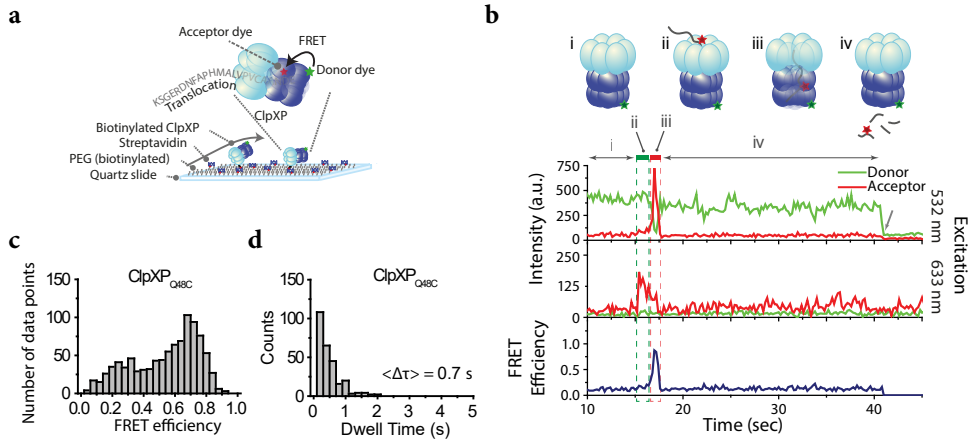


Figure 2.1: Single-molecule observation of ClpXP translocation.

(a) Schematics of the single-molecule fingerprinting platform. Donor-labeled ClpXP is immobilized on a PEG-coated slide via biotin-streptavidin conjugation. $ClpX_6$ recognizes an acceptor-labeled substrate (K-38-CCy5-ssrA, **Table S2.1**) and translocates it into the $ClpP_{14}$ chamber upon which FRET occurs. **(b)** A typical fluorescence time trace. (i) The donor signal is from Cy3-labeled ClpXP (upper trace) upon green excitation (532 nm). (ii, green box) The sudden appearance of acceptor signal (time ~16 s) during acceptor-direct excitation with red (633 nm) reports on binding of acceptor-labeled substrate to ClpXP (middle trace). (iii, red box) The high FRET (time ~17 s) reports on the presence of the substrate in $ClpP_{14}$ (top and bottom traces). (iv) Loss of fluorescence signal indicates the release of the substrate. The arrow at time ~40 s indicates the photobleaching of Cy3. **(c)** FRET distribution of stage iii. **(d)** Dwell time distribution of stage iii.

(**Figure 2.1**a). We introduced cysteines to the Q48 residue of ClpP ($ClpP_{Q48C}$), A139 ($ClpP_{A139C}$) or F31 ($ClpP_{F31C}$), labeled them with maleimide-functionalized fluorophores (**Figure S2.1**a), and evaluated the suitability for FRET-based substrate detection. $ClpP_{Q48C}$ and $ClpP_{A139C}$ showed higher FRET than $ClpP_{F31C}$ (**Figure S2.1**b). Among the first two, $ClpP_{Q48C}$ was chosen for our final scanner due to its higher efficiency of fluorophore labeling (see Methods).

The donor fluorophores on $ClpP_{Q48C}$ are located near the center of the $ClpP_{14}$ chamber, which is ~12 nm away from the substrate entry portal of $ClpX_6$ (**Figure S2.1**a).[25, 26] This distance is longer than the Förster radius of a standard single-molecule FRET pair (~5 nm). This physical separation enables us to selectively detect signals from only the fluorophores (acceptors) on a protein substrate that have been translocated through a $ClpX_6$ central channel. We obtained FRET time traces reporting on translocation, as shown in **Figure 2.1**b, by presenting a labeled peptide substrate to immobilized ClpXP complexes. The sudden appearance of acceptor signal during direct acceptor excitation indicates binding of acceptor-labeled peptide to ClpXP (**Figure 2.1**b, middle trace, stage ii). The subsequent appearance of a high FRET state indicates translocation of the substrate by $ClpX_6$ into the $ClpP_{14}$ chamber (**Figure 2.1**b, stage iii). When a slowly-hydrolyzable ATP analogue (ATPγS) was used, the probability of high-FRET appearance decreased by one order of magnitude (**Figure S2.2**d). Loss of FRET signal occurs upon the release of the dye-labeled peptide fragment (**Figure 2.1**b, stage iv). When a cleavage inhibitor (DFP, diisopropyl fluorophosphate) was used[27] (**Figure S2.2**a and b), the dwell time of high FRET increased 3.5-fold (**Figure S2.2**c).

Our single-molecule fingerprinting concept requires detection of the order of fluorophores on a single substrate. To demonstrate fingerprinting, we functionalized a peptide with one type of fluorophore (Cy3) at the N-terminal site and a second type of fluorophore (Cy5) on an internal cysteine residue. We monitored the order in which the two fluorophores passed through Alexa488-labeled $ClpP_{14}$ (**Figure 2.2**a). 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 since an ssrA-tagged substrate is translocated through ClpX6 starting from its C-terminus. **Figure 2.2**b depicts a representative time trace obtained from a substrate ($_{Cy3}NH_3$-17-C_{Cy5}-ssrA). The simultaneous appearance of Cy3 and Cy5 signals upon direct excitation with 532 nm and 637 nm (**Figure 2.2**b, middle and bottom, t ~ 40 s, indicated with arrows) indicates binding of a substrate containing both labels. In the FRET trace (**Figure 2.2**b, top), Alexa488-Cy5 FRET (marked * in the time trace) was observed before Alexa488-Cy3 FRET (marked °). This observation confirms that the ClpXP fingerprinter reads an ssrA-tagged substrate from the C-terminal to the N-terminal site.

We applied this fingerprinting scheme to the titin I27 domain. We labeled two Cys residues of titin (Cys64 and Cys80) with acceptor fluorophores. Because we did not have control over which dyes were attached to which Cys residues, we tagged both residues with the same dye, Cy5. Using Cy3 as a donor, we observed two separate FRET peaks within the time trajectories (**Figure S2.3**a). The time interval between the two peaks was elongated when ATPγS was mixed with ATP (**Figure S2.3**b), indicating that the two peaks represented the sequential probing of Cys80 and Cys64 residues.

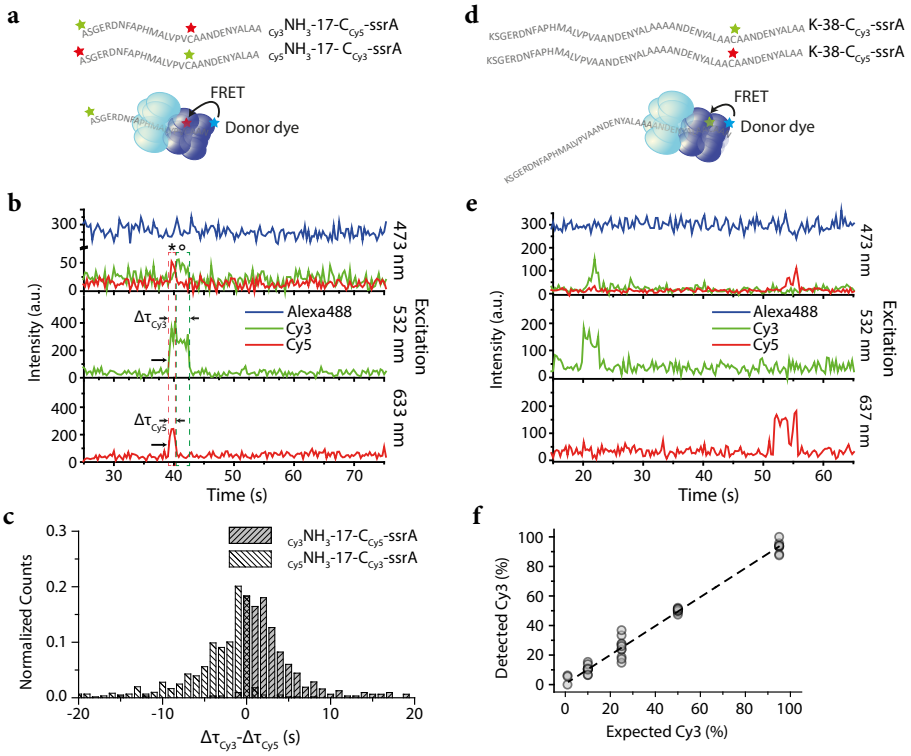


Figure 2.2: Single-molecule fingerprinting

(a) Substrates with two acceptor dyes were labeled at the N-terminal end and on cysteine residues (Supplementary Table 1). **(b)** A typical time trace for Cy3-NH3-17-CCy5-ssrA from three-color ALEX (top) showed FRET between Alexa488 and Cy3; and Alexa488 and Cy5 upon excitation with blue laser light (473 nm). Concurrent signals from Cy3 (middle) and Cy5 (bottom) upon direct excitation, respectively with green (532 nm) and red (637 nm). For clarity, an arbitrary offset of 200 a.u. was applied to the Alexa488 trace, and the sum of the Cy3 and Cy5 signals was plotted (middle). For the original trace, see Supplementary Figure 1e. **(c)** Comparison of acceptor dwell times. The dwell time of Cy5 ($\Delta\tau_{Cy5}$) was subtracted from that of Cy3 ($\Delta\tau_{Cy3}$) for each event from Cy3-NH3-17-CCy5-ssrA (grey). The mean of the distribution is 3.5 ± 4.90 (sec). In white is the same analysis for Cy5-NH3-17-CCy3-ssrA. The mean of the distribution is -3.9 ± 5.67 (sec). **(d)** Substrates labeled with one acceptor dye. Cysteine residues of the substrates (K-38-C-ssrA) were labeled with either Cy3 or Cy5. **(e)** A time trace from substrates in [d]. At time ~20 s, a Cy3-labeled substrate binds (top, Alexa488-Cy3 FRET) and 532 nm direct excitation (middle). At time ~50 s, a Cy5-labeled substrate bind (top, Alexa488-Cy5 FRET) and 637 nm direct excitation (bottom). An offset of 200 a.u. applied to Alexa488. **(f)** The percentage of processed Cy3-labeled substrates was plotted against the expected percentage of Cy3-labeled substrates. The line is a linear fit (slope of 0.98 ± 0.02, intercept 0.58 ± 0.79, R2 = 0.99). Data points are from 100-s recordings, repeated ten times per condition (each n = 24.4 ± 1.50).

We extracted the length of time that Cy3 and Cy5 acceptor fluorophores were engaged with ClpXP ($\Delta\tau_{Cy3}$, $\Delta\tau_{Cy5}$). We observed positive differences in dwell time ($\Delta\tau_{Cy3\text{-}Cy5} = \Delta\tau_{Cy3} - \Delta\tau_{Cy5}$, $< \Delta\tau_{Cy3\text{-}Cy5} > = 3.5$ sec) for a substrate with N-terminal Cy3-labeling and internal Cy5-labeling (**Figure 2.2**c, grey, $_{Cy3}NH_3$-17-C_{Cy5}-ssrA). For a substrate with exchanged dye positions ($_{Cy5}NH_3$-17-C_{Cy3}-ssrA), we observed negative differences (**Figure 2.2**c, white, $< \Delta\tau_{Cy5\text{-}Cy3} > = -3.9$ sec). Thus, dye-labeled amino acids located closer to the C-terminal ssrA tag were retained in the ClpXP complex for shorter amounts of time than labeled amino acids located more closely to the N-terminus. We can conclude that our fingerprinter can detect dyes in an order matching the amino-acid sequence. The ordered disappearance of the Cy3 and Cy5 signals further implies that uncleaved or partially cleaved substrate does not accumulate within the $ClpP_{14}$ chamber, which would otherwise hamper accurate fingerprinting.

A single-molecule fingerprinter should perform without any bias to fluorophores and with high dynamic range. To determine the sensitivity of our fingerprinter, we performed a population study in which $ClpP_{14}$ was labeled with donor fluorophore (Alexa488) and substrate peptides were singly labeled with either Cy3 or Cy5 as an acceptor fluorophore (**Figure 2.2** d and 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 relationship between the percentage of Cy3-labeled substrates we detected versus the expectation, with an offset of 0.58 ± 0.79 % and a slope of 0.98 ± 0.02 (adjusted R2 = 0.99) (**Figure 2.2**f). We conclude that both FRET pairs are detected with equal sensitivity, and that our FRET scanner has the potential to detect low abundance proteins.

Our previous computational analysis 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 their order.[10] A uniform speed of the scanner, represented by $ClpX_6$, is crucial to extract distance information. To determine whether the processing time of ClpXP is proportional to the length of protein substrates, we determined the processing times (the dwell time of fluorescence signals emitted by Cy5 labels on substrates, upon direct excitation) for three peptides (29, 40, and 51 amino acids (AA) in length; see Supplementary Table 1) and monomeric (119 AA) and dimeric (210 AA) versions of titin (all labeled at Cys, see Table 1). Plotting the total time ($\Delta\tau$, see **Figure 2.3**a) that a substrate was bound and processed by ClpXP versus the length of the substrates showed a linear increase with an average processing speed of 23.9 amino acids per second (**Figure 2.3**b and **Figure S2.2**), which agrees with previous results obtained from both bulk[28] and single-molecule assays.[11, 12, 29] We obtained a similar processing speed of 14.5 amino acids per second (translocation of 16 amino acids for 1.1 seconds) from the doubly-labeled titin (**Figure S2.3**b). In **Figure 2.3**b, the y-axis offset of 4.2 s reports on the initial docking phase and the eventual retention within $ClpP_{14}$. Our data indicates that the ClpXP fingerprinter has the potential to determine both the order and spacing distance of labeled residues.

Uni-directional translocation is also of utmost importance for our new technology. Backtracking of ClpXP would result in insertion errors in the observed fingerprint and thus reduce the detection precision. To evaluate the occurrence of backtracking, we

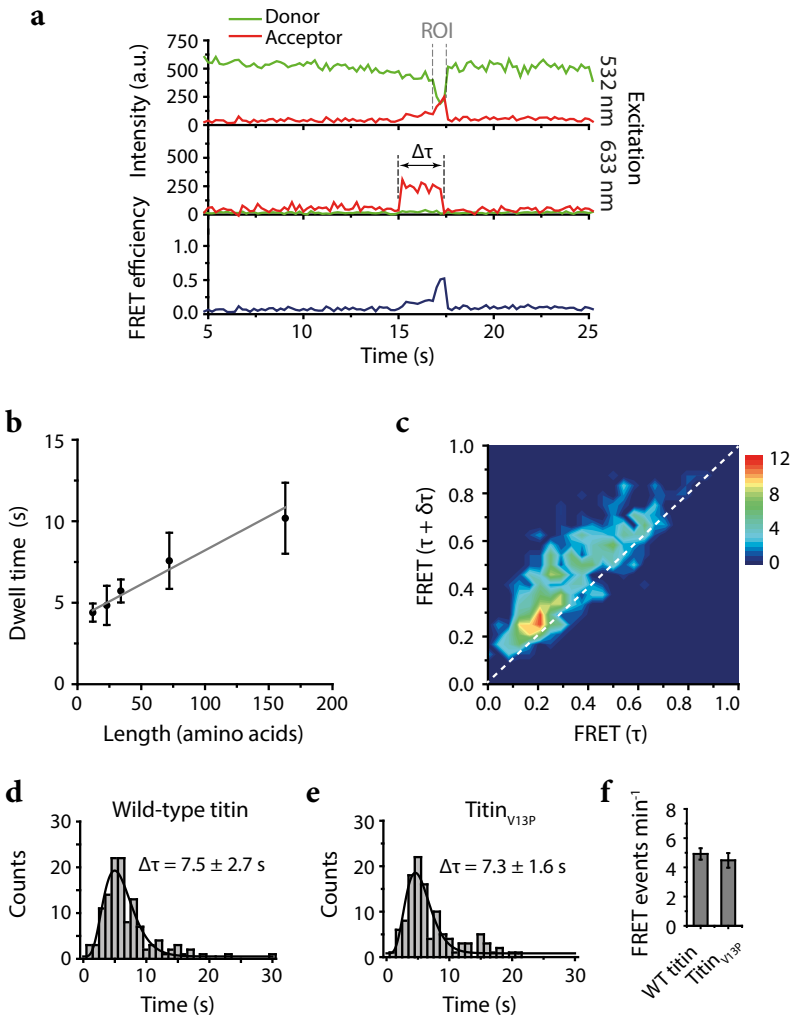


Figure 2.3: ClpXP performs uni-directional scanning with a constant speed.
(a) Representative time trace. ROI (Region of Interest) is where the FRET efficiency gradually increases. Δτ: the total docking time. **(b)** Total dwell time vs. substrate length. The average time, <Δτ>, was obtained by fitting data in Supplementary Figure 4 with a gamma distribution. 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). The substrate length is the number of amino acids between the C-terminus and a dye the most proximal to the N-terminus. Error bars obtained by bootstrapping with 1000 resamples. A linear fit results in an offset of 4.0 ± 0.20 s and a speed of 23.9 ± 2.86 amino acids/s. **(c)** Transition density plot. FRET change was analyzed by measuring FRETt = τ and FRETt = τ + δτ, with δτ = 0.4 s, for every point in ROI. The dotted line represents FRETδτ = FRETτ + δτ. K-38-C-ssrA was used. **(d)** and **(e)** Total dwell times (Δτ) for wild-type titin and titinV13P. Δτ = 7.5 ± 2.7 s and 7.3 ± 1.6 s were obtained respectively by fitting with a gamma distribution. Errors obtained by bootstrapping with 1000 resamples. Wild-type titin, n = 123. TitinV13P, n = 112. **(f)** The number of traces showing FRET events for wild-type (WT) titin and titinV13P. Error bars are standard deviations from 15 measurements.

determined the change in FRET over time during processing of peptide substrates. We created a 2D heat map by plotting the change of FRET over a given time interval $\delta\tau$. In **Figure 2.3**c, FRET ($t = \tau + \delta\tau$) versus FRET ($t = \tau$) is deposited for every time point along a time trace reporting on translocation (**Figure 2.3**a, ROI (region of interest)). We set $\delta\tau = 0.4$ s, a time scale longer than our time resolution (0.2 s) but shorter than the average translocation time (0.7 s, **Figure 2.1**d), to visualize the gradual increase of FRET. Any backtracking of $ClpX_6$ along the substrate would result in momentary FRET decrease during translocation, which would appear as FRET ($t = \tau + \delta\tau$) values lower than FRET ($t = \tau$) (population below the diagonal). We observed FRET ($t = \tau + \delta\tau$) higher than FRET ($t = \tau$) (upper diagonal population) for a major fraction (92.5 %) of the data points. The remaining fraction is likely due to the backtracking of ClpX, the statistical noise of the fluorescence signals, and the photoblinking of acceptor dyes. This degree of experimental error is predicted not to interfere with the ability to extract length information according to our computational simulation.[10]

A single-molecule protein fingerprinter should be able to process any structural element of a protein. Single-molecule force spectroscopy studies of ClpXP have shown that $ClpX_6$ stalls on substrates with rigid secondary structures[30,31], which would inhibit the extraction of sequence information. We therefore explored the possibility of disrupting such tightly folded structures to enable fingerprinting. Perturbation of cysteine residues in the titin protein has been shown to interfere with the secondary structure of the protein, making it behave as an unstructured polypeptide chain.[32,33] We purified the I27 domain of both wild-type titin, known to make $ClpX_6$ stall[30], and titinV13P, a variant that is still folded but is degraded at a rate close to denatured titin.[33] By fluorophore labeling the cysteine residues of wild-type titin and titinV13P, we sought to determine the degree of structural influence of the cysteine-dye conjugation on ClpXP processing. We obtained equivalent total dwell times for processing stable wild-type titin ($\Delta\tau = 7.5 \pm 2.7$ s, **Figure 2.3**d) and titinV13P ($\Delta\tau = 7.3 \pm 1.6$ s, **Figure 2.3**e). A similar number of both substrates was processed by ClpXP within our time interval of observation (**Figure 2.3**f), indicating that $ClpX^6$ can process labeled wild-type and V13P substrates with the same efficiency. These results suggest that preparing substrates for sequencing by labeling cysteine residues (and likely lysine residues as well) might sufficiently destabilize their protein structures. This will allow for fingerprinting of any protein regardless of structural stability.

2.4 Discussion

We have demonstrated a FRET-based detection platform utilizing an AAA+ protease as a scanner of peptides and proteins. In our approach, we conjugate fluorophores to thiol groups of cysteine residues and amine groups of the N-terminal site (which can be extended to lysine residues) because these chemical groups can be labeled with high efficiency and specificity. Our platform, however, is not limited to these two modifications. With appropriate chemistry, one could target other residues or even post-translational modifications. Detection of these moieties could be implemented by extending our current three-color FRET scheme to four-color FRET.[21]

Several outstanding perspectives remain in order for our method to be directly

applied to a protein sequencing technology. First, for proteomics analysis, our sequencing technique has to work for all cellular proteins without sequence bias. ClpX, a core of our platform, only recognizes substrates displaying specific sequence tags including ssrA. The substrate selectivity of ClpX would need to be broadened, perhaps through targeted mutations in the substrate-recognition loops of the ClpX channel, or the use of engineered adaptor proteins (e.g. modified SspB) that non-specifically deliver substrates to ClpX. Second, a challenge of cellular protein analysis is to detect low-abundance proteins within a complex sample such as a clinical tissue sample. The depth of the sequencing coverage might be increased by removing housekeeping proteins chromatographically.[34] Third, to cover the whole proteome in a reasonable amount of time, the throughput should be enhanced. Under the standard conditions used in this work (10 nM substrate, 512x512 pixel camera, $ClpX_6$), we obtained ~10 productive reactions per minute per imaging area. By using a CMOS camera that has a larger number of pixels (e.g. 2084x2084 pixels from (Juette et al.[35]) as well as a zero-mode waveguide platform that allows for single-molecule imaging of a higher concentration of substrate (e.g. 1 μM)[36], the throughput would be improved by a factor of ~1000. We also observed that productive reactions (a trace ending with high FRET) make up only ~10% of the total population (**Figure S2.2**D). We suspect that this low yield is due to the lack of the N-terminal domain of ClpX in hexameric linked ClpX. By using wildtype, monomeric ClpX and also introducing an adaptor protein that facilitates substrate binding, we expect that the percentage of the productive population will reach near 100%. When our sequencer is improved with these changes, we expect to cover 1x of a single human-cell proteome (~108 proteins) in nearly 10 hours (10 events /min * 16 * 100 * 10 ≈ 107 events / hour).

Our method has the potential to scan full-length proteins from end to end without the need for fragmentation. Sequencing substrates are processed at a constant speed, allowing for accurate protein identification.[10] In this proof-of-concept study we show our capability to detect low-frequency sub-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 into an invaluable asset to clinical diagnostics.

2.5 Materials And Methods

2.5.1 $ClpX_6$ purification and biotinylation

To ensure proper immobilization and hexamer formation of ClpX at low concentrations, $ClpX_6$ (ΔN), a covalently linked hexamer containing a single biotinylation site, was used throughout the experiments. $ClpX_6$ (ΔN) was overexpressed and purified as described.(37) In brief, ClpX6 protein expression and biotinylation was induced in a E. coli BLR(DE3) strain at $O.D._{600}$ ~0.6 by adding 1.0 mM IPTG and 100 μM of biotin to increase BirA-mediated biotinylation efficiency. The culture was incubated overnight at 18 °C. 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 1 mM PMSF and lysed by French press twice at 20 psi. $ClpX_6$(ΔN) 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).

2.5.2 ClpP mutations, purification and labeling

Point mutations were constructed in ClpP by overlap extension PCR to produce the cysteine-free variant $ClpP_{C91S\text{-}C113S}$, and the subsequent variants $ClpP_{Q48C}$, $ClpP_{A139C}$ and $ClpP_{F31C}$. The variants were overexpressed in E. coli BL21(DE3)pLysS at $O.D._{600}$ ~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 efficiency of 5.9, 1.1, and 1.7 dyes per tetradecameric $ClpP_{Q48C}$, $ClpP_{A139C}$, and $ClpP_{F31C}$, respectively, was measured by spectrophotometry (DeNovix DS-11 FX).

2.5.3 ClpP inactivation

Purified $ClpP_{Q48C}$ was chemically inactivated as described previously (1). Briefly, $ClpP_{Q48C}$ (4 μM) was inactivated in PD buffer containing 10 mM DFP (Sigma). The reaction was incubated for 6 h at 4 °C and then dialyzed twice: 1x 2 h and 1x overnight against 1x PBS (pH 7.4). $ClpP_{Q48C_DFP}$ was labeled with monoreactive maleimide donor dye, Cy3, for 4 hours at 4 °C. A 10x molar excess of dye was used in PBS pH 7.4 under nitrogen. Free dye was removed using Pierce™ Dye Removal Columns (Thermo Fisher). A labeling efficiency of 1.8 dye per tetradecameric $ClpP_{Q48C_DFP}$ was measured by spectrophotometry (DeNovix DS-11 FX).

2.5.4 ClpXP cleavage reaction

To assess the enzymatic activity of donor-labeled ClpXP, 0.9 µM ClpX and 2.9 µM of ClpP (WT or variants) in PD buffer (25 mM HEPES pH 8.0, 5 mM $MgCl_2$, 40 mM KCl, 0.148 % NP-40, 10 % glycerol) were incubated at 30 °C in the presence of 10 µM titin$_{V13P}$-ssrA and 5 mM ATP. Samples were taken at t = 0 min and 30 min and analyzed using 4 - 20 % precast SDS-PAGE gels (Thermo Scientific) and coomassie staining.

2.5.5 Substrate preparation

Titin-I27 (wild-type, V13P and dimer) with the C-terminal ssrA tag was expressed in E. coli BL21AI at O.D.$_{600}$ ~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 4 M 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 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 up to 95 % were measured by spectrophotometry (DeNovix DS-11 FX) (See **Table S2.1** and **Table S2.2** for the full list of substrates.)

2.5.6 Single-molecule sample preparation

To reduce the nonspecific binding of proteins, acidic piranha-etched quartz slides (G. Finkenbeiner) were passivated with two rounds of polyethylene glycol (mPEG--Succinimidyl Valerate, MW 5000 Laysan, followed by MS(PEG)$_4$, Piercenet) as described previously (38). After assembly of a microfluidic flow chamber, slides were incubated with 5 % Tween-20 for 10 min[39], and excess 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.[40] 30 nM $ClpX_6$ and 90 nM $ClpP_{14}$ (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)). Donor-labeled $ClpP_{14}$ added into a chamber without $ClpX_6$ led to very few non-specifically immobilized ClpP protein complexes, ruling out any non-specific adsorption of $ClpP_{14}$ to the surface (**Figure S2.1**B). All experiments were performed at room temperature (23 ± 2 °C).

2.5.7 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).

2.5.8 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 with an exposure time of 0.1 s was recorded. The time traces were extracted from the CCD image series using an IDL (ITT Visual Information Solution) 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).

2.6 Supporting Information

2.6.1 Supporting Figures

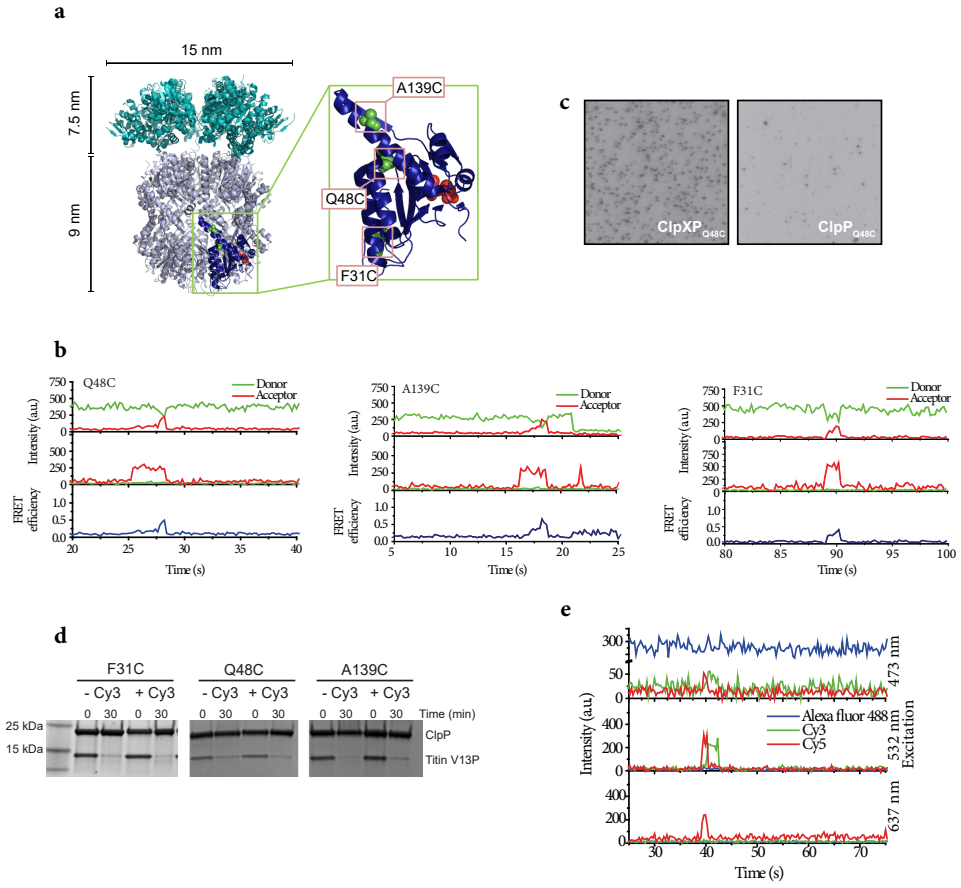


Figure S2.1: ClpXP modifications

(a) Co-crystal structure of ClpXP, obtained by manually combining crystal structures from PDB:1YG6 (ClpP$_{14}$) and PDB: 3HTE (ClpX$_6$). Highlighted in red are the two cysteine residues present in wild-type ClpP. Highlighted in green are cysteines introduced into three variants: ClpP$_{Q48C}$, ClpP$_{A139C}$and ClpP$_{F31C}$. **(b)** Representative time trace from ClpP$_{Q48C}$ (left), ClpP$_{A139C}$ (middle), and ClpP$_{F31C}$ (right). ClpP$_{Q48C}$ and ClpP$_{A139C}$ exhibit higher FRET efficiency than ClpP$_{F31C}$. **(c)** CCD images (donor channel) showing immobilization of donor-labeled ClpP$_{Q48C}$ in complex with ClpX$_6$ (left) or in the absence of ClpX$_6$ (right). Each spot represents a single donor-labeled ClpP$_{14}$ molecule. **(d)** Degradation of titin$_{V13P}$ by ClpXP. The degradation efficiency of unlabeled and labeled ClpP variants was compared at time 0 and 30 min. **(e)** Three-color time trace. The original time trace used in **Figure 2.2**b to present a three-color FRET event. Note the original, not summed, levels of Cy3 and Cy5 signals in the middle panel. Cy3 transfers its energy to Cy5 via FRET.

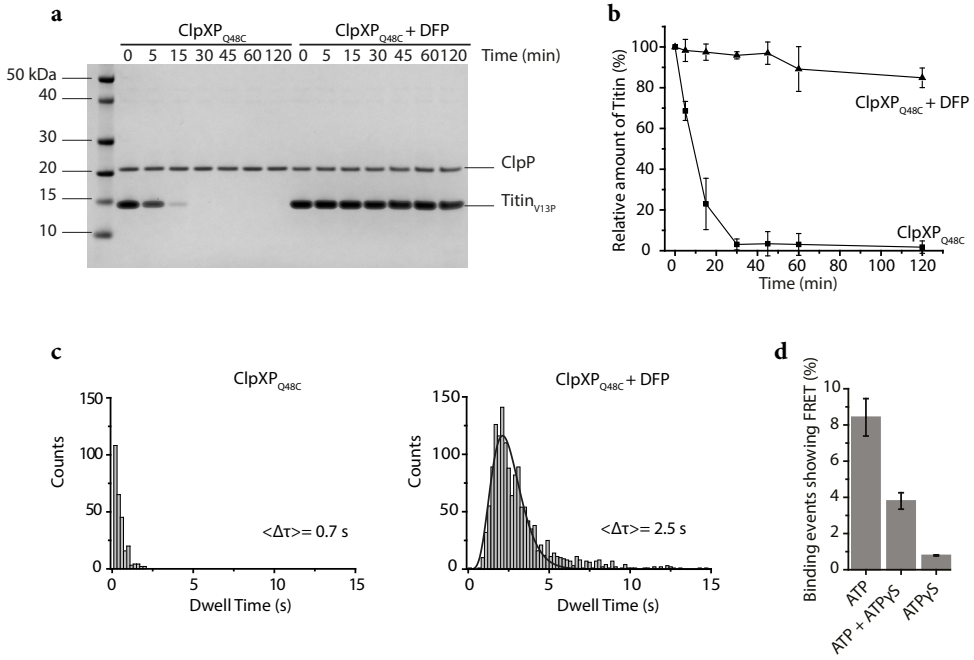


Figure S2.2: Evaluation of chemically inactivated ClpX and ClpP.
(A) Degradation of titinV13P by ClpPQ48C or chemically inactiviated ClpPQ48C+ DFP. **(B)** The change of the relative amount of titin substrate when incubated with ClpXP$_{Q48C}$ or chemically inactiviated ClpXP$_{Q48C}$+ DFP. **(C)** Dwell-time histograms of single-molecule experiments of K-38-C-ssrA with ClpXP$_{Q48C}$ or chemically inactiviated ClpXP$_{Q48C}$+ DFP, the dwell-time of the high FRET increase 3.5-fold when ClpP$_{Q48C}$ + DFP was used. **(D)** ClpX was inactivated by using an ATP analogue ATPγS, increasing the ATPγS concentration reducies the number of binding events showing FRET.

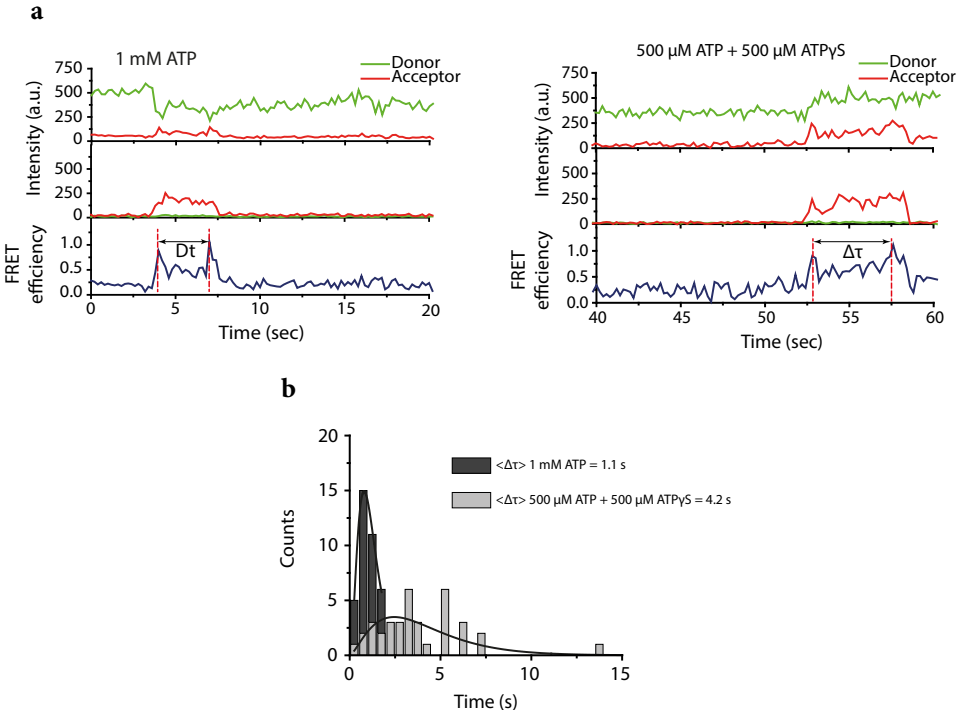


Figure S2.3: Fingerprinting cysteine residues in Titin.

(A) Representative time traces of the titin substrate with both cysteines labeled with Cy5 (Cys64 and Cys80). A mixture of ATP and ATPγS (right trace) was used to elongate the dwell-time between two of the cysteines. **(B)** The dwell-time histograms of the elongated interval between the two peaks of 1 mM ATP and 500 µM ATP + 500 µM ATPγS.

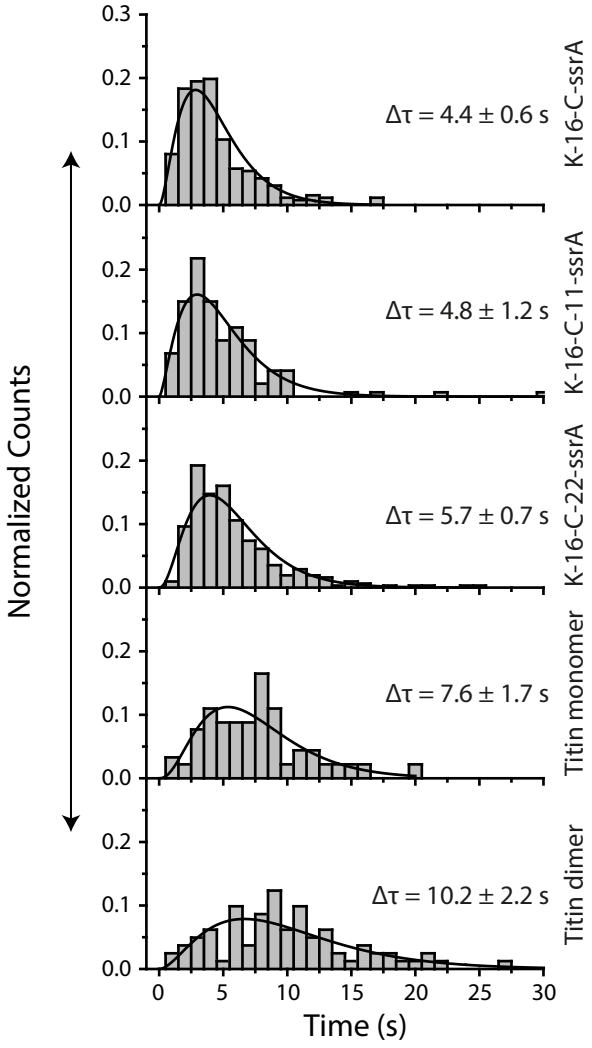


Figure S2.4: Distribution of total dwell time for different substrates.
Total dwell times (Δτ; see the definition in **Figure 2.3**A) were determined for ssrA-tagged peptides of increasing lengths and monomeric and dimeric titin. Total dwell times for all substrates showed gamma-like distributions. Errors were obtained by bootstrapping with 1000 resamples.

2.6.2 Supplementary Table

Table S2.1: Single-molecule peptide substrates.

Construct	Amino acid Sequence (N → C)	length in amino acids	Molecular Weight (kDa)
NH_3-17-C-ssrA	ASGERDNFAPHMALVPV**C**AAN-DENYALAA	29	3.018
K-16-C-ssrA	KSGERDNFAPHMALVPV**C**AAN-DENYALAA	29	3.075
K-16-C-11-ssrA	KSGERDNFAPHMALVPV**C**AAN-DENYALAAAANDENYALAA	40	4.180
K-16-C-22-ssrA	KSGERDNFAPHMALVPV**C**AAN-DENYALAAAANDENYALAAAAN-DENYALAA	51	5.284
K-38-C-ssrA	KSGERDNFAPHMALVPVAAN-DENYALAAAANDENYA-LAA**C**AANDENYALAA	51	5.284

Table S2.2: Single-molecule Protein substrates.

Construct	Amino acid Sequence (N → C)	length in amino acids	Molecular Weight (kDa)
Titin-ssrA	MRGSHHHHHHGLVPRGS- LIEVEKPLYGVEVFVGE- TAHFEIELSEPDVHGQW- KLKGQPLAASPD**C**EIIED- GKKHILILHN**C**QLGMTGE- VSFQAANTKSAANLKV- KELRSAANDENYALAA	119	13.050
Titin-V13P-ssrA	MRGSHHHHHHGLVPRGS- LIEVEKPLYGVEPFVGE- TAHFEIELSEPDVHGQW- KLKGQPLAASPD**C**EIIED- GKKHILILHN**C**QLGMTGE- VSFQAANTKSAANLKV- KELRSAANDENYALAA	119	13.048
Titin-Titin-ssrA	MRGSHHHHHHGLVPRGS- LIEVEKPLYGVEVFVGE- TAHFEIELSEPDVHGQW- KLKGQPLAASPD**C**EIIED- GKKHILILHN**C**QLGMTGE- VSFQAANTKSAANLKV- KELRSLIEVEKPLYGVE- VFVGETAHFEIELSEPDVH- GQWKLKGQPLAASPD**C**EI- IEDGKKHILILHN**C**QLG- MTGEVSFQAANTKSAAN- LKVKELRSAANDENYALAA	210	23.056

2.7 References

1 Harper JW & Bennett EJ (2016) Proteome complexity and the forces that drive proteome imbalance. Nature 537(7620):328-338.

2 Zubarev RA (2013) The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 13(5):723-726.

3 Talapatra A, Rouse R, & Hardiman G (2002) Protein microarrays: challenges and promises. Pharmacogenomics 3(4):527-536.

4 Nivala J, Marks DB, & Akeson M (2013) Unfoldase-mediated protein translocation through an α-hemolysin nanopore. Nature biotechnology:1-5.

5 Rosen CB, Rodriguez-Larrea D, & Bayley H (2014) Single-molecule site-specific detection of protein phosphorylation with a nanopore. Nature biotechnology:1-3.

6 Zhao Y, et al. (2014) Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. Nature nanotechnology 9:466-473.

7 Ohshiro T, et al. (2014) Detection of post-translational modifications in single peptides using electron tunnelling currents. Nature Nanotechnology 9:835-840.

8 Kennedy E, Dong Z, Tennant C, & Timp G (2016) Reading the primary structure of a protein with 0.07 nm3 resolution using a subnanometre-diameter pore. Nature Nanotechnology.

9 Swaminathan J, Boulgakov AA, & Marcotte EM (2015) A theoretical justification for single molecule peptide sequencing. PLoS computational biology 11:e1004080.

10 Yao Y, Docter M, van Ginkel J, de Ridder D, & Joo C (2015) Single-molecule protein sequencing through fingerprinting: computational assessment. Physical biology 12:055003.

11 Aubin-Tam M-E, Olivares AO, Sauer RT, Baker Ta, & Lang MJ (2011) Single-Molecule Protein Unfolding and Translocation by an ATP-Fueled Proteolytic Machine. Cell 145:257-267.

12 Maillard RA, et al. (2011) ClpX(P) Generates Mechanical Force to Unfold and Translocate Its Protein Substrates. Cell 145:459-469.

13 Sen M, et al. (2013) The ClpXP protease unfolds substrates using a constant rate of pulling but different gears. Cell 155:636-646.

14 Thompson MW, Singh SK, & Maurizi MR (1994) 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. The Journal of biological chemistry 269:18209-18215.

15 Barkow SR, Levchenko I, Baker Ta, & Sauer RT (2009) Polypeptide translocation by the AAA+ ClpXP protease machine. Chemistry & biology 16:605-612.

16 Burton RE, Siddiqui SM, Kim YI, Baker Ta, & Sauer RT (2001) Effects of protein stability and structure on substrate processing by the ClpXP unfolding and degradation machine. The EMBO journal 20:3092-3100.

17 Kolygo K, et al. (2009) Studying chaperone-proteases using a real-time approach based on FRET. Journal of structural biology 168:267-277.

18 Keiler KC, Waller PR, & Sauer RT (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. Science (New York, N.Y.) 271:990-993.

19 Baker Ta & Sauer RT (2012) ClpXP, an ATP-powered unfolding and protein-degradation machine. Biochimica et biophysica acta 1823:15-28.

20 Kapanidis AN, et al. (2004) Fluorescence-aided molecule sorting: analysis of structure and interactions by alternating-laser excitation of single molecules. Proceedings of the National Academy of Sciences of the United States of America 101:8936-8941.

21 Lee J, et al. (2010) Single-Molecule Four-Color FRET. Angewandte Chemie 122:10118-10121.

22 van Oijen AM (2011) Single-molecule approaches to characterizing kinetics of biomolecular interactions. Curr Opin Biotechnol 22(1):75-80.

23 Lamichhane R, Solem A, Black W, & Rueda D (2010) Single-molecule FRET of protein-nucleic acid and protein-protein complexes: Surface passivation and immobilization. Methods 52(2):192-200.

24 Bae W, Choi MG, Hyeon C, Shin YK, & Yoon TY (2013) Real-time observation of multiple-protein complex formation with single-molecule FRET. J Am Chem Soc 135(28):10254-10257.

25 Flanagan JM, Wall JS, Capel MS, Schneider DK, & Shanklin J (1995) Scanning transmission electron microscopy and small-angle scattering provide evidence that native Escherichia coli ClpP is a tetradecamer with an axial pore. Biochemistry 34:10910-10917.

26 Kim DY & Kim KK (2003) Crystal structure of ClpX molecular chaperone from Helicobacter pylori. The Journal of biological chemistry 278:50664-50670.

27 Maurizi MR, Clark WP, Kim SH, & Gottesman S (1990) ClpP represents a unique family of serine proteases. J Biol Chem 265(21):12546-12552.

28 Martin A, Baker Ta, & Sauer RT (2008) Protein unfolding by a AAA+ protease is dependent on ATP-hydrolysis rates and substrate energy landscapes. Nature structural & molecular biology 15:139-145.

29 Shin Y, et al. (2009) Single-molecule denaturation and degradation of proteins by the AAA+ ClpXP protease. Proceedings of the National Academy of Sciences of the United States of America 106:19340-19345.

30 Cordova JC, et al. (2014) Stochastic but Highly Coordinated Protein Unfolding and Translocation by the ClpXP Proteolytic Machine. Cell 158:647-658.

31 Nivala J, Mulroney L, Li G, Schreiber J, & Akeson M (2014) Discrimination among protein variants using an unfoldase-coupled nanopore. ACS nano 8:12365-12375.

32 Iosefson O, Nager AR, Baker Ta, & Sauer RT (2015) Coordinated gripping of substrate by subunits of an AAA+ proteolytic machine. Nature chemical biology.

33 Kenniston Ja, Baker Ta, Fernandez JM, & Sauer RT (2003) Linkage between ATP consumption and mechanical unfolding during the protein processing reactions of an AAA+ degradation machine. Cell 114:511-520.

34 Han X, Aslanian A, & Yates JR, 3rd (2008) Mass spectrometry for proteomics. Curr Opin Chem Biol 12(5):483-490.

35 Juette MF, et al. (2016) Single-molecule imaging of non-equilibrium molecular ensembles on the millisecond timescale. Nat Methods 13(4):341-344.

36 Levene MJ, et al. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299(5607):682-686.

37 Martin A, Baker Ta, & Sauer RT (2005) Rebuilt AAA + motors reveal operating principles for ATP-fuelled machines. Nature 437:1115-1120.

38 Chandradoss SD, et al. (2014) Surface passivation for single-molecule protein studies. Journal of visualized experiments : JoVE:e50549.

39 Pan H, Xia Y, Qin M, Cao Y, & Wang W (2015) A simple procedure to improve the surface passivation for single molecule fluorescence studies. Phys Biol 12(4):045006.

40 Singh SK, et al. (2001) Functional domains of the ClpA and ClpX molecular chaperones identified by limited proteolysis and deletion analysis. The Journal of biological chemistry 276:29420-29429.

2

3

High-Resolution Single-Molecule FRET via DNA eXchange (FRET X)

Mike Filius, Sung Hyun Kim, Ivo Severins, and Chirlmin Joo

Nano Letters
Nano Lett. 2021, 21, 7, 3295–3301
DOI: 10.1021/acs.nanolett.1c00725



3

3.1 Abstract

Single-molecule FRET is a versatile tool to study nucleic acids and proteins at the nanometer scale. However, currently, only a couple of FRET pairs can be reliably measured on a single object, which makes it difficult to apply single-molecule FRET for structural analysis of biomolecules. Here we present an approach that allows for the determination of multiple distances between FRET pairs in a single object. We use programmable, transient binding between short DNA strands to resolve the FRET efficiency of multiple fluorophore pairs. By allowing only a single FRET pair to be formed at a time, we can determine the pair distance with sub-nanometer precision. The distance between other pairs are determined by sequentially exchanging DNA strands. We name this multiplexing approach FRET X for FRET via DNA eXchange. Our FRET X technology will be a tool for the high-resolution analysis of biomolecules and nano-structures.

3.2 Introduction

X-ray crystallography, nuclear magnetic resonance and cryo-electron microscopy are the golden standard for determining the structure of biomolecules.[1,2] However, minute conformational changes of biomolecules often cannot be observed with these techniques since a certain conformation may be stabilized by the required sample preparation.[3] Single-molecule FRET can be used to determine the structure of molecules—including rare conformations—with sub-nanometer resolution. However, the use of single-molecule FRET for the analysis of complex molecular structures, e.g. protein tertiary structures, has been limited since it requires resolving the FRET efficiency of multiple dye pairs.[4,5] Currently, single-molecule FRET analysis allows us to deal with only one or two FRET pairs in a single measurement.[6,7] Therefore, structural analysis using single-molecule FRET requires the preparation of a protein library consisting of many different combinations of dye locations, rigorous modeling and simulations following the data acquisition.[8–11]

Single-molecule multiplexing has been demonstrated with photoswitchable fluorophores. In this approach, a molecule of interest is labeled with a single donor and multiple identical acceptor fluorophores. By using photoswitchable acceptor fluorophores, only one of the acceptors is active at a given time.[12] This method, called switchable FRET, allows for the detection of multiple FRET pairs in a single nanoscale object and thereby the determination of structures within and interactions between biomolecules ranging from proteins to DNA. However, the stochastic nature of the photoswitching is one of the main obstacles for the wide adaptation of the method. An alternative way of switching between on and off states of fluorescent probes is by using fluorophores that bind a target only for short period of time, as with point accumulation in nanoscale topography (PAINT).[13–15] For example, fluorophores are attached to short DNA oligos that bind the complementary target strands for several hundreds of milliseconds. This transient binding is central to the super-resolution technique DNA-based point accumulation for imaging in nanoscale topography (DNA-PAINT).[16–19]

Here we propose a new structural analysis tool that can resolve the FRET efficiency of multiple pairs in a single target molecule. By using programmable, transient binding between short DNA strands, a single FRET pair is formed at any given time allowing for accurate distance determination between the momentarily formed fluorophore pair. By repeating the imaging cycle, we can resolve multiple points of interest (POIs) in a single nanoscale object. We demonstrate the proof of concept of sub-nanometer resolution single-molecule structural analysis on various DNA nanostructures.

3.3 Results

To demonstrate the concept of FRET via DNA strands, we designed an assay where an acceptor (Cy5)-labeled single-stranded (ss) DNA molecule was immobilized on a quartz slide through biotin-streptavidin conjugation (**Figure 3.1**a). The measurements yielded a distinct fluorescence signal in single-molecule total internal refection microscopy images upon binding of a donor-labeled imager strand on

the immobilized target strand (**Figure 3.1**b). The base sequence and length of the imager strand sequence was chosen such that the binding events between the two DNA strands would have a short dwell-time to allow for frequent replenishment of

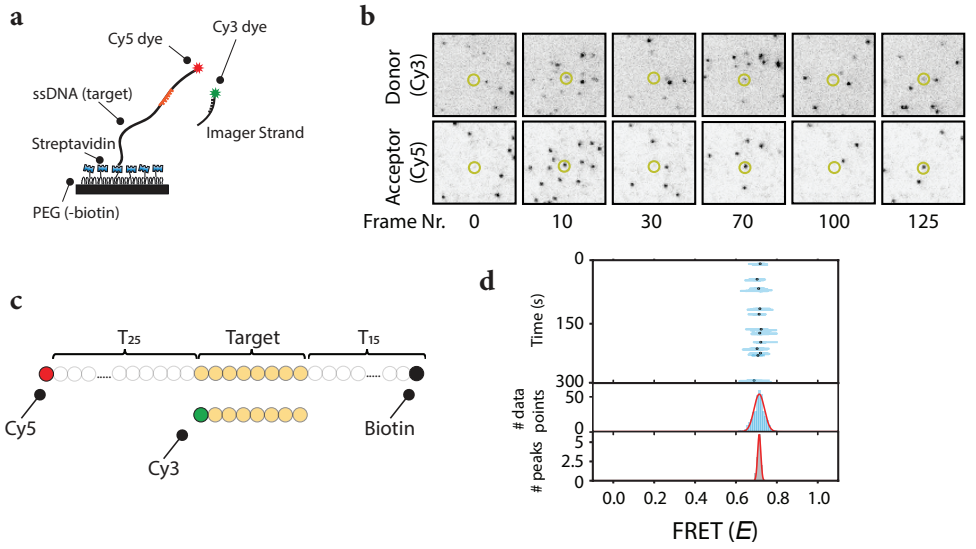


Figure 3.1: Repetitive binding of short DNA imager strands allows for high detection precision for Single-Molecule FRET.

(a) Schematic representation of the single-molecule FRET assay. An acceptor (Cy5, red star) labeled single-stranded target DNA construct is immobilized on a PEGylated surface through biotin-streptavidin conjugation. Binding of the donor (Cy3, green star) labeled imager strand results in short FRET events and is observed using total internal reflection microscopy. **(b)** A series of CCD snapshots obtained from a single-molecule movie with 100 ms exposure time. The top row represents the donor channel, and the bottom row represents the acceptor channel. Each dot represents a single molecule. Dynamic binding of the imager strands can be observed over time (highlighted molecule). **(c)** Schematic representation of the ssDNA constructs. Upon binding of the imager strand, the donor fluorophore is separated from the acceptor by a 25-nt thymine linker. **(d)** Single-molecule FRET kymograph from a time trace from one single molecule (ROI, highlighted molecule from panel b). The kymograph shows the FRET efficiency for each data point in a binding event (blue lines) and the mean FRET efficiency from all data points per binding event (dots) as a function of time. A FRET histogram that is built from the efficiency for each datapoint (d, middle panel) has a larger standard deviation (0.72 ± 0.05) compared to the standard deviation (0.72 ± 0.01) from a histogram that is built from the mean FRET values per binding events (d, bottom panel).

the imager strand (**Figure 3.1**b-c and **Figure S3.1**a), and thus for the same POI to be probed multiple times.[20] At the same time, the dwell-time was chosen to be long enough, several hundred milliseconds or longer (**Figure S3.1**), for precise determination of the FRET efficiency.

To visualize the FRET efficiency of each dye pair appearing in a single region of interest (ROI, highlighted as yellow circles in **Figure 3.1**b), we built a FRET kymograph (**Figure 3.1**d and **Figure S3.2**a). The kymograph shows the FRET efficiency per data point (**Figure 3.1**d, lines) and the mean FRET efficiency from all data points per binding event (**Figure 3.1**d, dots). The histograms built with the FRET efficiencies per data points (**Figure 3.1**d, middle panel) and per binding event (**Figure 3.1**d, bottom panel) show a single FRET population, indicating that imager binding is highly specific to the target site. The ensemble kymograph built from all 363 molecules for this construct shows a similar mean FRET and a standard deviation of 0.71 ± 0.01 (**Figure S3.2**b).

Analysis of complex biomolecules using single-molecule FRET requires the detection of multiple FRET pairs in a single object. One of the main benefits of FRET via DNA strands over conventional FRET measurements is to use a transiently binding DNA imager strand which can be exchanged by will. Each POI labelled with an orthogonal sequence for the imager binding can be sampled without any crosstalk by means of solution exchange. The absence of crosstalk between the POIs allows for accurate determination of the FRET efficiency of each POI. To illustrate this, we designed a ssDNA construct with two target sequences, each of which can interact with a donor-labeled imager strand for 2-3 seconds giving different FRET efficiencies (**Figure 3.2**a and **Figure S3.1**a and b). The distance between POI B and the acceptor is kept the same for both constructs (35 nt), but the distance for POI A is altered among the constructs (20 nt for **Figure 3.2**b and g and 25 nt for **Figure 3.2**c and h). When we used the same single imager strand for both POIs separated by a 5-nt spacer (**Figure 3.2**b), two FRET peaks were observed (**Figure 3.2**d), reporting on the location of each POI. However, when the two POIs were placed with no linker sequence in between (**Figure 3.2**c), the FRET histogram became unresolvable (**Figure 3.2**e). These results demonstrate that it is not feasible to determine the pair distances of several POIs with high precision using a single imager strand. It is noted that we used an experimental condition to test the resolving power of our approach by structurally compacting the target ssDNA molecule subjected in a buffer of high ionic strength, 100 mM $MgCl_2$ (**Figure S3.3**).

To achieve higher spatial resolution, we sought to detect the different POIs independently so that the overlapping FRET peaks can be obtained separately and fitted more precisely. As illustrated in **Figure 3.2**f, each POI was measured using a unique short DNA imager strand. After recording the binding events for the first POI for several minutes, the imager strand was exchanged by washing the microfluidic chamber and injecting a unique DNA imager strand for the second POI (**Figure 3.2**f). This process can be repeated for any number of POIs. We name this method FRET X for FRET via DNA eXchange.

To demonstrate the concept of FRET X, we measured POIs separated by a 5-nt thymine (**Figure 3.2**g) linker and POIs in closer proximity with no linker in between (**Figure 3.2**h) using two unique imager strands. In case of the 5-nt linker, in the

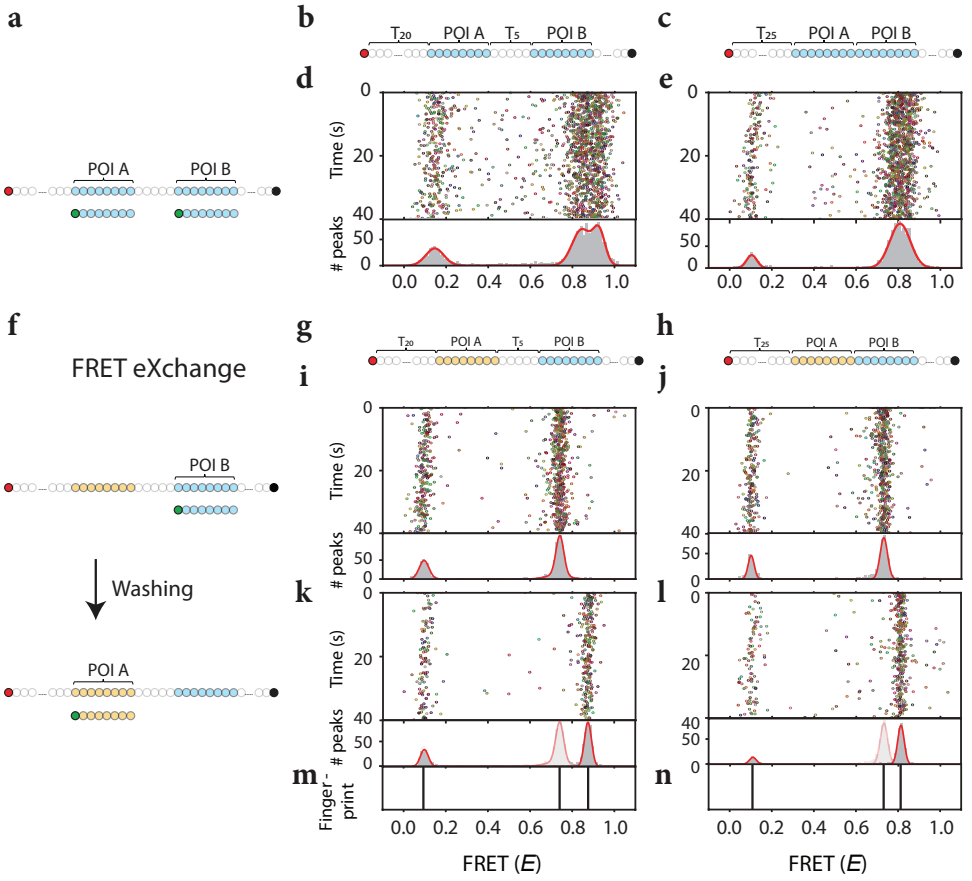


Figure 3.2: FRET by eXchange of unique imager strands allows for high spatial resolution of multiple POIs in a single nanoscale object.

a) Schematic representation of the single-molecule experiments with two target sequences. A single imager strand is used that can bind to both of the POIs in a target molecule. An acceptor (Cy5, red circle) labeled ssDNA construct contains two POIs. Binding of the donor (Cy3, green circle) labeled imager strand results in either high FRET (when binding to POI A) or mid FRET (when binding to POI B). **b** and **c)**Schematic representation of the target constructs in which two POIs are separated by a 5 Thymine linker (panel B) or in which the two POIs are directly connected to each other (panel C). The distance from the acceptor was kept the same for POI B (35 nt) among the two constructs, but was altered for POI A (20nt for panel B and 25nt for panel C). **(d)** Single--molecule kymograph of the ssDNA target construct from panel B. Top panel shows the binding events obtained for all molecules in a single field of view. Bottom panel shows a FRET histogram consisting of a donor only peak and two additional FRET peaks reporting on the location of each POI with respect to the acceptor fluorophore. **(e)** The single-molecule kymograph of the ssDNA target construct from panel C. Using the same imager strand for both POI does not allow for the detection of the position of both POIs when they are in close proximity. The FRET histogram shows a broad peak at 0.81. **(f)** Schematic workflow of FRET by eXchange of imager strand (or FRET X). A ssDNA target constructs consists of two POIs with unique DNA binding sequences, allowing us to measure the POIs one at a time. In a first round of detection, the imager strand for

POI B (blue circles) is added and imaged for 5 minutes. Then the microfluidic chamber is washed and an imager strand for POI A (orange circles) is added. **(g and h)** Schematic representation of the FRET X target constructs, in which two unique POI sequences blue circles (POI B) or orange circles (POI A) are separated by a 5 nt thymine linker (panel g) or in which the two POIs are directly adjacent (panel H). **(i and j)** Single-molecule kymographs for the FRET X for constructs in panels G and H. FRET X imaging allows for the determination of each POI in a separated round. In a first round of FRET X imaging we observe similar FRET efficiencies for POI B, 0.76 ± 0.05 (panel I) and 0.75 ± 0.01 (panel J). **(k and l)** Single-molecule kymographs for the second round of FRET X imaging of the constructs in panels G and H. **(k)** For a construct in which the POIs were separated by a 5 nt thymine linker we observed a FRET efficiency of resulted in a FRET efficiency of 0.87 ± 0.02 for POI A. **(l)** FRET X allows for the accurate detection of POIs even when they are in closer proximity. We observed a distinct FRET peak in the second FRET X imaging round for POI A of 0.81 ± 0.02 (panel l) and can be clearly separated from POI A (panel J, 0.75 ± 0.02). **(m and n)** The Gaussian fits of individual histograms for each POI obtained using the FRET X approach allows for the determination of the center of a peak with <0.005 precision. The center of the peaks are plotted in a separate panel, which we name this the FRET fingerprint of a nanoscale object. Mean FRET efficiencies and standard deviation are calculated from 3 independent experiments.

first round of FRET X detection we determined the FRET peak to be at 0.76 for POI B (**Figure 3.2**i). In the second round of FRET X imaging using the imager strand complementary to the POI A, we observed a single FRET peak at 0.87 reporting on the POI A (**Figure 3.2**k). As shown in **Figure 3.2**j and l, FRET X allows for the accurate detection of both POIs even when they are in closer proximity. We note that the conformation of the partially hybridized template strand is different between **Figure 3.2**d and e and **Figure 3.2**i-l due to the sequence difference of an unoccupied binding site, which consequently leads to slightly different FRET efficiencies.

Our FRET X approach allows for the detection of only a single POI for a prolonged time, until another imager strand is introduced. Therefore, while each histogram showed a wide distribution of ~0.05 (**Figure 3.2**i and j, the standard deviation) of the peak, the Gaussian fit can be used to resolve the center of a peak with high precision of <0.005 (standard error of mean), where the achievable precision depends on the number of binding events (**Figure S3.4**a and b). The resolved FRET values for each POI are plotted as the FRET fingerprint of the measured object (**Figure 3.2**m and n).

Structural analysis of complex biomolecules using single-molecule FRET requires the detection of multiple FRET pairs in a single object. To demonstrate the potential of FRET X, we designed a DNA nano-structure consisting of two POIs and tested whether FRET X can obtain distance information for each POI in a single object. The DNA nano-structure is in a triangular shape that consists of an acceptor reference point, and a POI is placed at each corner of the triangle (**Figure 3.2**a). To avoid the photobleaching of the acceptor dye, we designed a unique sequence near the 3' end of the construct where a complementary acceptor-labeled imager strand can transiently bind. To increase the probability of energy transfer between donor and acceptor fluorophores, the acceptor imager strand was designed to have a higher binding rate and lower dissociation rate than the donor imager strands.[21,22] We estimated the time-dependant FRET detection rates for both static and dynamic acceptor strand. The static acceptor showed a faster decrease in the FRET detection rate due to photobleaching (**Figure S3.6**).

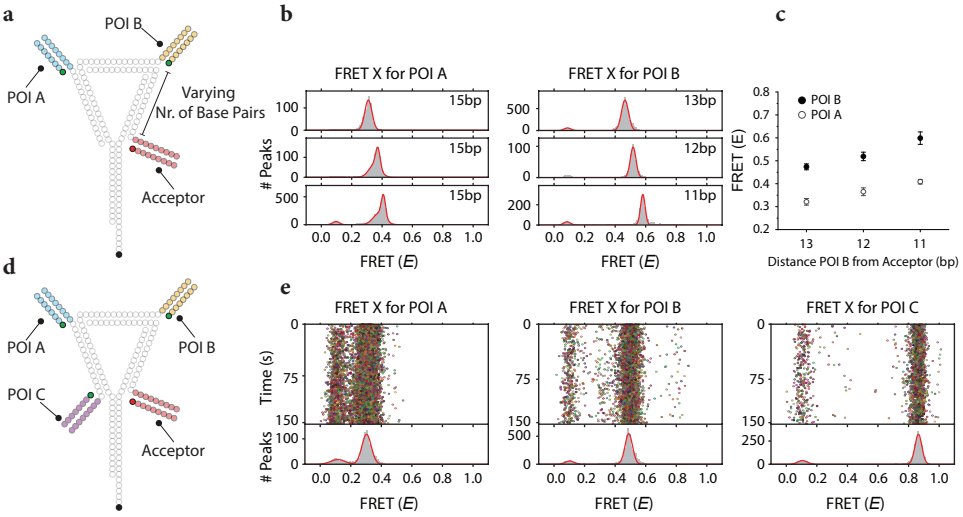


Figure 3.3: Structural analysis of a complex DNA nanostructure using FRET X.
(a) Schematic representation of the dsDNA nanostructure used for the determination of several POIs in a single molecule. The DNA nanostructure consist of 2 POIs, one of which is fixed (POI A, blue circles) at 15 bp apart from the acceptor imager binding site. The second POI (POI B, orange circles) is separated by different linker lengths from the acceptor. The acceptor (Cy5)-labeled imager strand binds transiently to a unique binding site (red circles) to avoid photobleaching of the acceptor fluorophore. **(b)** FRET X histograms of the different POIs in the DNA nanostructure. For a DNA nanostructure with a 13 bp linker between POI B and the acceptor we observes a FRET efficiency of 0.31 ± 0.01 and 0.47 ± 0.02, for POI A and POI B, respectively (panel B, top row). Next, by only decreasing the linker length with steps of 1 bp between POI B and the acceptor, we observed an increase in FRET efficiency for POI B (0.52 ± 0.01 and 0.60 ± 0.02 for 12 bp and 11 bp, respectively). Furthermore, the FRET efficiency for POI A and the acceptor increases when the linker between POI B and acceptor is shorter, which hints global distortion of the nano-structure due to the shortening of one side of the triangle (left panel). **(c)**The mean FRET X efficiency for POI A (open circles) and POI B (solid circles) determined on different days. Mean FRET efficiencies and standard deviation are calculated from 3 independent experiments. **(d)** Schematic representation of a dsDNA nanostructure with 3 POIs. A third POI is added close to the acceptor binding site. **(e)** Kymographs obtained for each POI of the dsDNA nanostructure. In a first round of FRET X imaging we observed a FRET efficiency of 0.27 ± 0.01 for POI A. The second round of FRET X imaging resulted in a FRET efficiency of 0.47 ± 0.02 for POI B. In a final round of FRET X imaging, we observed a FRET efficiency of 0.86 ± 0.01 for POI C in the DNA nanostructure. Mean FRET efficiencies and standard deviation are calculated from 3 independent experiments.

In a first round of FRET X, we determined the FRET efficiency for POI A that is separated by a 15-bp linker from the acceptor and observed a distinct FRET peak at 0.31 (**Figure 3.3**b, top-left panel). Next, we washed the chamber and injected the imager strand for POI B and observed a FRET peak at 0.47 when POI B is 13-bp away from the acceptor (**Figure 3.3**b, top-right panel).

To determine the resolution of FRET X for the detection of multiple POIs in a nanoscale object, we changed the length between the acceptor and POI B by a step

of 1 bp. For each construct, we determined the FRET efficiency for both POIs and observed a clear change in FRET for POI B (**Figure 3.3**b, right panels and **Figure 3.3**c, solid circles). Furthermore, the FRET efficiency for POI A and the acceptor increased when the linker between POI B and acceptor was shorter (**Figure 3.3**b, left panels and **Figure 3.3**c, open circles).

To further demonstrate the ability of FRET X for the detection of multiple POIs, we added a third POI to the triangular DNA structure (**Figure 3.3**d). POI C was introduced at a site close to the acceptor reference point and gives a high FRET value (**Figure 3.3**e, right panel). For POI A and B we observed similar FRET efficiencies compared to their location in the structure with only 2 POIs (**Figure 3.3**e left and middle panels, and **Figure 3.3**b top panels).

To investigate the ultimate resolution of FRET X, we designed a series of ssDNA constructs in which the position of the donor imager binding site is altered by only a single nucleotide among the different imager strands (**Figure 3.4**a). The FRET X cycle was then repeated for all nine imager strands. The center of a peak of each histogram was determined by fitting with a single Gaussian function. The obtained fingerprint showed nine separated peaks, one for each donor-labeled nucleotide (**Figure 3.4**b

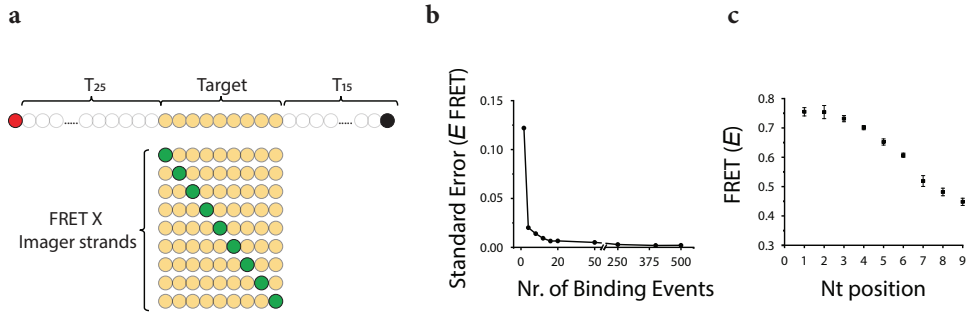


Figure 3.4: Single nucleotide resolution can be achieved with FRET X.
(a) Schematic representation of the single-molecule constructs used for the determination of different POIs separated by a single base pair. An acceptor (Cy5, red circle) labeled ssDNA target construct consisting of a 9 nt target sequence (orange circles) where each imager strand can bind. A series of donor (Cy3, green circles) labeled FRET X imager strands. The position of each POI (or nucleotide) in the target sequence will be determined one by one using our FRET X approach. **(b)** Standard error of the FRET X efficiency for imager strand 5 (**Figure S3.5**e) vs the number of binding events. We observe that we can determine the center of a Gaussian fit with a FRET X precision of ΔE ~0.01 after >10 binding events. **(c)** The mean FRET X efficiency for each of the POIs determined on different days. We find good reproducibility for FRET X. Mean FRET efficiencies and standard deviation are calculated from 3 independent experiments

and **Figure S3.5**), indicating that FRET X has a single-nucleotide resolution.

To determine the precision that can be obtained using our FRET X approach, the standard error of the FRET efficiency was plotted as a function of the number binding events. The chosen events were from an imager strand labeled at position 5 (**Figure S3.5**e) that yielded a FRET efficiency value of 0.65. Given our photon count rate of 5000 s-1 and the binding dwell time of 2 s (**Figure S3.1**a), we expect the theoretical limit[23] of the precision in FRET determination was ~0.005 (**Figure S3.4**c and d). Experimentally, however, we found that the center of a Gaussian fit can be determined with a precision of ΔE ~0.01 after obtaining >10 binding events (**Figure 3.4**b) due to other noise contributions such as electronic shot noise and backgrounds, stray light, and an uneven illumination profile. The reproducibility of FRET X was demonstrated by measuring all nine labeled imager strands on different days. As shown in **Figure 3.4**c, the standard deviation between the measurements made on different days is about 0.02 for each construct.

Finally, having the high-resolution analysis of different POIs in a nanoscopic object without photobleaching problems, we speculated that FRET X can be used reliably for population analysis at the single-molecule level, which requires repeated sampling of individual targets. To demonstrate the potential use of FRET X for population analysis at the single-molecule level, we designed two ssDNA constructs with structural differences and tested whether individual molecules can be distinguished when the two are mixed. The ssDNA constructs consist of two POIs, one of which is located at an identical position on the two DNA constructs. The second POI is connected to the side of one of the nucleotides in the backbone sequence and has a different location on the two constructs (**Figure 3.5**a and **Figure S3.7** and **Figure S3.8**). To avoid the photobleaching of the acceptor dye, we designed a unique sequence near the 3' end of the construct where a complementary acceptor labeled imager strand can transiently bind. We immobilized a mixture of the two constructs in a 1:1 ratio.

In a first round of FRET X, we determined the FRET efficiency for POI A and observed two distinct FRET populations reporting on the distinct distance between POI A and the acceptor reference point for the two different constructs (**Figure 3.5**b

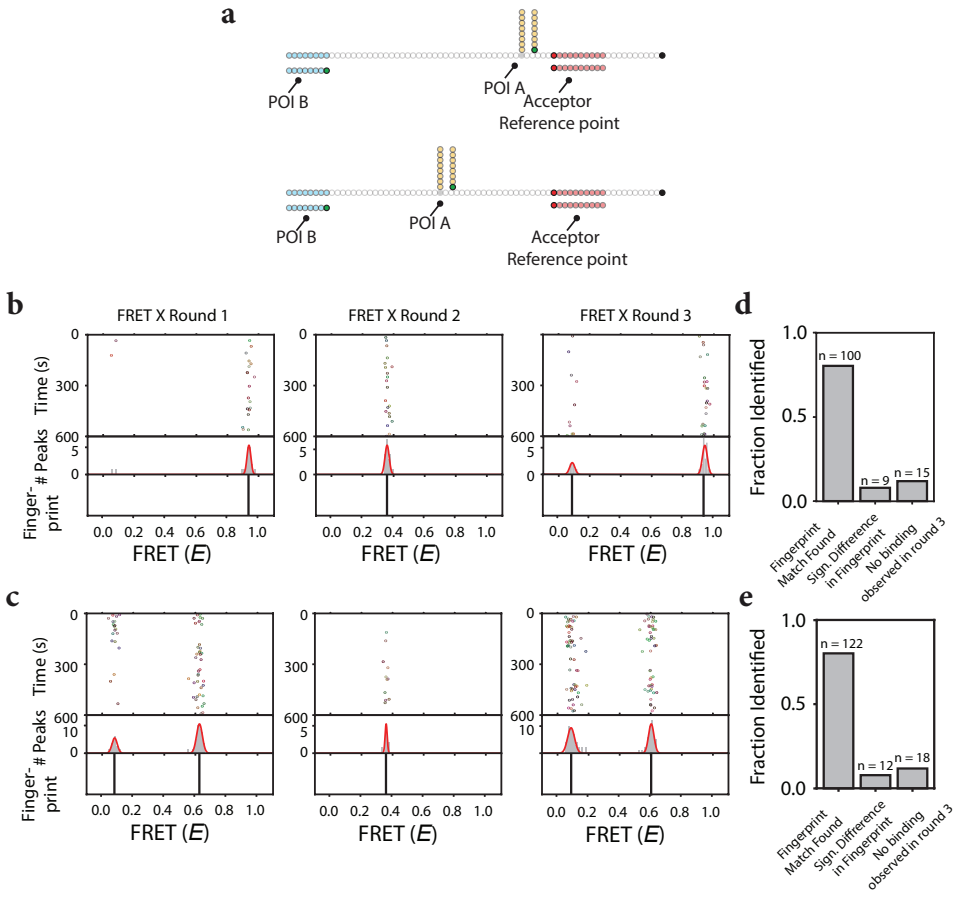


Figure 3.5: Population analysis on the individual molecule level using FRET X.
(a) Schematic representation of the DNA constructs used for population analysis. The ssDNA construct contains two POIs, of which one is fixed and has the same location relative to the acceptor on both constructs. The second POI is connected to the side chain of one of the nucleotides in the backbone sequence and has a different location on both constructs. **(b and c)** Kymographs of individual molecules obtained for an equal mixture of the ssDNA constructs immobilized on the slide surface. The FRET X cycle consisted of 3 rounds. In a first round of the FRET X cycle, we determined the unique fingerprint of POI A among the different constructs and observe a FRET efficiency of 0.94 for the high-FRET construct (panel b, left kymogrpah) and 0.63 for the medium-FRET construct (panel c, left kymograph). The second round of the FRET X cycle resulted in a single peak obtained from FRET between POI B and the acceptor, which is identical in both constructs (panels b and c, middle kymographs). In the last round of the FRET X cycle (panels b and c, right kymogrpahs) we confirmed the location of POI A and observed the same FRET peaks as in round 1. **(d and e)** Bar plots showing the fractions of fingerprint matches and non-matches for individual molecules that were identified as high- (panel d) or medium-FRET construct (panel e). We determined the mean FRET efficiency of the medium- or high-FRET fingerprint in round 1 and compared this with a detection uncertainty of ΔE ~0.07 with round 3 to find positives matches. The majority of molecules were identified identically between round 1 and 3, for the high- (panel d) and medium-(panel e) FRET ssDNA constructs.

and c, left panels). Next, we washed the microfluidic chamber and injected the imager strand for POI B. As expected, we observed a single peak for POI B, reporting on the same position of POI B for both constructs (**Figure 3.5**b and c, middle panels). In a final round of FRET X, we confirmed the location of POI A by injecting the imager strand for POI A back and observed the same FRET peaks as in the FRET X imaging round 1 (**Figure 3.5**b and c, right panels).

For each individual molecule, we determined the mean FRET efficiency for POI A in round 1 and 3 compared this with the FRET efficiency obtained for POI A in round 3 (**Figure S3.9**). The majority (>80 %) of the individual molecules in the mixture had a similar resolved FRET efficiency of POI A between rounds 1 and 3, for the high- (**Figure 3.5**d) or medium- (**Figure 3.5**e) FRET constructs. Only a small fraction of molecules did not show a match between the FRET X rounds due to a different resolved FRET efficiency for POI A or a lack of imager strand binding events (**Figure 3.5**d and e). Altogether, these results show that the FRET X method is capable of detecting the populations of individual DNA constructs at the single-molecule level.

3.4 Discussion

Here we present a proof-of-concept for FRET X, a novel tool for the detection of several FRET pairs in a single object, which can be used for the structural analysis of biomolecules. Our FRET X technique relies on the dynamic binding of fluorescently labeled short oligos to complementary docking sequences on a target object. Conventional single-molecule FRET techniques report on the changes in distance between a single dye pair on a single molecule. In contrast, FRET X uses orthogonal imager strands for different POIs which allows us to separate the detection in time, and consequently detect a large number of POIs on a single object. Both switchable FRET and FRET X uses a stochastic on-off method. However, unlike switchable FRET, FRET X allows for probing one, only one location, for a prolonged time, until another imager strand is introduced by an operator. Therefore, we can collect higher precision data for each location. We note that our FRET X technique can be integrated with another recently developed multiplexed FRET barcode technique[24], which allows simultaneous observation of multiple orthogonal probes, reducing the total measurement time.

Single-molecule FRET has recently been combined with super-resolution imaging using DNA-PAINT to allow for faster acquisition[21,25] and multiplexing based on FRET efficiency.[22] While FRET X is designed to report on position and distance information of different POIs in a single object, we envision that it can also be used to improve the resolution of current DNA-PAINT technologies.

3.5 Materials and Methods

3.5.1 Single-Molecule Setup

All experiments were performed on a custom-built microscope setup. An inverted microscope (IX73, Olympus) with prism-based total internal reflection was used. In combination with a 532 nm diode-pumped solid-state laser (Compass 215M/50mW, Coherent). A 60x water immersion objective (UPLSAPO60XW, Olympus) was used for the collection of photons from the Cy3 and Cy5 dyes on the surface, after which a 532 nm long pass filter (LDP01-532RU-25, Semrock) blocks the excitation light. A dichroic mirror (635 dcxr, Chroma) separates the fluorescence signal which is then projected onto an EM-CCD camera (iXon Ultra, DU-897U-CS0-#BV, Andor Technology). A series of EM-CDD images was recorded using custom-made program in Visual C++ (Microsoft).

3.5.2 Single-Molecule Data Acquisition

Single-molecule flow cells were prepared as previously described.[20,26] In brief, to avoid non-specific binding, quartz slides (G. Finkerbeiner Inc) were acidic piranha etched and passivated twice with polyethylene glycol (PEG). The first round of PEGylation was performed with mPEG-SVA (Laysan Bio) and PEG-biotin (Laysan Bio), followed by a second round of PEGylation with MS(PEG)4 (ThermoFisher). After assembly of a microfluidic chamber, the slides were incubated with 20 µL of 0.1 mg/mL streptavidin (Thermofisher) for 2 minutes. Excess streptavidin was removed with 100 µL T50 (50mM Tris-HCl, pH 8.0, 50 mM NaCl). Next, 50 µL of 75 pM Cy5 labeled ssDNA was added to the microfluidic chamber. After 2 minutes of incubation, unbound ssDNA was washed away with 100 µL T50. For experiments in Figure 1, 50 µL of 10 nM donor labeled imager strands in imaging buffer (50 mM Tris-HCl, pH 8.0, 500 mM NaCl, 0.8 % glucose, 0.5 mg/mL glucose oxidase (Sigma), 85 ug/mL catalase (Merck) and 1 mM Trolox (Sigma)) was injected. All single-molecule FRET experiments were performed at room temperature (23 ± 2 °C).

3.5.2.1 FRET X experiments (Figures 3.2, 3.3 and 3.4)

The triangular DNA nanostructures were annealed in 50 mM Tris (pH 8.0), 100 mM NaCl and 1 mM $MgCl_2$ using a thermocycler (Bio-Rad) at –1 °C/cycle for 5 min/cycle from 80 °C to 4 °C and then store at 4 °C. For FRET X imaging in Figures 2, 3 and 4, 50 µL of 75 pM target DNA strands were immobilized and the unbound DNA was washed away with 100 µL T50 after 2 minutes of incubation. Next, an imaging buffer containing the imager strand for POI A (**Figure 3.2** and **Figure 3.3**), or imager strand with internal nucleotides labeled at position 1 (**Figure 3.4**), was injected. After obtaining 2000 frames at 100 ms exposure time, the microfluidic chamber was washed with 1000 µL T50 and the imager strand for POI B (**Figure 3.2** and **Figure 3.3**), or internally labeled nucleotide position 2 (**Figure 3.4**), was injected. This cycle was repeated until all internally labeled nucleotides were measured for **Figure 3.4**.

3.5.2.2 FRET X experiments for Single molecule Population Analysis (Figure 5)

For buffer exchange and imaging of the same molecules in a single field of view for different rounds of FRET X imaging, tubing was connected to the inlet and outlet of the microfluidic chamber. One of the tubes was connected to a buffer reservoir and the other was connected to a syringe. By gently pulling on the syringe, the washing buffers and imaging solutions were exchanged without perturbing the sample stage.

For the branched DNA constructs experiments, 50 µL of 75 pM branched DNA target strand was immobilized for 2 minutes and unbound DNA was removed with 100 µL T50. For long term acquisition, a 50 µL imaging solution consisting of 100 nM acceptor imager strand and 10 nM of donor labeled imager strand for POI A was injected and the chamber was imaged for 15 minutes at 100 ms exposure time. Then the imaging solution for POI A was removed by washing with 1000 µL T50 and the imaging solution of POI B was added (50 µL of 100 nM acceptor imager strand and 10 nM of imager strand for POI B in imaging buffer). After this second round of imaging, the microfluidic chamber was washed with 1000 µL T50 and POI A was imaged again by injecting fresh imaging solution for POI A.

3.5.3 Data Analysis

CCD images were analyzed using a custom code written in IDL (ITT Visual Information Solution) to find the position of individual FRET pairs and to extract fluorescence time traces. When the same field of view is measured multiple times (**Figure 3.5**), drift correction between the measurements and trace extraction were performed by a custom-built code written in Python (Python 3.7). For visualization of single molecule fluorescence and FRET time traces, we used a custom code written in Matlab (Mathworks). For automated detection of individual fluorescence imager strand binding events, we used a custom Python code (Python 3.7) utilizing a two-state K-means clustering algorithm on the sum of the donor and acceptor fluorescence intensities of individual molecules to identify the frames with high intensities.[24] To avoid false positive detections, only binding events that lasted for more than three consecutive frames were selected for further analysis. FRET efficiencies for each imager strand binding event were calculated and used to build the FRET kymograph and histogram. Populations in the FRET histogram are automatically classified by using Gaussian mixture modeling and used to determine the presence of specific points of interest. The automated analysis code in Python is freely available at (https://github.com/kahutia/transient_FRET_analyzer2).

3.6 Supporting Information

3.6.1 Supporting Figures

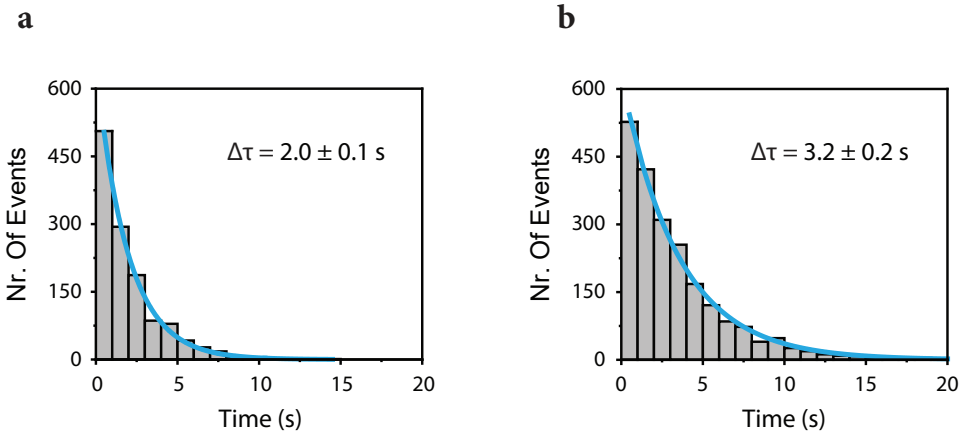


Figure S3.1: Single-molecule binding kinetics of FRET X imager strands.
a) Dwell-time histogram for the FRET X imager strand used for the detection of a POI in ssDNA target used in **Figure 3.1** and for the detection of POI B in **Figure 3.2.** Maximum likelood estimation gives 2.0 ± 0.1 s for a single exponential distribution (blue line). The number of datapoints: n = 1252. **b**) Dwell-time histogram for the FRET X imager strand used for the detection of POI A in Figure 2. Maximum likelood estimation gives 3.2 ± 0.2 s for a single exponential distribution (blue line). The number of datapoints: n = 2146.

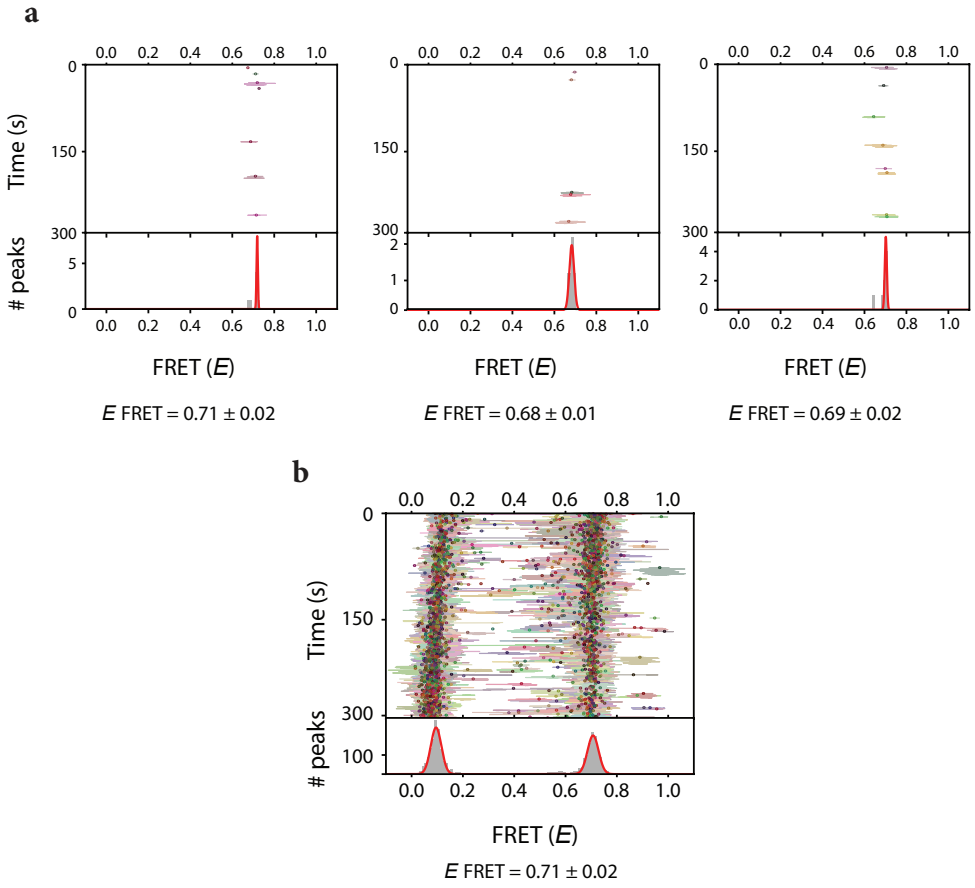


Figure S3.2: Single-Molecule FRET Kymographs.
a) Representative FRET kymographs obtained from individual molececule in a single field of view. **b**) An ensemble FRET kymograph obtained from all molecules in a single field of view. The FRET efficiencies were reported as the mean ± the standard deviation.

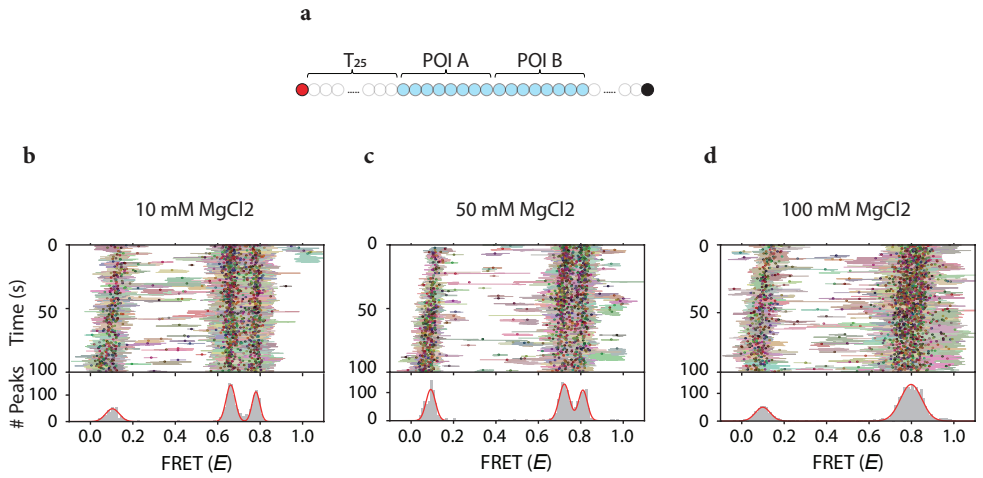


Figure S3.3: Effect of $MgCl_2$ concentration on the compactness of ssDNA.

a) Schematic representation of the ssDNA construct used for the determination of the effect of $MgCl_2$ on the compactness of the ssDNA construct using the same imager strand. **b-d**) Single molecule kymographs obtained with different concentrations of $MgCl_2$. At a lower concentrations of $MgCl_2$ (panel a and b) the two POIs can be clearly resolved. However, when using 100 mM $MgCl_2$ the histogram become unresolvable, which is an ideal condition to test the resolving power of FRET X.

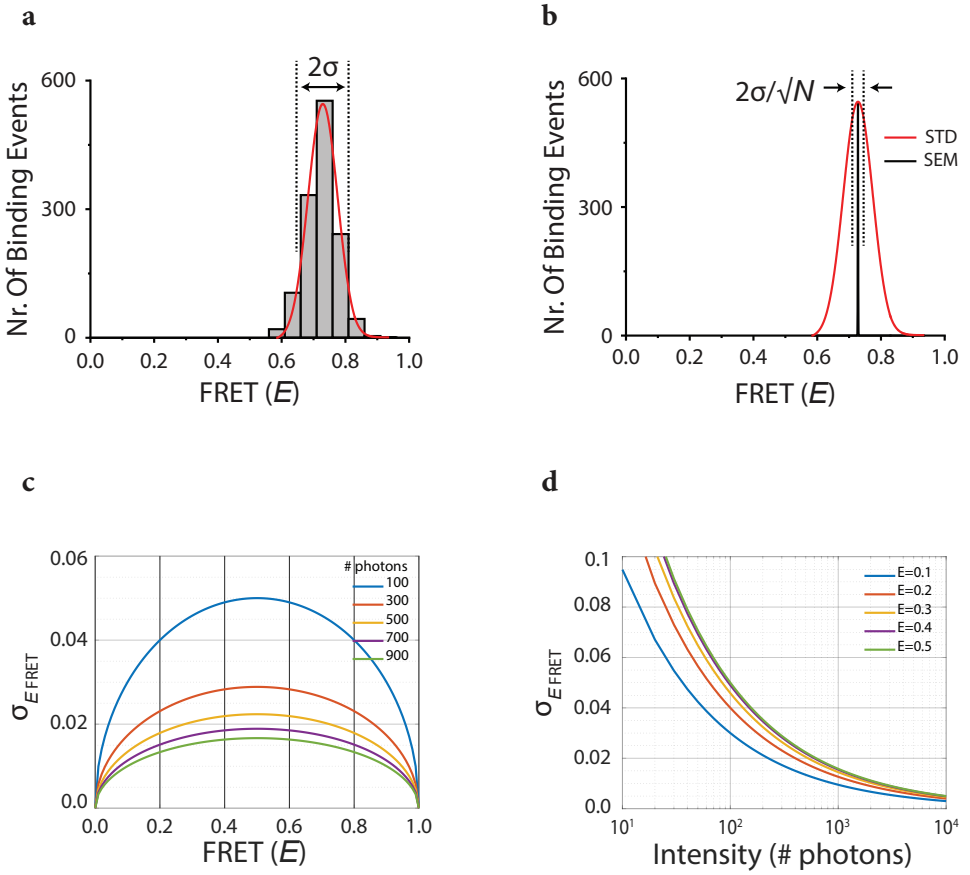


Figure S3.4: Standard deviation vs Standard error.

a) The standard deviation (σ) reports on the intrinstic broadness of a FRET histogram. **b**) The standard error measures the accuracy of determing the center of a peak. N is the number of imager strand binding events. **c-d**) The photon shot noise-based theoretical error limit of FRET at various total number of photons (sum of the donor and acceptor) and FRET efficiency. The FRET error is given by:

$$\sigma(E) = \frac{1}{N_D + N_A}\sqrt{E^2 \times \sigma^2(N_D) + (1-E)^2 \times \sigma^2(N_A)}$$

where E is the FRET efficiency, N_D and N_A are the number of photons in the donor and acceptor signals, respectively. σ^2 denotes the variance (Holden et al. (2010)).[23]

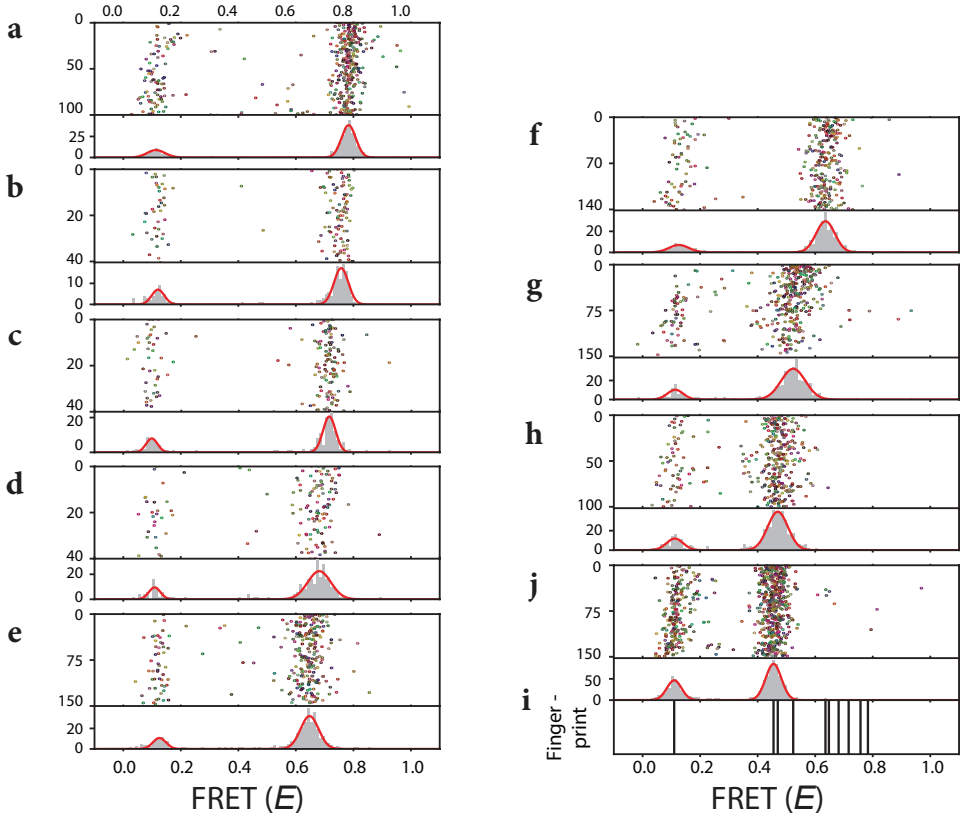


Figure S3.5: Kymographs for the FRET X precision on a ssDNA target.

a-j) Kymographs for each of the POIs determined using FRET X. The top kymograph (**a**) is obtained with the imager strand where the donor fluorophore binds closest to the acceptor (separated by a 25-thymine linker). Each next kymograph is obtained with a subsequent imager strand where the distance to the acceptor increases by a single basepair. We obtained nine separated FRET histogram, one for each of donor labeled base pairs using FRET X. **k**) We observed nine clearly separated peaks in the FRET fingerprint. The fingerprint show the center of each Gaussian fit that obtained using our FRET X approach.

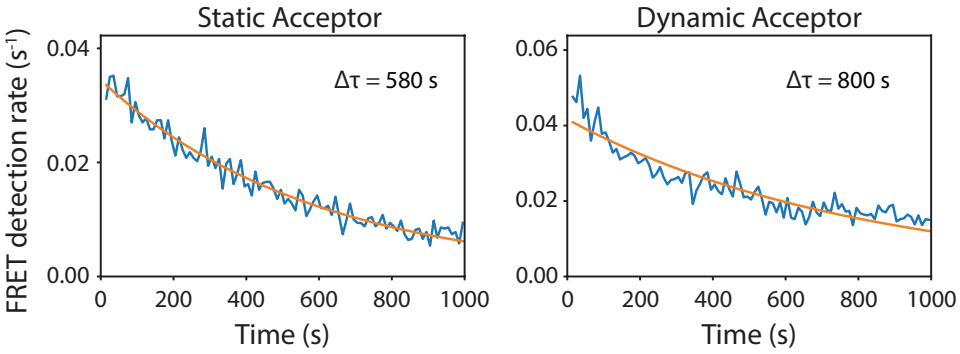


Figure S3.6: Comapison of the dyanamic and static acceptor.

The time-dependent FRET detection rates per POI were calculated with time windows of 10 s from 500 target DNA molecules under constant illumination. The static acceptor showed faster decrease in the FRET detection rate due to photo bleaching.

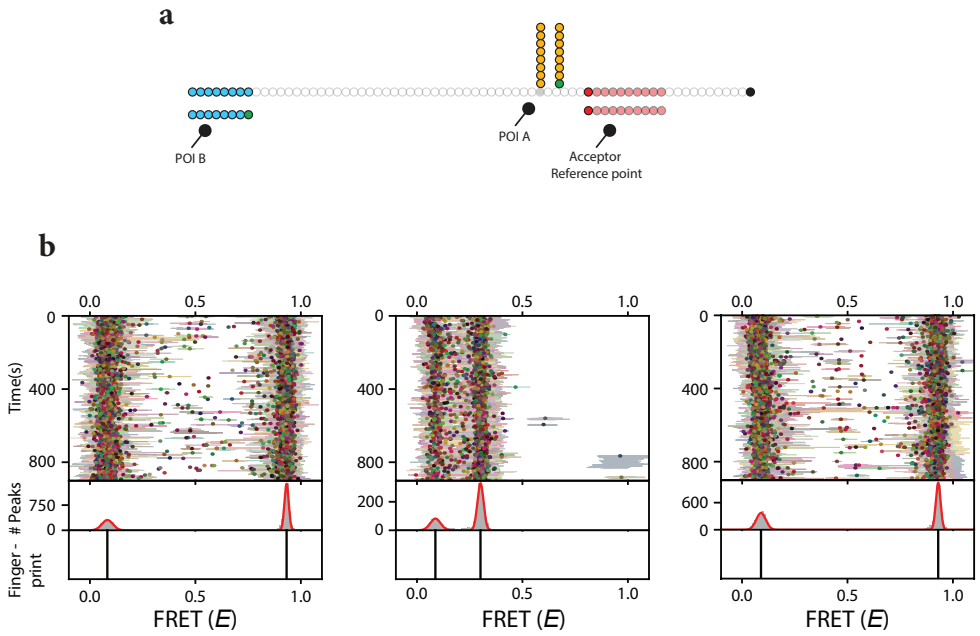


Figure S3.7: Single-Molecule FRET X analysis of complex ssDNA structure resulting in high FRET.

a) Schematic representation of complex ssDNA structure resulting in high FRET. Upon binding of the FRET X imager strand for POI A, the donor fluorophore is separated by a 5 nt polyT linker from the acceptor binding site. The imager strand for POI B is separated by a 40 nt polyT linker from the acceptor binding site. **b**) Ensemble FRET kymographs obtained from different rounds of FRET X imaging. In a first round of imaging (left panel) we obtained a high FRET peak reporting on the location of POI A relative to the acceptor biding site. After washing of the microfluidic cell we injected the imager strand for POI B (middle panel) and obtained a low FRET efficiency reporting on the distance of POI B to the acceptor binding site. In a last round of FRET X imaging (right panel) we confirmed the high FRET peak for POI A relative to the acceptor binding site.

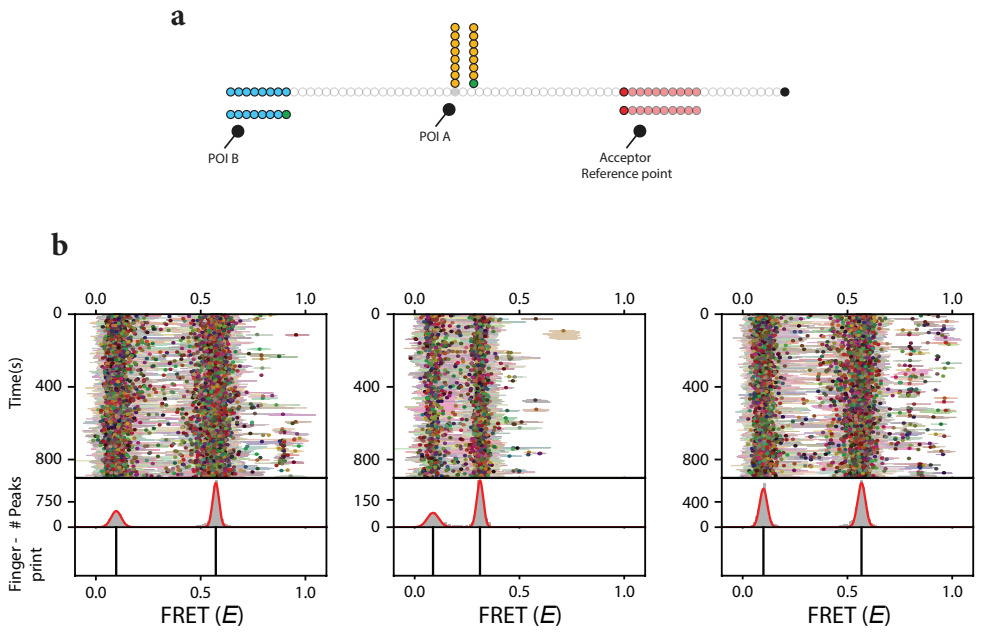


Figure S3.8: Single-Molecule FRET X analysis of complex ssDNA structure resulting in medium FRET.

a) Schematic representation of complex ssDNA structure resulting in medium FRET. Upon binding of the FRET X imager strand for POI A, the donor fluorophore is separated by a 20 nt polyT linker from the acceptor binding site. The imager strand for POI B is separated by a 40 nt polyT linker from the acceptor binding site. **b**) Ensemble FRET kymographs obtained from different rounds of FRET X imaging. In a first round of imaging (left panel) we obtained a medium FRET peak reporting on the location of POI A relative to the acceptor biding site. After washing of the microfluidic cell we injected the imager strand for POI B (middle panel) and obtained a low FRET efficiency reporting on the distance of POI B to the acceptor binding site. In a last round of FRET X imaging (right panel) we confirmed the medium FRET peak for POI A relative to the acceptor binding site.

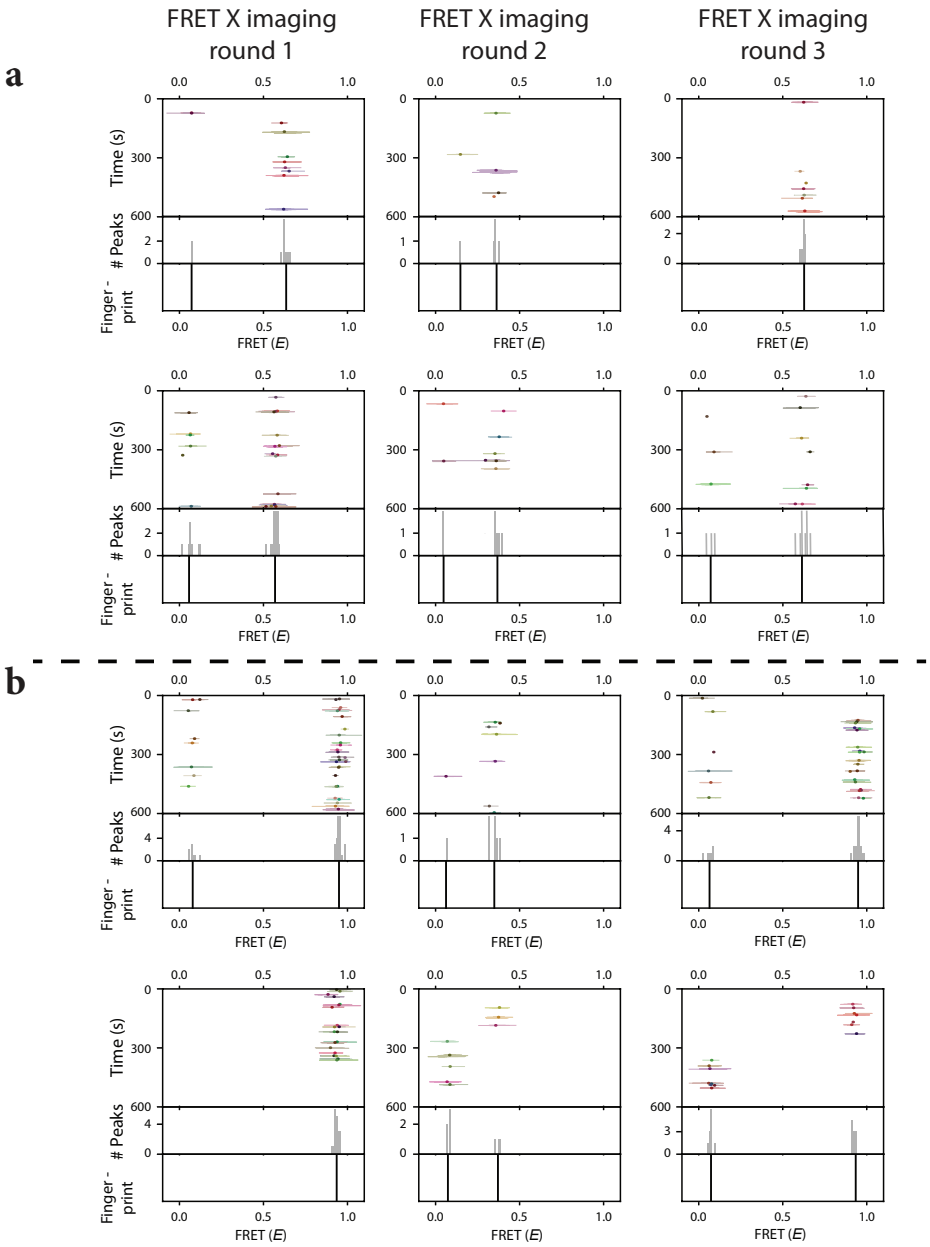


Figure S3.9: Representative kymographs of individual molecules in a mixture of structurally similar DNA constructs.

a-b) Representative FRET kymographs from individual complex ssDNA molecules in a mixture. Using our FRET X approach we can observe a difference between the medium FRET complex structure (**Figure S3.7**) or high FRET complex structure (**Figure S3.8**), in FRET X imaging round 1 (left Column) and round 3 (right column). The second round of FRET X imaging (middle column) shows a similar FRET peak for both constructs, reporting on the structural similarity among the constructs.

Table S3.1: Single-Molecule DNA constructs

DNA Strand	Nucleotide Sequence (5' → 3')	Modification
Figure 1 single molecule DNA target construct	TTTTT TTTTT TTTTT TTTTT TTTTT ATACA TCTAT TTTTT TTTTT TTTTT TTTTT TTTTT TTTTT	Cy5 - 5' end labeled, biotin - 3' end labeled
Figure 2B – DNA FRET target construct 5 nt spacing	TTTTT TTTTT TTTTT TTTTT ATACA TCTAT TTTTT ATACA TCTAT TTTTT TTTTT TTTTT	Cy5 - 5' end labeled, biotin - 3' end labeled
Figure 2C – DNA FRET target construct no spacing	TTTTT TTTTT TTTTT TTTTT TTTTT ATACA TCTAT ATACA TCTAT TTTTT TTTTT TTTTT	Cy5 - 5' end labeled, biotin - 3' end labeled
Figure 2G – FRET X target construct 5 nt spacing	TTTTT TTTTT TTTTT TTTTT ATACA TCTAT TTTTT TCTTC ATTAC TTTTT TTTTT TTTTT	Cy5 - 5' end labeled, biotin - 3' end labeled
Figure 2H – FRET X target construct no spacing	TTTTT TTTTT TTTTT TTTTT TTTTT ATACA TCTAT TCTTC ATTAC TTTTT TTTTT TTTTT	Cy5 - 5' end labeled, biotin - 3' end labeled
Imager strand 1	AGATGTAT	Cy3 - 3' end labeled
Imager strand 2	TAATGAAGA	Cy3 - 3' end labeled
Figure 4 – FRET X target construct single nt experiments	TTTTT TTTTT TTTTT TTTTT TTTTT AGAAGTAATG TTTTT TTTTT TTTTT	Cy5 - 5' end labelled, biotin - 3' end labeled
Figure 4 – Imager strand position 1	ATTACTTC**T**	Cy3 – internal labeled dT
Figure 4 – Imager strand position 2	ATTACTT**C**T	Cy3 – internal labeled dT
Figure 4 – Imager strand position 3	ATTACT**T**CT	Cy3 – internal labeled dT
Figure 4 – Imager strand position 4	ATTAC**T**TCT	Cy3 – internal labeled dT
Figure 4 – Imager strand position 5	ATTA**C**TTCT	Cy3 – internal labeled dT
Figure 4 – Imager strand position 6	CATT**A**CTTCT	Cy3 – internal labeled dT
Figure 4 – Imager strand position 7	AT**T**ACTTCT	Cy3 – internal labeled dT
Figure 4 – Imager strand position 8	A**T**TACTTCT	Cy3 – internal labeled dT
Figure 4 – Imager strand position 9	**A**TTACTTCT	Cy3 – internal labeled dT

DNA Strand	Nucleotide Sequence (5' → 3')	Modification
Branched structure 1	TTTCA ATGTA TTTTT TTTTT TTTTT T TTTT***X*** TTTTT TTTTT TTTTT TTTTT TCTTC ATTAC TATCT ACATA TTTTT	3' Biotin, internal branch to nt ***X***, branch sequence 5'-TATACATC-TAT-3'
Branched structure 2	TTTCA ATGTA TTTTT TTTTT TTTTT TTTTT TTTTT TTTTT TTTTT X TTTTT TCTTC ATTAC TATCT ACATA TTTTT	3' Biotin, internal branch to nt ***X***, branch sequence 5'-TATACATC-TAT-3'
Imager strand 3 (POI B branched structures)	TACATTGAA	Cy3 – 5' end labeled
Imager strand 4 (acceptor for branched structure)	GTAATGAAGA	Cy5 – 3' end labeled
DNA Nanostructure backbone + POI B	ATTCA TTCTC ATCCT CTGTC GGGTG TACCG TAAGG TGAAT AGTGA CTTTA TACAT CTA	
DNA Nanostructure left arm + POI A	AGAGG AGGAT TTCGG TACAC CCGAC AG	
DNA Nanostructure imager strand for POI A	TCCTCCT	Cy3 – 5' end labeled
DNA Nanostructure imager strand for POI B	AGATGTAT	Cy3 - 3' end labeled
DNA Nanostructure biotin strand for 13 bp linker	CTGAT TGTTA TCGAG GATGA GAATG AATTT TTTTT TTTTT TTT	Biotin – 3'end labeled
DNA Nanostructure right arm + acceptor for 13bp linker	TCTTC ATTAC TTTTC GATAA CAATC AGGTC ACTAT TCACC TTA	
DNA Nanostructure biotin strand for 12 bp linker	CTGAT GTTAT CGAGG ATGAG AATGA ATTTT TTTTT TTTTT	Biotin – 3'end labeled
DNA Nanostructure right arm + acceptor for 12 bp linker	TCTTC ATTAC TTTTC GATAA CATCA GGTCA CTATT CACCTTA	
DNA Nanostructure biotin strand for 11 bp linker	CTGAT GTTAT GAGGA TGAGA ATGAA TTTTT TTTTT TTTTTT	Biotin – 3'end labeled

DNA Strand	Nucleotide Sequence (5' → 3')	Modification
DNA Nanostructure right arm + acceptor for 11 bp linker	TCTTC ATTAC TTTTC ATAAC ATCAG GTCAC TATTC ACCTTA	
DNA Nanostructure transient acceptor imager strand	AGTAATGAA	Cy5 – 5' end labeled
DNA Nanostructure Left arm + POI A and POI C	AGAGG AGGAT TTCGG TACAC CCGAC AGTTT TCAAT GTA	
DNA Nanostructure imager strand POI C	TACATTGA	Cy3 - 3' end labeled

3.7 References

1 Shi, Y. A glimpse of structural biology through X-ray crystallography. Cell 159, 995–1014 (2014).

2 Nogales, E. & Scheres, S. H. W. Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. Mol. Cell 58, 677–689 (2015).

3 Henzler-Wildman, K. A. et al. Intrinsic motions along an enzymatic reaction trajectory. Nature 450, 838–844 (2007).

4 Algar, W. R., Hildebrandt, N., Vogel, S. S. & Medintz, I. L. FRET as a biomolecular research tool — understanding its potential while avoiding pitfalls. Nat. Methods 16, 815–829 (2019).

5 Lerner, E. et al. Toward dynamic structural biology: Two decades of single-molecule förster resonance energy transfer. Science (80-.). 359, (2018).

6 Hohng, S., Joo, C. & Ha, T. Single-Molecule Three-Color FRET. Biophys. J. 87, 1328–1337 (2004).

7 Clamme, J. P. & Deniz, A. A. Three-color single-molecule fluorescence resonance energy transfer. ChemPhysChem 6, 74–77 (2005).

8 Kalinin, S. et al. A toolkit and benchmark study for FRET-restrained high-precision structural modeling. Nat. Methods 9, 1218–1225 (2012).

9 Hellenkamp, B., Wortmann, P., Kandzia, F., Zacharias, M. & Hugel, T. Multidomain structure and correlated dynamics determined by self-consistent FRET networks. Nat. Methods 14, 176–182 (2017).

10 Peulen, T. O., Opanasyuk, O. & Seidel, C. A. M. Combining Graphical and Analytical Methods with Molecular Simulations to Analyze Time-Resolved FRET Measurements of Labeled Macromolecules Accurately. J. Phys. Chem. B 121, 8211–8241 (2017).

11 Craggs, T. D. & Kapanidis, A. N. Six steps closer to FRET-driven structural biology. Nat. Methods 9, 1157–1159 (2012).

12 Uphoff, S. et al. Monitoring multiple distances within a single molecule using switchable FRET. Nat. Methods 7, 831–836 (2010).

13 Giannone, G. et al. Dynamic superresolution imaging of endogenous proteins on living cells at ultra-high density. Biophys. J. 99, 1303–1310 (2010).

14 Schoen, I., Ries, J., Klotzsch, E., Ewers, H. & Vogel, V. Binding-activated localization microscopy of DNA l. Nano Lett. 11, 4008–4011 (2011).

15 Sharonov, A. & Hochstrasser, R. M. Wide-field subdiffraction imaging by accumulated binding of diffusing probes. Proc. Natl. Acad. Sci. U. S. A. 103, 18911–18916 (2006).

16 Jungmann, R. et al. Super-Resolution Microscopy by Fluorescence Imaging of Transient Binding on DNA Origami. Nano Lette 4756–4761 (2010).

17 Jungmann, R. et al. Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. Nat. Methods 2014, 11, 313– 318, DOI: 10.1038/nmeth.2835

18 Dai, M., Jungmann, R. & Yin, P. Optical imaging of individual biomolecules in densely packed clusters. Nat. Nanotechnol. 11, 798–807 (2016).

19 Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with DNA-PAINT. Nat. Protoc. 12, 1198–1228 (2017).

20 Filius, M. et al. High-Speed Super-Resolution Imaging Using Protein-Assisted DNA-PAINT. Nano Lett. 20, 2264–2270 (2020).

21 Auer, A., Strauss, M. T., Schlichthaerle, T. & Jungmann, R. Fast, Background-Free DNA-PAINT Imaging Using FRET-Based Probes. Nano Lett. 17, 6428–6434 (2017).

22 Deußner-Helfmann, N. S. et al. Correlative Single-Molecule FRET and DNA-PAINT Imaging. Nano Lett. 18, 4626–4630 (2018).

23 Holden, S. J. et al. Defining the limits of single-molecule FRET resolution in TIRF microscopy. Biophys. J. 99, 3102–3111 (2010).

24 Kim, S. H., Kim, H., Jeong, H. & Yoon, T.-Y. Encoding Multiple Virtual Signals in DNA Barcodes with Single-Molecule FRET. Nano Lett. 21, 1694–1701 (2021).

25 Lee, J., Park, S. & Hohng, S. Accelerated FRET-PAINT microscopy. Mol. Brain 11, 70 (2018).

4

Evaluation of FRET X for Single-Molecule Protein Fingerprinting

Carlos de Lannoy*, **Mike Filius***, Raman van Wee, Chirlmin Joo, and Dick de Ridder

* These Authors have contributed equally to this work

iScience
iScience (2021) 24, 103239, November 19
DOI: https://doi.org/10.1016/j.isci.2021.103239

4.1 Abstract

Single-molecule protein identification is a novel, as of yet unrealized concept with potentially ground-breaking applications in biological research. We propose a method called FRET X (Förster Resonance Energy Transfer via DNA eXchange) fingerprinting, in which the FRET efficiency is read out between exchangeable dyes on protein-bound DNA docking strands, and accumulated FRET efficiency values constitute the fingerprint for a protein. To evaluate the feasibility of this approach, we simulated fingerprints for hundreds of proteins using a coarse-grained lattice model and experimentally demonstrated FRET X fingerprinting on a system of model peptides. Measured fingerprints are in agreement with our simulations, corroborating the validity of our modeling approach. In a simulated complex mixture of >300 human proteins of which only cysteines, lysines and arginines were labeled, a support vector machine was able to identify constituents with 95 % accuracy. We anticipate that our FRET X fingerprinting approach will form the basis of an analysis tool for targeted proteomics.

4.2 Introduction

Proteins come in a wide variety of shapes, sizes and forms. Each is attuned to fulfill one or more of the many functions that are essential to living cells, including the catalysis of metabolic reactions, replication of genetic information, provision of structural support, transport of molecules and many more. To fully understand the biological processes taking place in a cell, it is critical to identify and quantify constituents of its proteome at any given time during the cell cycle.

Mass spectrometry (MS) is currently the gold standard for protein identification and quantification. Over the past decades, MS techniques have improved tremendously in terms of accuracy and dynamic range; however, detecting and distinguishing all proteins in complex samples remains challenging. Many biologically and clinically relevant proteins such as signaling molecules and disease biomarkers occur in such low abundance that they remain undetectable by MS.[1] Moreover, the proteome complexity increases through alternative splicing or posttranslational modifications, as a single gene can produce dozens of distinct protein varieties, referred to as proteoforms.[2] Not all of these proteoforms canbe distinguished by current approaches. As such, there is considerable incentive for the development of new protein sequencing methods that operate at the single-molecule level.[3,4]

Single-molecule techniques have boosted DNA sequencing, allowing for the identification of individual nucleic acid molecules, and are now routinely used for genome and transcriptome mapping of single cells.[5] However, the search for single-molecule protein sequencing techniques is not trivial due to the high complexity of protein molecules compared to DNA molecules. For example, the DNA code consists of only four nucleotides whereas there are twenty different amino acids for proteins. Furthermore, low abundant DNA molecules can be enzymatically amplified outside the cell whereas such an enzyme is absent for proteins.

Novel single-molecule protein analysis methods have been proposed to circumvent this additional complexity. Importantly, only a subset of the theoretically possible combinations of polypeptide chains occurs in nature, and a fraction of that subset is of importance in a given research setting. Therefore, proteins may be identified by reading out a signature of incomplete information, which is then compared to a database of relevant signatures. We refer to this approach as protein fingerprinting, and to said protein signatures as protein fingerprints. It has been shown that sufficiently distinct protein fingerprints only require the read-out of a small subset of residue types.[6–8] If cysteine and lysine residues were orthogonally labeled and read out sequentially, our simulations indicated that the majority of human proteins were uniquely identifiable.

Several novel protein fingerprinting methods based on the readout of a subset of residue types have recently been demonstrated, most of which require linearization of the polypeptide chain to allow for the determination of the residue order.[9,10] This linearization can be achieved by translocating the polypeptide chain through a nanopore[4] or by using a fluorescently labeled motor protein8 to read out the modified residues required for fingerprinting. Alternatively, the protein fingerprint can be obtained by labeling certain amino acids and determining their location through several Edman degradation cycles.[11] Although full length proteins are difficult to analyze due to

the limited number of Edman cycles that can be performed, its utility for analyzing shorter peptides has been shown in a proof of concept. All these approaches have in common that they probe each protein only once, while the accuracy would increase if the same molecule could be measured multiple times.

In this study, we present a novel protein fingerprinting method that builds further on the concept of residue-specific labeling of selected amino acids and obtains a protein fingerprint by determining the location of amino acids in the 3D structure of a protein. As the size of most proteins lies in the low nanometer range, our protein fingerprinting approach requires a technique that can determine the location of residues with sub-nanometer resolution. Single-molecule FRET is well suited for this and comes with the benefit that several thousands of molecules can be imaged at the same time, if full length proteins can be immobilized in a microfluidic chamber.[12] Here we verify the feasibility of a single-molecule FRET-based protein fingerprinting method. We first demonstrate that experimentally obtained fingerprints for four model peptides are distinct and are reproduced by our simulation method. Then we show that simulated fingerprints of 313 human proteome constituents can be identified with 95% accuracy. If mislabeling of residues is assumed to occur, this accuracy decreases to 91%. This supports the notion that FRET fingerprinting allows for the reliable identification of proteins in complex mixtures.

4.3 Approach

4.3.1 FRET X for protein fingerprinting

To realize protein fingerprinting using single-molecule FRET, a resolution sufficient to determine the location of multiple amino acids in the protein structure is required. However, single-molecule FRET analysis is limited to just one or two FRET pairs in a single measurement.[13,14] Recently, our group developed a concept to allow for the detection of multiple FRET pairs in a single nanoscopic object. Our technique, FRET X (FRET via DNA eXchange), employs transient hybridization of DNA strands labeled with a fluorophore to temporally separate FRET events that originate from different FRET pairs. We have shown that FRET X can resolve the distance between multiple FRET pairs with sub-nanometer accuracy.[15] Here, we apply FRET X for protein fingerprinting. By detecting target amino acids one by one, FRET X produces a unique fingerprint, allowing identification of the protein from a reference database.

Figure 4.1 illustrates the workflow for protein fingerprinting using FRET X. A subset of amino acids of a protein of interest is labeled with orthogonal DNA sequences, which serve as docking strands for their complementary imager strands (**Figure 4.1**a). One of the termini is labeled with a unique DNA sequence, which functions as a reference point and facilitates immobilization of the full-length protein to a microfluidic chip. To obtain a FRET fingerprint for one of the amino acids, fluorescently labeled imager strands for the terminal reference sequence and for the particular amino acid (e.g. Cysteine, **Figure 4.1**b) are added. The imager strands for the reference point are labeled with an acceptor fluorophore, while those for the cysteines carry a donor. FRET can occur only when both imager strands are simul-

taneously bound. The transient and repetitive binding of imager strands reports on the relative location of a residue to the reference point. Furthermore, since the pool of fluorophores is continuously replenished, the effect of photobleaching is mitigated and we can probe each residue multiple times, thereby increasing the precision. After obtaining a sufficient number of FRET events, the FRET fingerprint can be constructed, reporting on the distance of each target amino acid from the reference point. Then the microfluidic chamber is washed and a new imaging solution is injected to probe a second amino acid (e.g. Lysine) (**Figure 4.1**c). The FRET X cycle can be repeated for any number and type of amino acids, as long as they are labeled with orthogonal DNA sequences. The detection of multiple types of amino acids improves the uniqueness of a protein fingerprint, thereby enhancing the chance of identification. The resolved FRET efficiencies for each amino acid are combined to generate a protein fingerprint, with which a protein can be identified against a reference database (**Figure 4.1**d).

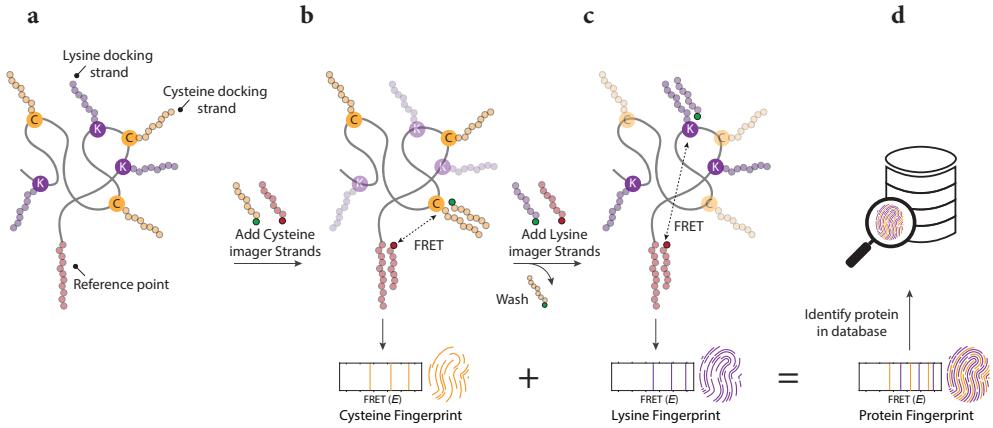


Figure 4.1: The concept of FRET X for protein fingerprinting.

(a) A subset of amino acids (here cysteines and lysines) are labeled with orthogonal DNA sequences which function as docking sites for complementary, fluorescently labeled imager strands. Another orthogonal DNA sequence is conjugated to one of the protein termini, which serves as an acceptor docking site and facilitates immobilization of the protein to a microfluidic device. **(b)** In the first round of FRET X imaging, imager strands that hybridize with the cysteine docking site (yellow circles) and those that hybridize with the reference point (red circles) are injected in the microfluidic chamber. Both the donor and acceptor labeled imager strands transiently interact with their complementary docking strands. When both are present at the same time, FRET can occur and the FRET efficiency is determined between a cysteine and the reference point. Each of the three FRET pairs is separately probed, giving rise to a number of FRET efficiencies (*E*), which constitute the cysteine fingerprint. **(c)** The chamber is washed and FRET X imaging is repeated to probe the lysine fingerprint. This FRET X cycle can be repeated to probe additional amino acids and generate additional fingerprints. **(d)** The FRET efficiencies for individual amino acids are combined to produce a protein fingerprint that can be mapped against a reference database to identify the protein.

4.3.2 Fingerprinting simulations

The usefulness of our method hinges on its ability to discern FRET X fingerprints derived from many different proteins. We run simulations to assess this scenario. Simulating the FRET X fingerprint for a given protein is a complex endeavor, as the fingerprint incorporates both sequence and structural information. While protein structure prediction has seen major advancements recently, cutting-edge methods[16,17] remain too computationally costly to assess many proteins. Furthermore, they cannot account for the presence of conjugated DNA tags. Instead, we opted to use a computationally much less intensive lattice modelling approach[18], in which each residue is represented as a single pseudo-atom, restricted in space to only occupy the vertices of a lattice (**Figure S4.1**). Such structures can be efficiently energy-minimized by a Markov chain Monte Carlo process. Despite their simplicity, past investigations have shown that lattice models can reproduce native protein folding behavior.[19–23]

The attachment of DNA tags to selected residues, as required to accurately model our approach, has not previously been included in lattice models. The precise effect of DNA tags on protein structure is unclear. However, we find that implementation at the coarse granularity required by lattice models may be built on three basic assumptions: that tags prefer to reside on the exterior of the protein, require sufficient unoccupied space to avoid steric hindrance and repel each other if situated closely together. In the lattice models thus produced, FRET values can then be estimated from the simulated dye positions. To simulate the read-out of FRET efficiencies at a given resolution, we bin efficiencies using the resolution as bin width. As we have shown in previous work that a resolution of one FRET percentage point (0.01 *E*) is achievable, we set the resolution of fingerprints to 0.01 E in simulations, unless otherwise noted. As FRET X allows for orthogonal read-out of multiple residue types, the sampling can be repeated to produce the FRET fingerprints associated with different residue types. Analogously to experimentally obtained fingerprints, simulated FRET fingerprints for several residue types are then combined to serve as features for automated classification algorithms.

4.4 Results

4.4.1 Experimental FRET X fingerprinting of model peptides

To demonstrate the concept of protein fingerprinting using FRET X and to compare results with computational predictions, we designed an assay where DNA labeled peptides were immobilized on a PEGylated quartz surface via biotin-streptavidin conjugation (**Figure 4.2**a). Each peptide contains an N-terminal lysine for the attachment of a DNA-docking strand, to allow for the transient binding of an acceptor (Cy5)-labeled imager strand. Additionally, an orthogonal DNA-docking strand was conjugated to a cysteine residue in the peptide to facilitate transient binding of the donor (Cy3)-labeled imager strands (**Figure 4.2**a). The donor and acceptor imager strands were designed to exhibit a dwell time of ~ 2 s (**Figure S4.2**), so that dyes could be frequently replenished. Furthermore, to increase the probability of the presence of

the acceptor imager strand upon donor imager strand binding and allow for FRET detection, we injected 10-fold molar excess of the acceptor imager strand over the donor imager strand. Short-lived FRET events were recorded with single-molecule total internal reflection microscopy upon binding of both donor and acceptor labeled imager strands to the immobilized target peptide.

Next, we plotted a kymograph to visualize the FRET efficiency of each binding event in a target peptide (**Figure 4.2**b). The FRET efficiency for each data point (**Figure 4.2**b, lines) and the mean efficiency per binding event are calculated (**Figure 4.2**b, circles). A histogram of the mean FRET efficiency per binding event shows distinct FRET populations. Gaussian distributions were fit to resolve peak centers with high resolution[14], which then constitute the fingerprint of the peptide (**Figure 4.2**b, bottom panel).

To demonstrate the ability of FRET X to distinguish different peptides with varying FRET pair separations, we designed four model peptides. These peptides had an incrementing distance, in steps of 10 amino acids, between donor and acceptor docking strands (**Figure 4.2**c). First, we performed single-molecule experiments to obtain experimental FRET fingerprints and found a clearly discernible peak for each peptide (**Figure 4.2**d and **Figure S4.3**). Then we simulated FRET fingerprints for the same sequences using our simulation pipeline and found a similar trend. We only fine-tuned the parameters for the repulsion effect between tags to minimize the difference with experimental values (**Figure 4.2**e). In both simulations and experiments we observe a monotonous decrease in FRET efficiency for increasing FRET pair separation. Furthermore, the experimentally obtained fingerprints generally correlate well with values found by simulations (**Figure 4.2**f). Since for each peptide the minimum inter-peptide difference in FRET (E) is larger than the maximum standard deviation, we find that we can distinguish these four peptides by their FRET fingerprint.

4.4.2 Fingerprinting simulation of protein spliceoforms

We set out to evaluate the performance of our method for targeted proteomics, based on simulations. For this we sought to identify the different spliceoforms of the apoptosis regulator Bcl-2 (UniProt ID: Q07817), which are potential biomarkers for cancer[23] and are likely to produce different fingerprints. While BCL-XL is an anti-apoptotic regulator, both BCL-XS and BCL-Xb are pro-apoptotic factors.[24,25] The ratio between these factors is important for cell fate. We simulated simultaneous labeling of cysteine (C) and lysine (K) to create C+K fingerprints for each of the spliceoforms, BCL-XL, BCL-XS, and BCL-XB (**Figure 4.3**a and b). As the spliceoforms differ in the numbers and locations of C and K residues, we expected their fingerprints to be dissimilar. This was indeed the case in simulation (**Figure 4.3**c). Fingerprints do vary across individual molecules of the same spliceoform; however, the fingerprints remain sufficiently characteristic to identify each spliceoform by eye (**Figure S4.4**A).

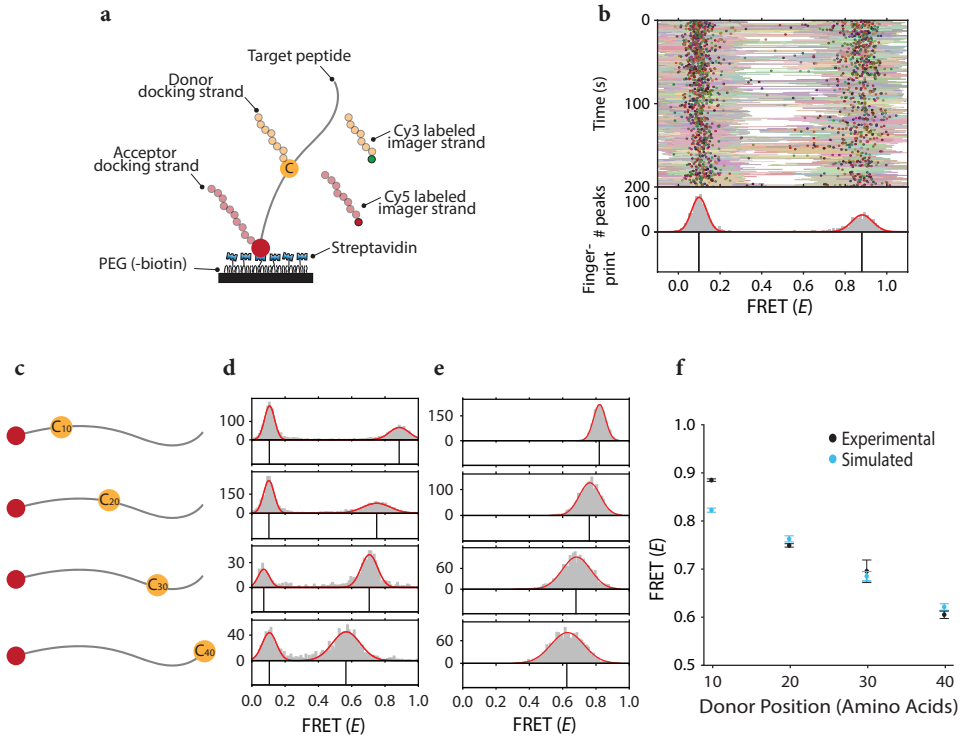


Figure 4.2: Model peptides can be fingerprinted with FRET X.

(a) Depiction of the experimental system for peptide fingerprinting. The target peptide is immobilized through conjugation of its N-terminal biotin with the streptavidin on the PEGylated surface. The donor (Cy3) labeled imager strand (green) can bind to the DNA docking site on the cysteine, while the acceptor (Cy5) labeled imager strand (red) can hybridize to the docking site on the lysine. Simultaneous binding generates short FRET events and is observed with total internal reflection microscopy. **(b)** Representative kymograph for a peptide with a cysteine that is 10 amino acids separated from the acceptor binding site. The FRET efficiency for each data point in a binding event (lines) and the mean FRET efficiency from all data points in a binding event (dots) are indicated as a function of time. A Gaussian distribution (0.88 ± 0.05) is fitted on a histogram of average FRET efficiencies per FRET event. The means of the Gaussians are plotted in a separate panel (bottom) and are referred to as the FRET fingerprint of the peptide. The FRET population on the left is caused by donor leakage into the acceptor channel. **(c)** Our four model peptides have a lysine at the N-terminus and a cysteine at position 10, 20, 30 or 40. **(d)** Experimental distributions and fingerprints for each peptide show a downward trend in FRET (E) for increasing FRET pair separation (0.89 ± 0.06, 0.75 ± 0.08, 0.72 ± 0.03, 0.57 ± 0.08). **(e)** The simulated distributions and fingerprints for the four peptides show a similar downward trend. **(f)** Experimental and simulated data correlate well. Standard deviation of experimental data points is over four kymographs (each consisting of hundreds of events). Experiments were performed on separate days.

We also trained and tested a support vector machine (SVM) classifier on 10 replicates in a 10-fold cross validation scheme and attained an accuracy of 100%.

We then simulated a more difficult scenario, in which we attempted to classify fingerprints for six spliceoforms of PTGS1 (UniProt ID: P23219).[26] Although the higher number of C and K residues made discrimination of fingerprints by eye harder, an SVM trained and tested in a 10-fold cross validation scheme was still able to separate the six spliceoforms with 100% accuracy (**Figure S4.4**B).

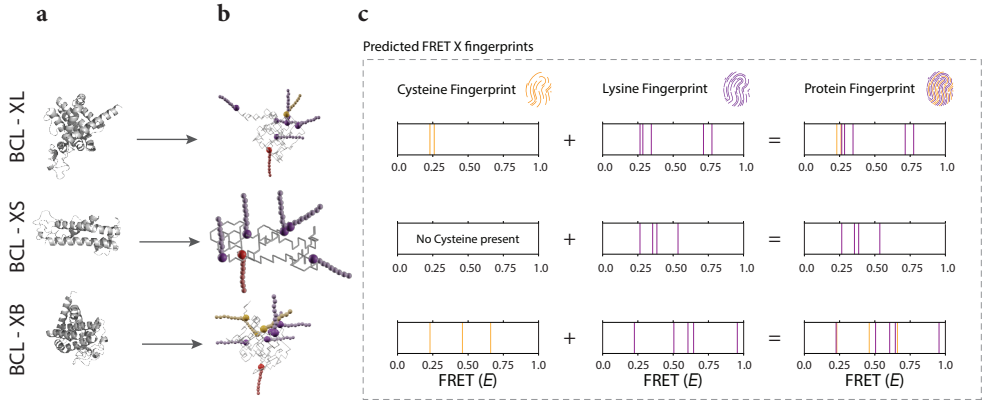


Figure 4.3: Representative FRET (E) fingerprints for three spliceoforms of BCL-X.
(a) Fully atomic structure for BCL XL, Xs and Xb (from top to bottom) as predicted by the RaptorX structure prediction tool. **(b)** Energy-optimized lattice model structures with DNA-docking strands attached to cysteines (orange) and lysines (purple). The reference acceptor docking strand (red) is added to the N-terminus of the proteins. **(c)** The simulated fingerprint for spliceoform of the BCL proteins. Fingerprints are based on averaged donor-acceptor distances in 100 structural snapshots of Markov chain-generated lattice model structures.

4.4.3 Analysis of simulated protein mixtures

To evaluate a test case displaying a complexity closer to that found in a single cell, we selected all UniProt human proteome (ID: UP000005640) entries that were linked to a single-chain structure in the RCSB protein database and for which lattice modeling was able to find a configuration without steric hindrance of docking strands (n = 313). Based on available targeted residue labeling chemistries and relative residue frequencies in naturally occurring proteins, we simulated labeling schemes involving cysteine (C), lysine (K) and arginine (R). For each protein we generated fingerprints based on 10 separately simulated molecules, after which we trained and tested an SVM classifier in a 10-fold cross validation scheme. Here we report overall classifier accuracy. To identify the subset of proteins for which our method works well, we also report the number of well-identifiable proteins, i.e. those for which more than 5 of the replicates were identified correctly.

We find that our classifier performs at 45% accuracy on C-labeled proteins. Of 313 proteins, 126 were well-identifiable, indicating that labeling only C-residues is sufficient to consistently recognize this subset of proteins (**Figure 4.4**a, orange circle). 57 proteins did not contain C residues and are thus impossible to identify using C-labeling. The remaining 130 poorly identifiable proteins generally produced fingerprints containing few FRET values or highly variable fingerprints, the latter indicating a lack of structure stability.

When C+K or C+K+R residues were labeled, accuracy rose to 82% and 95% respectively (**Figure 4.4**b). As expected, fingerprints are more likely to obtain a characteristic signature if distances for more residue types are tracked. Numbers of well-identifiable fingerprints also rose to 278 and 312 out of 313 respectively. Regardless of which residue types are labeled, we find that proteins containing more tagged residues can be identified with higher accuracy (**Figure 4.4**c).

4.4.4 Robustness against suboptimal experimental conditions

To investigate the effect of labeling errors, we ran simulations for a suboptimal labeling scenario, with a 90% probability of labeling the target residue and a certain non-zero probability to label non-target residues (C: 1%, K:1%, R:0.5%, Supplementary Table 3). For C and K these probabilities were based on experimentally determined efficiencies and specificities found in literature.[27–29]

Overall, we find that labeling errors incur a modest decrease in classifier performance; for C, C+K and C+K+R labeling, accuracy drops from 45%, 82% and 95% to 39%, 74% and 91% respectively (Figure 4B). This indicates that FRET fingerprints - particularly those gained from C+K+R labeling - contain the redundant information required to mitigate the effect of imperfect labeling (**Figure 4.4**c). We also investigated the effect of decreased measurement resolution, however only after reducing resolution far beyond experimentally attainable levels - past 0.10 *E* - did we find severe reductions in accuracy (**Figure S4.5**).

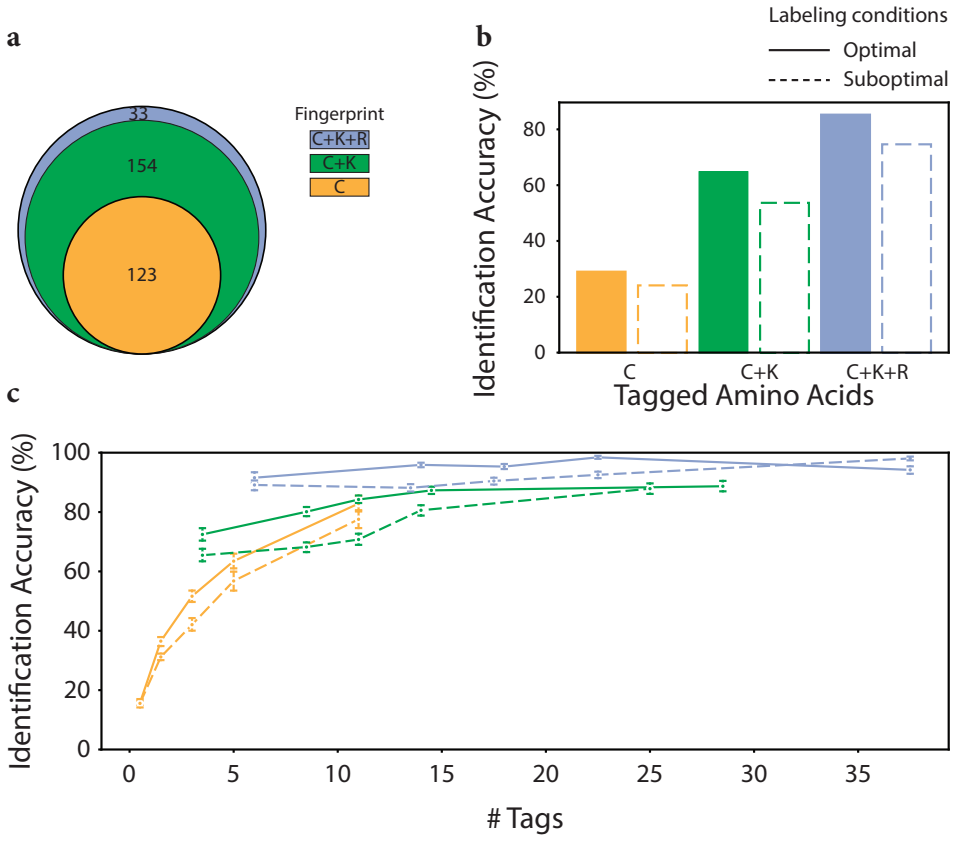


Figure 4.4: FRET X fingerprinting simulation results assuming optimal and suboptimal experimental conditions.

FRET X fingerprint classifier cross-validation performance measures are shown for three combinations of tagged residue types - C, C+K, and C+ K+R - and two labeling qualities - "optimal", where all targeted residues and no off-target residues were labeled, and "suboptimal", where erroneous labeling occurred following the rules in **Table S4.3**. **(a)** Venn diagram showing numbers of proteins that were found to be well-identifiable, i.e. that were correctly identified in more than 5 of 10 cross-validation folds. The total number of proteins is 313. **(b)** The identification accuracy of proteins under optimal and suboptimal labeling conditions. **(c)** Average classifier accuracy as a function of the number of tagged residues in structures, aggregated in five groups with similar numbers of tags. Whiskers denote two standard deviations.

4.5 Discussion

Here we present a new protein fingerprinting approach that determines the location of amino acids within a protein structure using FRET X. We provide evidence of its ability to identify proteins in heterogeneous mixtures using simulations and demonstrate its technical feasibility by producing experimental fingerprints for designed peptides.

We experimentally demonstrate fingerprinting of peptides of 40 amino acids long and observe a monotonous decrease in FRET efficiency. This trend is supported by simulations and suggests that our model peptide has a relatively linear conformation. These peptides do not exhaust the lower end of the FRET-efficiency domain, which implies that larger peptides and proteins with increased FRET pair separation can be fingerprinted. While most proteins are considerably larger than 40 amino acids, they usually adopt a globular structure, which reduces the FRET pair separation. The average protein is estimated to have a diameter of 5 nm[30], while the FRET dyes (Cy3-Cy5) used here are expected to be accurate at distances of up to ~7 nm.[11,12] Therefore, our FRET fingerprinting approach could be suitable for the identification of a large set of human proteins. This notion is substantiated by the simulations run using our lattice model, which shows that also for larger proteins the FRET fingerprints remain discernible.

We show that simulated fingerprints are sufficiently unique and reproducible to consistently identify the majority of the proteins in our simulation pool. Moreover, this result could be achieved by labeling up to three types of amino acids: cysteine, lysine and arginine, all of which can be targeted for specific labeling using existing chemistries.[4,27] Interestingly, even if only cysteine is labeled we find that a subset of proteins remained consistently identifiable, although labeling additional residue types does increase accuracy, the number of identifiable proteins and robustness against labeling errors. It should also be noted that the set of residue types targeted for FRET X fingerprinting can be expanded even further; labeling of e.g. methionine[31] or tyrosine[32] may be employed to further increase accuracy or taylor our method to the detection of a given target protein. For our simulations we investigated proteins for which the structure had already been determined; however, in our experimental system, a microfluidic chamber with non-physiological conditions, proteins may adopt a different structure or a set of several different structures, creating a discrepancy between simulated and experimental fingerprints. However, it is primarily the uniqueness and reproducibility of a fingerprint that is important for protein identification, not necessarily its predictability from a known structure. Furthermore, we expect that as the diversity of a sample decreases from several hundreds to tens of different proteins through sample fractionation, the fingerprint uniqueness and thereby the fraction of correctly identified proteins sharply increases. Adequate sample preparation and purification to reduce sample complexity will therefore be important for more targeted approaches.

A far-reaching goal of the proteomic community is to detect and analyze all proteoforms that can be derived from a single protein encoding gene.[2] Most proteoforms have subtle differences, e.g. alternative splicing or post translational modification, and are difficult to detect with current technologies, such as ELISA, MS or native MS.[33] We have shown that FRET X has the ability to distinguish peptides based on the location of a single cysteine, a subtlety akin to those found in many isoforms, and we have shown two cases in which clinically relevant spliceoforms are well distinguishable based on their simulated FRET X fingerprints. This suggests that our FRET X fingerprinting platform would be a suitable complementary technique for the detection of clinically relevant proteoforms.

4.6 Materials and Methods

4.6.1 Peptide Labeling

Custom designed polypeptides were obtained from Biomatik (Canada) and had a constant backbone sequence, differing only in the cysteine substitutions. Cysteine residues of the polypeptides were reduced with 40-fold molar excess Tris(2-carboethyl)phosphine (TCEP) for 30 minutes and then donor-labeled with 6-fold molar excess monoreactive maleimide-(5') functionalized DNA in 50 mM HEPES pH 6.9 overnight at room temperature. The acceptor docking strand was labeled onto a single lysine that is located at the N-terminus of the peptide. For this, Dimethyl sulfoxide (DMSO) was added to 50 % (v/v) and the pH was increased to pH 7.5 through the addition of NaOH. Next, we added monoreactive N-Hydroxysuccinimide (NHS)-ester functionalized Dibenzocyclooctyne (DBCO) (Sigma Aldrich, Germany) in a 25-fold molar excess and incubated for 6 hours at room temperature. Free NHS-DBCO was removed by using C18 bed micropipet tips (Pierce) according to manufacturer's protocol. Finally, monoreactive Azidobenzoate-(5') functionalized-DNA was added in 5-fold molar excess and incubated overnight at room temperature. See **Table S4.1** and **Table S4.2** for the full list of substrates.

4.6.2 Single-Molecule Setup

All experiments were performed on a custom-built microscope setup. An inverted microscope (IX73, Olympus) with prism-based total internal reflection was used. In combination with a 532 nm diode-pumped solid-state laser (Compass 215M/50mW, Coherent). A 60x water immersion objective (UPLSAPO60XW, Olympus) was used for the collection of photons from the Cy3 and Cy5 dyes on the surface, after which a 532 nm long pass filter (LDP01-532RU-25, Semrock) blocks the excitation light. A dichroic mirror (635 dcxr, Chroma) separates the fluorescence signal which is then projected onto an EM-CCD camera (iXon Ultra, DU-897U-CS0-#BV, Andor Technology). A series of EM-CDD images was recorded using a custom-made program in Visual C++ (Microsoft).

4.6.3 Single-Molecule Data Acquisition

Single-molecule flow cells were prepared as previously described.[34,35] In brief, to avoid non-specific binding, quartz slides (G. Finkerbeiner Inc) were acidic piranha etched and passivated twice with polyethylene glycol (PEG). The first round of PEGylation was performed with mPEG-SVA (Laysan Bio) and PEG-biotin (Laysan Bio), followed by a second round of PEGylation with MS(PEG)4 (ThermoFisher). After assembly of a microfluidic chamber, the slides were incubated with 20 µL of 0.1 mg/mL streptavidin (Thermofisher) for 2 minutes. Excess streptavidin was removed with 100 µL T50 (50mM Tris-HCl, pH 8.0, 50 mM NaCl). Next, 50 µL of 75 pM DNA-labeled peptide was added to the microfluidic chamber. After 2 minutes of

incubation, unbound peptide and excess Azide-DNA from the earlier click reaction was washed away with 200 µL T50. Then, 50 µL of 10 nM donor labeled imager strands and 100 nM acceptor labeled imager strands in imaging buffer (50 mM Tris-HCl, pH 8.0, 500 mM NaCl, 0.8 % glucose, 0.5 mg/mL glucose oxidase (Sigma), 85 ug/mL catalase (Merck) and 1 mM Trolox (Sigma)) was injected. All single-molecule FRET experiments were performed at room temperature (23 ± 2 °C).

4.6.4 Data analysis

Fluorescence signals are collected at 0.1-s exposure time unless otherwise specified. Time traces were subsequently extracted through IDL software using a custom script. Through a mapping file, the script collects the individual intensity hotspots in the acceptor channel and pairs them with intensity hotspots in the donor channel, after which the time traces are extracted. During the acquisition of the movie, the green laser is used to excite the Cy3 donor fluorophores. For automated detection of individual fluorescence imager strand binding events, we used a custom Python code (Python 3.7, Python Software Foundation, https://www.python.org) utilizing a two-state K-means clustering algorithm on the sum of the donor and acceptor fluorescence intensities of individual molecules to identify the frames with high intensities.[31] To avoid false positive detections, only binding events that lasted for more than three consecutive frames were selected for further analysis. FRET efficiencies for each imager strand binding event were calculated and used to build the FRET kymograph and histogram. Populations in the FRET histogram are automatically classified by Gaussian mixture modeling.

4.6.5 Simulations

Fingerprinting simulations were generated using a lattice folding model written in Python 3.7. Simulation and analysis code are freely available at https://git.wageningenur.nl/lanno001/FRET_X_fingerprinting_simulation. A protein folding simulation was implemented to incorporate DNA-tags attached to certain residues and account for their effect on the protein structure. Lattice models were used because of the far lower computational power needed for folding simulations compared to fully atomistic models allowing unrestricted movement, which is attained by reducing each amino acid to a pseudo-atom and restricting its possible positions to the vertices of a lattice. Such models have previously been used in applications where low computational requirements were essential.[16–20] The procedure starts with a fully atomistic native structure, which is converted to a lattice structure with tagged residues marked. This structure is then refolded by making local modifications and calculating the effect these have on the model energy (Etot), as calculated by an energy function. Modifications that decrease Etot are accepted, whereas those that increase Etot are more likely to be discarded the more they increase Etot. The procedure ends when all DNA-tags fit in the structure without causing steric hindrance. Aspects of the modeling procedure are described in more detail below.

4.6.6 Lattice structure

The lattice modeling procedure employed here largely resembles those in previously published applications.[19] In particular, the model developed by Abeln et al.[19] was used as a starting point, however the cubic lattice was replaced by a novel body-centered cubic (BCC) lattice (**Figure S4.6**). The octahedral unit cell of a BCC lattice borders eight neighboring cells through its hexagonal faces and four through its square faces. However, only connections through hexagonal faces are considered, as this allows all bonds to be of the same length. As a result, only even coordinates in the lattice are valid vertices for residue placement.[33] This implementation increases the number of contacts that each non-endpoint residue can make from four to six (not including immediately neighboring residues) and increases the number of directions into which a bond may extend. The resulting increased flexibility allows lattice models to more closely resemble native folds. Moreover, alpha helices are represented better as the BCC lattice allows structures that make one regular turn per five residues.

4.6.7 Tag implementation

As the precise effect of the presence of DNA-tags on protein structure is unclear, we relied on several basic assumptions to include them in the model. First, we assume that DNA-tags prefer to reside in the periphery of a protein due to their polar backbones. Thus, labeling an internal residue should alter local structure to accommodate sufficient space from the residue to the surface, while tagging a residue that already resides on the protein surface should affect the structure less severely. This was implemented by adding a substantial energy penalty if a tagged residue did not have space for a DNA tag to reach the periphery of the structure without clashing with the main chain. Secondly, we assume that tags will electrostatically repel each other. This is represented by introducing a minimum angle and dihedral between tag pairs that are spatially close together in a given configuration (**Figure S4.7**). To parameterize this effect, we compared predicted fingerprints of 40-residue model peptides to the presented experimental data and found that values are reproduced well if at least a 70° angle and dihedral are enforced between tags situated within 20 Å of each other.

4.6.8 Simulated labeling scenarios

Two labeling scenarios are employed in this work. Under the optimal scenario, all target residues are labeled and no off-target labeling takes place. Under the suboptimal scenario, both labeling efficiency and specificity are decreased, following a similar procedure to Ohayon et al.7; each target residue has a 90% chance of being labeled by its dedicated chemistry, while some off-target labeling probability is defined for one or more other residue types. Where possible, efficiency and specificity parameters are based on literature (**Table S4.3**).

4.6.9 Structure collection

We base the lattice models used in our fingerprinting simulations on fully atomistic structures as stored in the RCSB PDB. To obtain a dataset of relevant structures, we analysed all available PDB entries corresponding to entries in the Uniprot human proteome set (UP000005640). Of the 20,381 entries in the proteome, 7,133 solved structures were found. We further filtered this list on structure quality, retaining only those with an R-free value below 0.21, and removed structures with non-canonical residues as our model contains no energy modifiers for these residues. Lastly, quaternary structure is expected to be lost during sample preparation, thus to avoid having to model the effect of losing other chains on the tertiary structure of the target chain, we removed structures which were crystalized as a complex of multiple chains. After these filtering steps, 746 structures remained for our simulations.

A lattice models is derived from a fully atomistic structure by reducing it to its Cα positions and placing each Cα-atom on the nearest lattice vertex, while remaining connected to its neighboring Cα-atom, starting from the residue with the lowest index. Alpha helices are forced to remain intact on the lattice, by first translating involved Cα-atoms to a lattice-compliant helix and then minimizing the distance between their respective lattice positions simultaneously.

As no PDB structures are available for the 40-residue model peptides labeled in practical experiments, starting structures for these peptides were stretched configurations. Starting structures for BCL-X and PTGS1 spliceoforms were generated using the RaptorX structure prediction server.[36]

4.6.10 Folding simulation

After initialization of the lattice model, a Markov Chain Monte Carlo (MCMC) procedure is employed to minimize the structure energy E_{tot}.

$$E_{tot} = E_{AA} + E_{sol} + E_{ss} + E_{tag} + E_{reg}$$

Residue interaction and residue-solvent interaction terms E_{AA} and E_{sol} are summed pairwise interaction terms between contacting residues or residue-solvent contacts, the magnitudes of which are obtained empirically.[37] The secondary structure formation energy term E_{ss} is adapted from Abeln et al.[19] and incurs an arbitrarily high energy bonus of -25 if an alpha helix or beta sheet is formed, but only if a given residue also was part of such a secondary structure in the native fold. An alpha helical residue incurs this bonus if the exact shape of the helix is formed (i.e. residue i up to i+4 take the same relative orientation at each step), while a bonus for beta sheet formation is applied if non-neighboring beta-sheet residues are adjacent to each other. The tag energy term E_{tag} incurs an arbitrarily high energy penalty of 100 for each residue impeding the shortest route from a tagged residue to the periphery of the structure. Lastly, the regularization term E_{reg} incurs a penalty for large structural reorganizations occurring in a single MCMC step, as we found that this helps to retain the native fold as much as possible.

4.6.11 Fingerprint extraction

To account for the fact that a structure may adopt several conformations over the course of measurements, fingerprints are based on a series of structure snapshots. After the folding simulation has finished and the structure which accommodates all DNA-tags without steric hindrance is found, another 1,000 MCMC steps are performed. During these steps, snapshots are taken at intervals of 10 steps, thus measuring 100 slightly different conformations. For each snapshot, dye positions are chosen randomly from all accessible lattice directions. If tags are found to be closer than 20Å to each other, a minimum angle and dihedral angle of 70 degrees each between those tags is enforced (**Figure S4.7**). Distances between donor and acceptor dye positions are estimated from the snapshots and averaged, after which the FRET efficiency is calculated as follows:

$$E_{FRET} = \frac{1}{1 + (R / R_0)^6}$$

Here R is the modeled distance between donor and acceptor dye and R_0 is the Förster radius, which characterizes the used FRET dye pair (R_0 assumed constant at 54 Å for the Cy3-Cy5 FRET pair[12]). Finally, all FRET values are binned and normalized over the number of snapshots to produce the final fingerprint. The bin width is used here to represent the observation resolution. Resolution is fixed at 0.01 unless otherwise noted, as previous work has shown that such a resolution can be achieved using FRET X.[15] If multiple residue types are tagged, each residue type generates its own fingerprint which is binned separately.

4.6.12 Classification

To classify simulated fingerprints a support vector machine (SVM) was implemented using the scikit-learn package (v0.23.2.)[36]. In a ten-fold cross validation procedure, the SVM was fitted to a training set consisting of 90% of produced fingerprints and tested on a held-out test set. As a higher resolution is also more sensitive to noise by unstable fingerprints, the resolution is tuned during training in steps of 0.01 E to produce the highest training accuracy. To evaluate classifier performance, we calculated test accuracy, i.e. the number of correct classifications over total number of test examples. As this measure obscures whether classification mistakes are consistently made for certain proteins or are randomly distributed, we also determined which proteins were correctly classified in more than half of replicates, which we denote as well-identifiable proteins.

4.7 Supporting Information

4.7.1 Supporting Figures

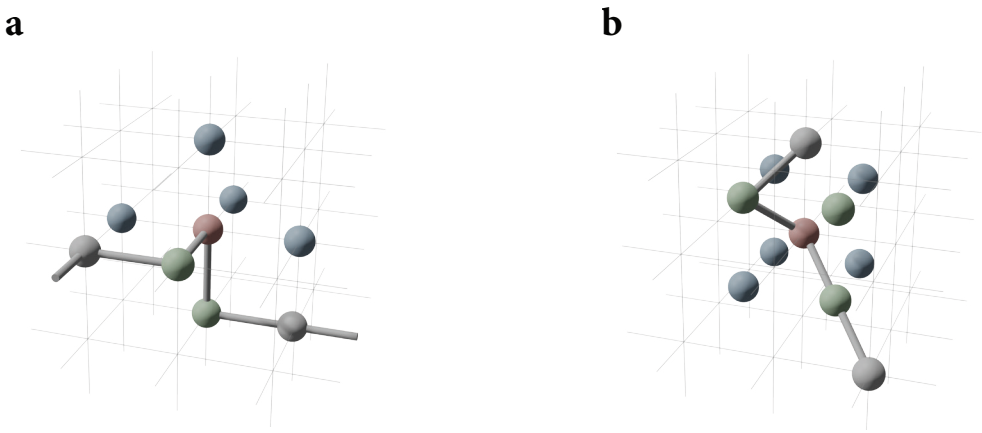


Figure S4.1: Pseudo-atoms on a cubic lattice (A) and a body-centered cubic lattice (B). Shown are one main pseudo-atom (red) and its direct neighbors (green). For the main pseudo-atom, all possible adjacent pseudo-atoms (green) are depicted. Figures were generated in Blender 2.93.0.

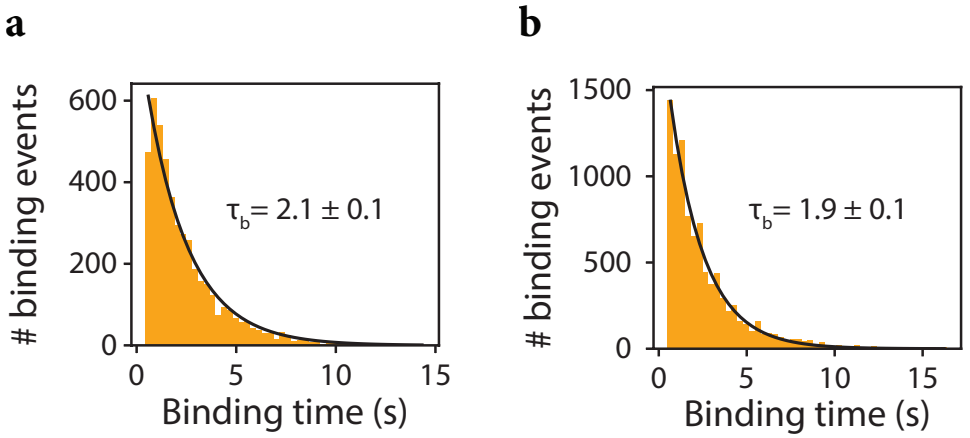


Figure S4.2: Single-molecule binding kinetics of FRET X imager strands.
(a) Dwell-time histogram for the FRET X donor imager strand (**Table S4.2**) fitted with a maximum likelihood estimation for a single exponential distribution (black line). Average ± standard deviation of four different estimations gives: 2.14 ± 0.07 s. The number of datapoints for this distribution: n = 4687 and peptide used was K1C40. **(b)** Dwell-time histogram for the FRET X acceptor imager strand (**Table S4.2**) fitted with a maximum likelihood estimation for a single exponential distribution (black line). Average ± standard deviation of four different estimations gives: 1.92 ± 0.02 s. The number of datapoints for this distribution: n = 9477 and peptide used was K1C40.

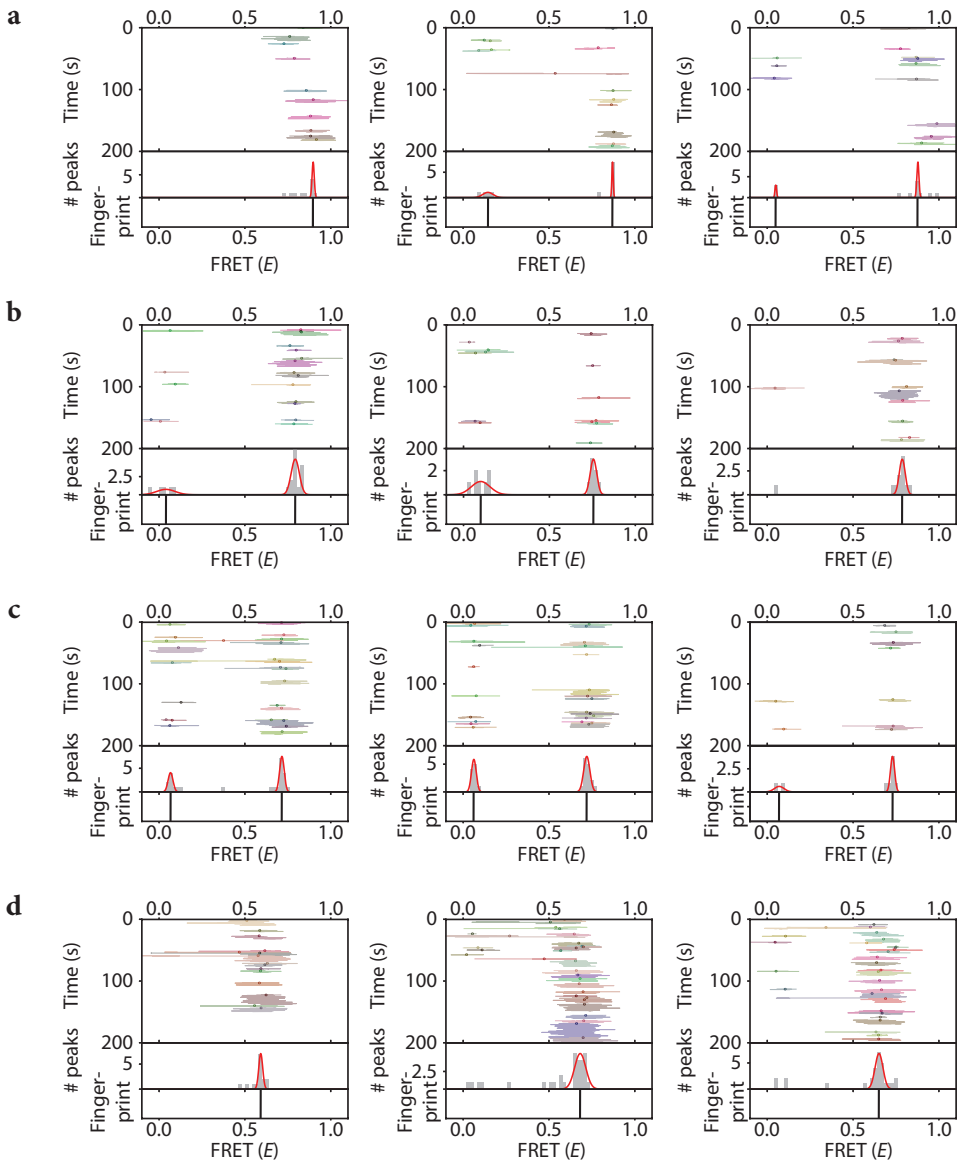


Figure S4.3: Representative kymographs of individual peptides.

(a-d) Representative single-molecule FRET kymographs for each of the four peptides. The downward FRET (*E*) trend remains at the single-molecule level and the distribution of each individual molecule can be fitted with high precision (s.d. ≤ 0.03 for each distribution). The ensemble of many identical single-molecules results in the FRET-fingerprint (**Figure 4.2**b).

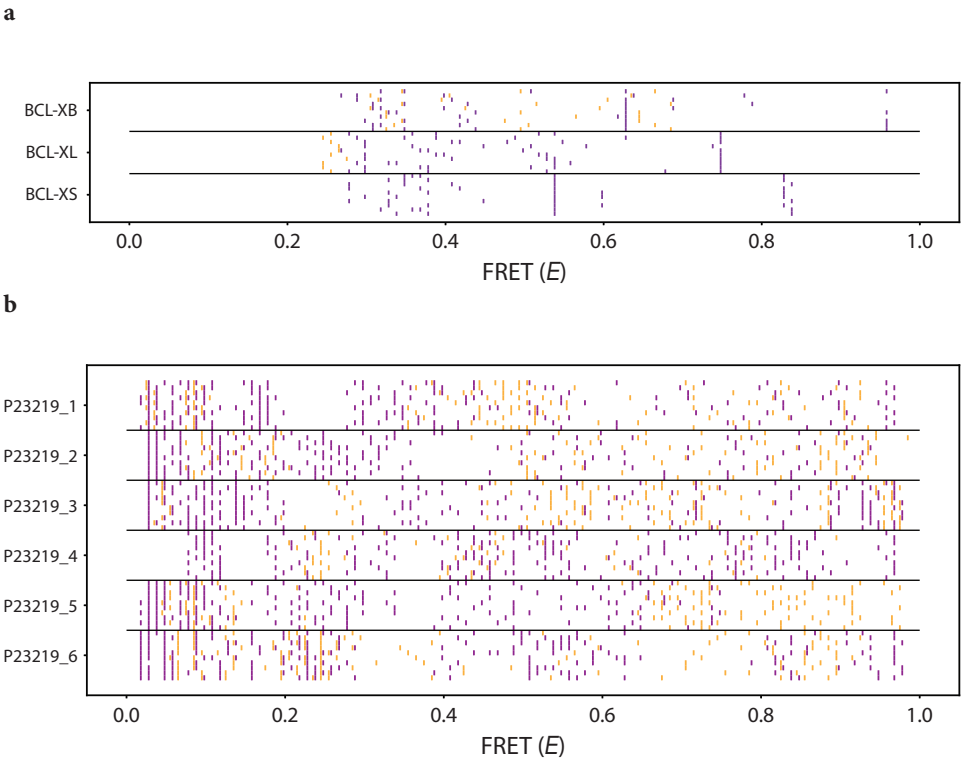


Figure S4.4: Simulated FRET X fingerprints for spliceoforms of BCL and PTGS1.
(a) FRET X fingerprints for ten simulated molecules, one per horizontal line, of three BCL-X spliceoforms: BCLXL, BCLXS and BCLXB. Cysteine and lysine-derived values are colored orange and purple respectively. **(b)** FRET X fingerprints for ten simulated molecules of six PTGS1 spliceoforms.

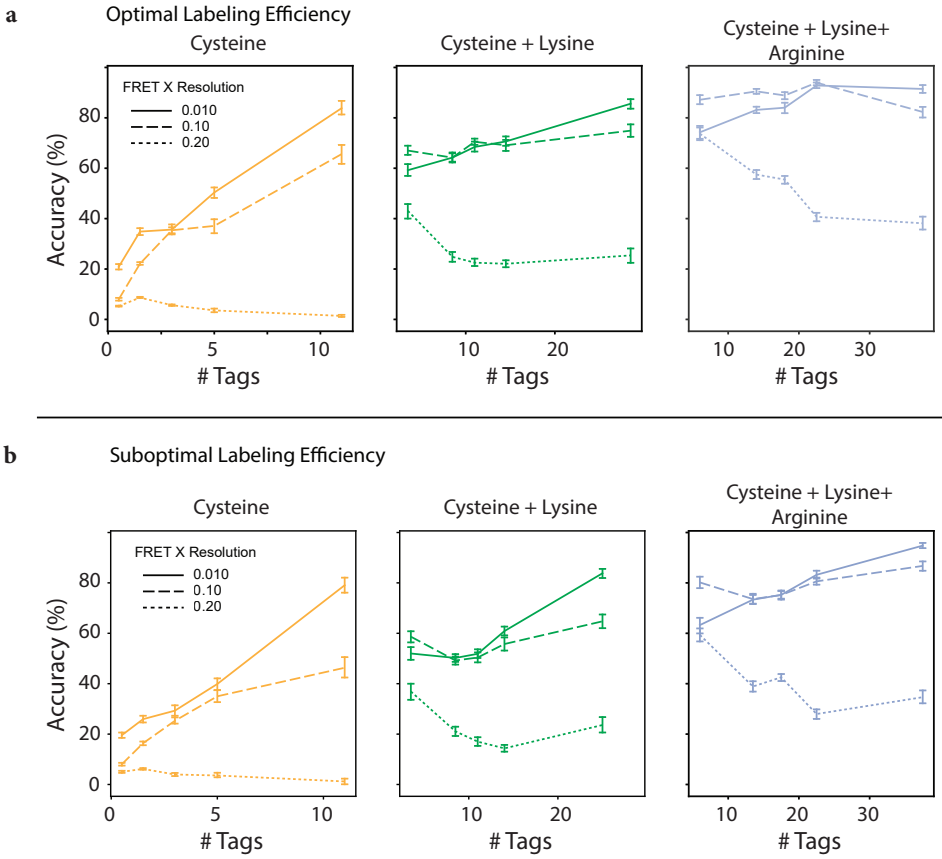


Figure S4.5: SVM classifier accuracy on simulated fingerprints for 313 proteins at different resolutions.

Average classifier accuracy versus the number of tagged residues in structures, aggregated in five groups with similar numbers of tags, at different resolutions. Data are shown for **(a)** optimal labeling quality (i.e. 100% efficiency, 100% specificity) and **(b)** suboptimal labeling quality (see supplementary table 3 for efficiency and specificity), for three combinations of tagged residues (C, C + K, and C+ K + R). Whiskers denote two standard deviations.

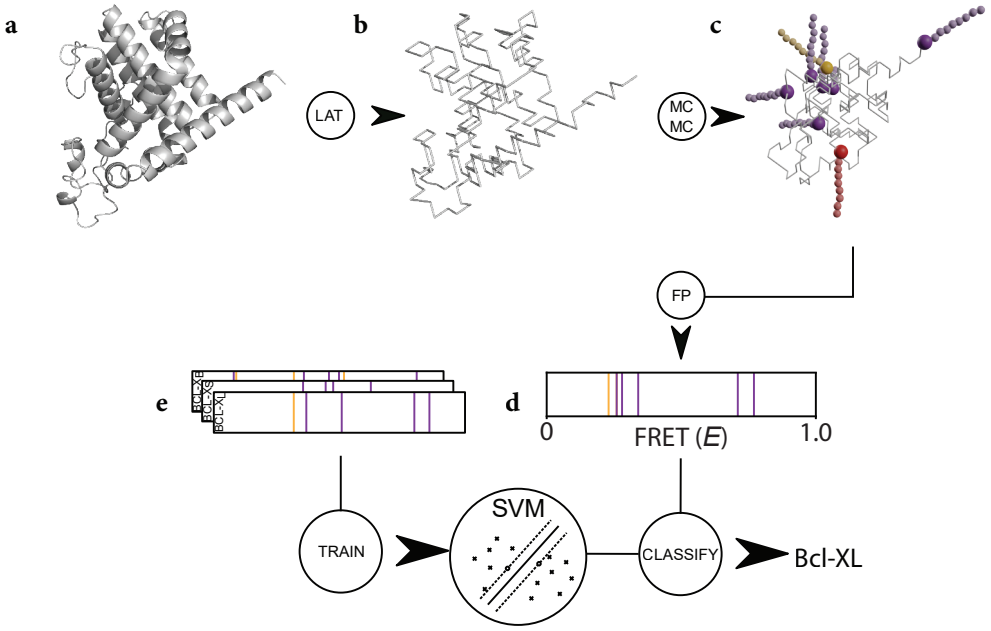


Figure S4.6: Schematic of FRET X fingerprinting simulation and classification pipeline used in this work.

(a) Simulation starts from a fully atomistic structure, **(b)** which is first converted into a lattice model. In the lattice model all residues are reduced to their Cα positions. **(c)** Residues to which docking strands must be attached are marked, after which the structure is randomly mutated using a Markov chain Monte Carlo (MCMC) process, until docking strands no longer experience steric hindrance from the rest of the structure. **(d)** The MCMC process then continues while snapshots are taken at regular intervals. Donor-acceptor dye distance for each dye pair is averaged over all snapshots and translated into a FRET efficiency. Combined, the FRET efficiencies form the final fingerprint for this molecule. **(e)** A support vector machine (SVM) is trained on a set of fingerprints with known identities, after which it can be used to classify unseen fingerprints.

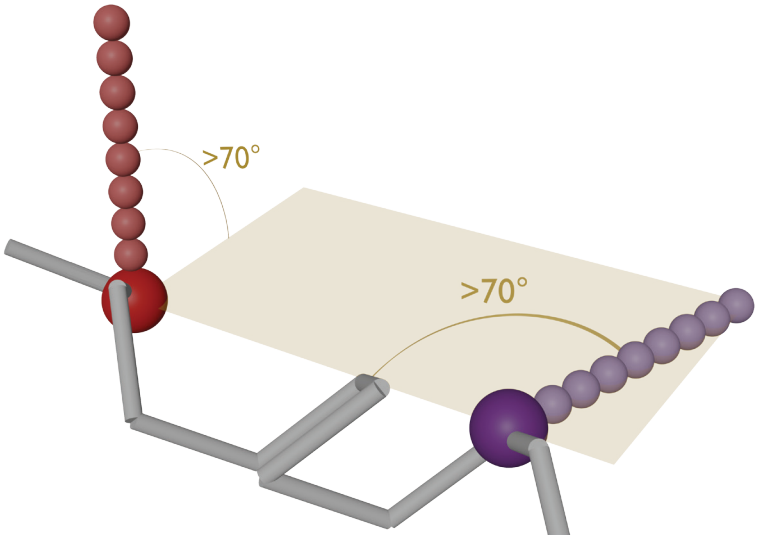


Figure S4.7: Illustration of the tag repulsion implementation of the lattice model.
If the distance between two tagged pseudo-atoms is found to be less than 20 Å, both the angle and the dihedral angle should be larger than 70° to obtain a valid tag position. Figure was generated in Blender 2.93.0.

4.7.2 Supporting Tables

Table S4.1: Single-Molecule Pepitide constructs.

Peptide	Amino acid Sequence (N → C)	Modification	Supplier
K1C10	KAGERDNFA**C**HMALVPVAAN-DENYALAAAANDENYALAAA	Biotin-Ahx @ N-terminus	Biomatik (CAN)
K1C20	KAGERDNFAPHMALVP-VAA**C**DENYALAAAAN-DENYALAAA	Biotin-Ahx @ N-terminus	Biomatik (CAN)
K1C30	KAGERDNFAPHMALVP-VAANDENYALAAA**C**N-DENYALAAA	Biotin-Ahx @ N-terminus	Biomatik (CAN)
K1C40	KAGERDNFAPHMALVP-VAANDENYALAAAAN-DENYALAA**C**	Biotin-Ahx @ N-terminus	Biomatik (CAN)

Table S4.2: DNA Constructs

DNA Construct	Nucleotide Sequence (5' → 3')	Modification	Supplier
Donor imager strand	AGATGTAT	3' Cy3	Ella Biotech (GmbH)
Acceptor imager strand	AATGAAGA	3' Cy5	Ella Biotech (GmbH)
Donor docking sequence	TATACATCTAT	5' Maleimide	Biomers.net (GmbH)
Acceptor docking sequence	TTCTTCATTACT	5' Azidobenzoate	Biomers.net (GmbH)

Table S4.3: Labeling probabilities under suboptimal conditions.

Chemistry	Target Amino Acid	P(target labeling)	Off-Target amino acid	P(Off-target labeling)	Reference
Maleimide	Cysteine (C)	90%	K	1%	Boutureira et al. [27]
NHS ester--mediated derivatization	Lysine (K)	90%	S,Y,T	1%	Abello et al.[28]
Arginine derivatization	Arginine (R)	90%	Any	0.5%	Thompson et al.[29]

4.8 References

1 Zubarev, R. a. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 13, 723–726 (2013).

2 Aebersold, R. et al. How many human proteoforms are there? Nat. Chem. Biol. 14, 206–214 (2018).

3 Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. Nat. Nanotechnol. 13, 786–796 (2018).

4 Alfaro, J. A. et al. The emerging landscape of single-molecule protein sequencing technologies. Nat. Methods 18, 604–617 (2021).

5 Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: Current state of the science. Nature Reviews Genetics vol. 17 175–188 (2016).

6 Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. PLoS Comput. Biol. 11, 1076–1082 (2015).

7 Ohayon, S., Girsault, A., Nasser, M., Shen-Orr, S. & Meller, A. Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. PLoS Comput. Biol. 15, e1007067 (2019).

8 Yao, Y., Docter, M., Van Ginkel, J., De Ridder, D. & Joo, C. Single-molecule protein sequencing through fingerprinting: Computational assessment. Phys. Biol. 12, 10–16 (2015).

9 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).

10 van Ginkel, J. et al. Single-molecule peptide fingerprinting. Proc. Natl. Acad. Sci. 201707207 (2018) doi:10.1073/pnas.1707207115.

11 Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. Nat. Biotechnol. 36, 1076–1082 (2018).

12 Lerner, E. et al. FRET-based dynamic structural biology: Challenges, perspectives and an appeal for open-science practices. eLife vol. 10 (2021).

13 Hohng, S., Joo, C. & Ha, T. Single-Molecule Three-Color FRET. Biophys. J. 87, 1328–1337 (2004).

14 Clamme, J. P. & Deniz, A. A. Three-color single-molecule fluorescence resonance energy transfer. ChemPhysChem 6, 74–77 (2005).

15 Filius, M., Kim, S. H., Severins, I. & Joo, C. High-Resolution Single-Molecule FRET via DNA eXchange (FRET X). Nano Lett. 21, 3295–3301 (2021).

16 Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710 (2020).

17 Xu, J., McPartlon, M. & Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. Nat. Mach. Intell. 1–9 (2021) doi:10.1038/s42256-021-00348-5.

18 Kolinski, A. & Skolnick, J. Reduced models of proteins and their applications. doi:10.1016/j.polymer.2003.10.064.

19 Abeln, S., Vendruscolo, M., Dobson, C. M. & Frenkel, D. A simple lattice model that captures protein folding, aggregation and amyloid formation. PLoS One 9, 85185 (2014).

20 Coluzza, I., Muller, H. G. & Frenkel, D. Designing refoldable model molecules. Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top. 68, 046703 (2003).

21 Bianco, V., Pagès-Gelabert, N., Coluzza, I. & Franzese, G. How the stability of a folded protein depends on interfacial water properties and residue-residue interactions. J. Mol. Liq. 245, 129–139 (2017).

22 Dijkstra, M. J. J., Fokkink, W. J., Heringa, J., van Dijk, E. & Abeln, S. The characteristics of molten globule states and folding pathways strongly depend on the sequence of a protein. Mol. Phys. 116, 3173–3180 (2018).

23 Van Gils, J. H. M. et al. The hydrophobic effect characterises the thermodynamic signature of amyloid fibril growth. PLoS Comput. Biol. 16, e1007767 (2020).

24 Kale, J., Osterlund, E. J. & Andrews, D. W. BCL-2 family proteins : changing partners in the dance towards death. Nat. Publ. Gr. 25, 65–80 (2017).

25 Shiraiwa, N. et al. An additional form of rat Bcl-x, Bcl-xβ, generated by an unspliced RNA, promotes apoptosis in promyeloid cells. J. Biol. Chem. 271, 13258–13265 (1996).

26 Garcia-blanco, M. A., Baraniak, A. P. & Lasda, E. L. Alternative splicing in disease and therapy. 22, 535–546 (2004).

27 Boutureira, O. & Bernardes, G. J. L. Advances in Chemical Protein Modification. Chem. Rev. 115, 2174–2195 (2015).

28 Abello, N., Kerstjens, H. A. M., Postma, D. S. & Bischoff, R. Selective acylation of primary amines in peptides and proteins. J. Proteome Res. 6, 4770–4776 (2007).

29 Thompson, D. A., Ng, R. & Dawson, P. E. Arginine selective reagents for ligation to peptides and proteins. J. Pept. Sci. 22, 311–319 (2016).

30 Erickson, H. P. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. Biological Procedures Online vol. 11 32–51 (2009).

31 Lin, S. et al. Redox-based reagents for chemoselective methionine bioconjugation. 602, 597–602 (2017).

32 Alvarez Dorta, D., Deniaud, D., Mével, M. & Gouin, S. G. Tyrosine Conjugation Methods for Protein Labelling. Chemistry - A European Journal vol. 26 14257–14269 (2020).

33 Leney, A. C. & Heck, A. J. R. Native Mass Spectrometry: What is in the Name? J. Am. Soc. Mass Spectrom. 28, 5–13 (2017).

34 Chandradoss, S. D. et al. Surface passivation for single-molecule protein studies. J. Vis. Exp. 4–11 (2014) doi:10.3791/50549.

35 Filius, M. et al. High-Speed Super-Resolution Imaging Using Protein-Assisted DNA-PAINT. Nano Lett. 20, 2264–2270 (2020).

36 Pedregosa FABIANPEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. Journal of Machine Learning Research vol. 12 http://scikit-learn.sourceforge.net. (2011).

5

Single-Molecule Protein Identification Using FRET X

Mike Filius, Raman van Wee, Carlos Victor de Lannoy, Ilja Westerlaken, Cecilia de Agrela Pinto, Dick de Ridder, and Chirlmin Joo

This chapter is in preparation for publication.

5.1 Abstract

Proteins are the primary functional actors of the cell, hence their identification is pivotal to advance our understanding of cell biology and disease. Current protein analysis methods are limited in their ability to distinguish proteoforms and are not sensitive enough to detect low abundance species. In this proof-of-concept study, we demonstrate FRET X (FRET by DNA eXchange) to localize individual amino acids within a protein structure and thereby identify a protein. Our scheme relies on transient binding of short fluorescently labeled DNA strands to probe the amino acids on a single protein one by one. We validate our method on a series of alpha--synuclein mutants and show that different constituents in a heterogeneous mixture can be discriminated. In addition, we demonstrate that proteins with multiple FRET pairs and globular proteins can be fingerprinted with FRET X. We anticipate that our technology will be used for highly sensitive protein identification in biological and translational research.

5.2 Introduction

Much protein regulation occurs beyond the genome and transcriptome level. Via mechanisms such as alternative splicing and post-translational modifications (PTMs), a single protein encoding gene can produce hundreds of unique protein products, or "proteoforms".[1,2] Subtle differences between proteoforms can have major implications on cell functioning and expression of aberrant proteoforms is implicated in many diseases.[3–5] Therefore, analysis at the proteomic level is required to completely unveil elemental cellular processes. Mass-spectrometry (MS) and enzyme-linked immunosorbent assay (ELISA) have been central to further our understanding of the proteome in the last decade[6–8], yet they have their limitations in terms of dynamic range, sensitivity and the inability to distinguish highly similar proteoforms.[9,10] To overcome the limitations that are inherent to these bulk protein identification methods, highly sensitive protein sequencing is needed.

While single-molecule sequencing of DNA and RNA is omnipresent, the nature of proteins creates several challenges that have thus far precluded single-molecule protein sequencing.[11,12] The increased number of building blocks in the polymer backbone from 4 nucleobases to 20 different amino acids complicates their discrimination and hinders specific labeling. The protein sequencing task is further impeded by the absence of a polymerase-like enzyme that can replicate proteins. Thirdly, protein folding and interactions are much less predictable than nucleic acid basepairing. As a workaround for these challenges, multiple groups have proposed protein fingerprinting, in which partial sequence information is used to generate a unique protein fingerprint.[13–16] By mapping this fingerprint against a reference database a protein can be identified.

Our group has recently introduced a fingerprinting approach, which relies on determining the position of amino acids within the protein structure.[17] In this approach, the protein fingerprint is constructed by using FRET X (FRET by DNA eXchange).[18] FRET X uses short fluorescently labeled DNA imager strands to allow for the detection of multiple FRET pairs in a single nanoscopic object (e.g. protein). One or more residue types are labeled with unique short DNA donor docking strands and an orthogonal acceptor docking strand is conjugated to one of the protein termini. Next, the complementary donor and acceptor labeled imager strands are added to the immobilized proteins. When both imager strands are present simultaneously on a single molecule, FRET occurs, reporting on the distance between the amino acid and the reference point. The binding and unbinding cycle of imager strands allows us to probe multiple FRET pairs, one at a time, with high resolution.[18] Our computational simulations indicated that by structurally fingerprinting just two different amino acids (Cys & Lys), the majority of the human proteome can be identified with FRET X.[17]

Here we validate the concept of FRET X for single-molecule protein fingerprinting. We demonstrate that our FRET X approach can obtain high resolution fingerprints of model proteins. We show that multiple constituents can be discriminated in a complex mixture and that species with more than one labeled amino acid can be identified. We demonstrate that FRET fingerprinting is compatible with globular proteins. This work lays the foundation for single-molecule protein identification with FRET X.

5.3 Results

5.3.1 Single-molecule protein fingerprinting using FRET X

Our protein fingerprinting approach relies on determining the location of amino acids residues relative to a reference point. To demonstrate the concept of protein fingerprinting using FRET X, we designed a single molecule FRET assay where a DNA labeled model protein is immobilized on a microfluidic device through biotin--streptavidin conjugation (**Figure 5.1**a). For our first set of experiments we used the human alpha-synuclein (aSyn) protein to which we conjugated a biotinylated single stranded (ss) DNA strand via an aldehyde encoding sequence (**Figure 5.1**b).[19,20] This ssDNA is used to immobilize the protein and functions as a docking site where complementary an acceptor (Cy5)-labeled imager strands can transiently bind. Additionally, the cysteine residues were labeled with an orthogonal DNA sequence to allow transient binding of donor (Cy3)-labeled imager strands (**Figure 5.1**a). The donor and acceptor imager strands were designed to have dwell times ($\Delta\tau$) of 0.51 ± 0.1 s and 2.1 ± 0.1 s (**Figure S5.1**), respectively. The dwell time of both imager strands were sufficiently short to ensure repetitive binding of the imager strands and allow precise determination of the FRET efficiency.[18] Furthermore, to increase the probability of the presence of the acceptor imager strand upon donor imager strand binding and allow for FRET detection, we injected 5-fold molar excess of the acceptor imager strand over the donor imager strand.

To demonstrate the ability of FRET X to fingerprint proteins, we constructed six aSyn model proteins, containing genetically introduced cysteines (**Figure 5.1**b). The aSyn proteins were designed to have a difference in distance of their single cysteine relative to the reference point that was attached near the C terminus of the protein. For all of the aSyn mutants, we observed short-lived FRET events upon binding of both the donor and acceptor imager strands to DNA labeled immobilized protein. We constructed a kymograph for each recording (**Figure 5.1**c and **Figure S5.2**), where the lines indicate the FRET efficiency (E) for each datapoint and the dots are the mean FRET efficiency per event. The mean FRET efficiencies were plotted in a histogram, which was fitted with a Gaussian function to resolve the centre of each peak with high precision. The center values of each peak are plotted in a separate panel and together constitute the fingerprint of a protein (**Figure 5.1**c, bottom panel). The different aSyn variants yielded clearly distinct distributions and fingerprints (**Figure 5.1**d). The minimum difference in FRET is larger than the maximum standard deviation among the different aSyn mutants (Figure 1E). We observed discernible FRET efficiencies for two mutants where a distance difference is only 5 amino acids (Cys_{124} and Cys_{129}), suggesting FRET X is capable of fingerprinting proteins with a resolution of ~5 amino acids.

As a single amino acid can recur multiple times in a protein sequence, protein fingerprinting using FRET X requires the detection of multiple FRET pairs in a single protein molecule. FRET X employs transient and repetitive binding of short DNA imager strands to probe the location of multiple residues, one at a time. To verify the notion that species with multiple FRET pairs can be fingerprinted with FRET X, we designed two aSyn model proteins, each of which contains two cysteines.

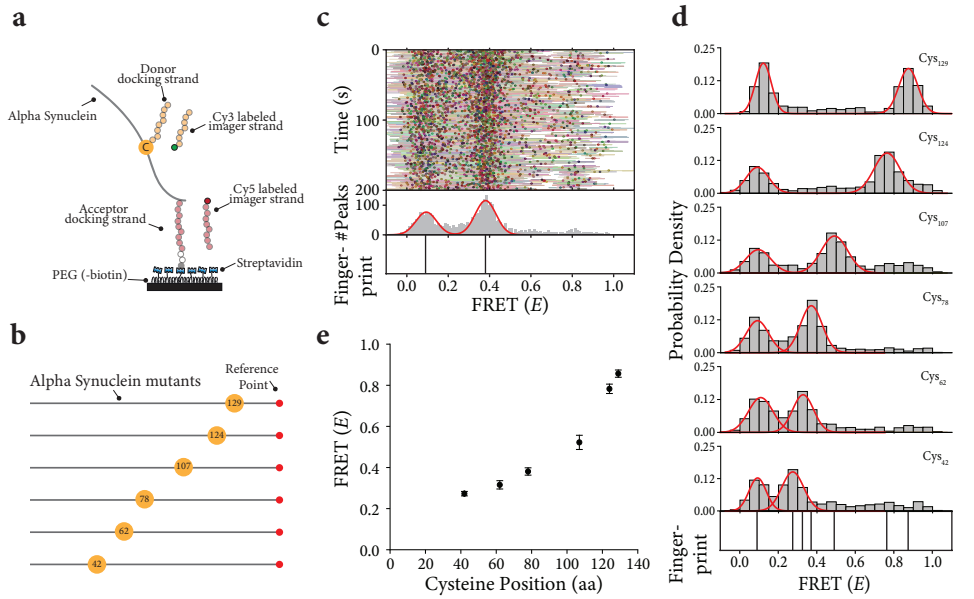


Figure 5.1: Repetitive binding of short DNA imager strands allows for high-resolution protein fingerprints.

(**a**) Schematic representation of the single-molecule assay. The model protein, alpha-synuclein, is engineered to with an C-terminal aldehyde encoding sequence. This motif is used for the conjugation of a biotinylated ssDNA strand (red circles) to facilitate immobilization of the target protein to the PEGylated quartz surface. The donor (Cy3) labeled imager strand (orange) can bind to the DNA docking site on the cysteine, while the acceptor (Cy5) labeled imager strand (red) can hybridize to the docking site that is conjugated to the C terminus of the protein. Simultaneous binding generates short FRET events and is observed with total internal reflection microscopy. (**b**) Schematic representation of the aSyn constructs. Each of the constructs is engineered to contain a single cysteine (orange circle) and c-terminal aldehyde encoding sequence for the attachment of the acceptor docking strand (red circle). (**c**) Representative kymograph for a peptide with a cysteine that is located at residues Cys78, and is 64 amino acids away from the acceptor binding site. The FRET efficiency for each data point in a binding event (lines) and the mean FRET efficiency from all data points in a binding event (dots) are indicated as a function of time. A Gaussian distribution (0.37 ± 0.13) is fitted on a histogram of average FRET efficiencies per FRET event. The means of the Gaussians are plotted in a separate panel (bottom) and are referred to as the FRET X fingerprint of the peptide. The FRET population on the left is caused by donor leakage into the acceptor channel. (**d**) Single-molecule FRET X histograms for each of the aSyn constructs shown in panel b. We observed six clearly separated peaks in the FRET fingerprint. The fingerprint show the center of each Gaussian fit that obtained using our FRET X approach. (**e**) The mean FRET X efficiency for each of the cysteines determined on different days. We find good reproducibility for FRET X. Mean FRET efficiencies and standard deviation are calculated from three independent experiments in panel e.

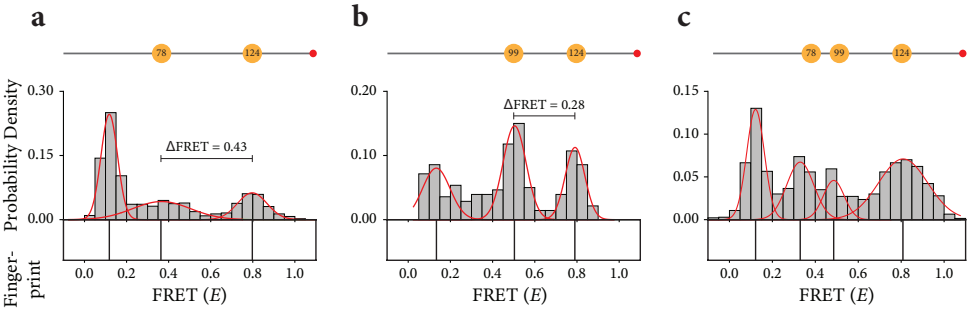


Figure 5.2: Single-molecule protein fingerprinting of aSyn.

(**a-c**) Top panels are schematic representation of the double and triple cysteine variants of the aSyn model substrate. The cysteines (orange circles) are labelled with a DNA donor docking strand, and the C terminal aldehyde motif (ref circle) is labelled with a DNA an acceptor docking strand. The aSyn constructs were engineered to have two (panels a and b) or three cysteines (panel c), one of the cysteines is at a fixed position (Cys_{124}) among the three constructs. The second and third cysteines are located at different positions in the protein substrate. Bottom panels; FRET X histograms reporting on the relative distance of the cysteines to the acceptor point. (**a**) For an aSyn construct two cysteines that are separated by 46 amino acids (Cys_{78} and Cys_{124}), we observe FRET efficiencies of 0.37 ± 0.30 and 0.80 ± 0.16. (b) Next, by decreasing the distance between the two cysteines to only 25 amino acids (panel b, Cys_{99} and Cys_{124}), we still observed two clear FRET peaks (0.51 ± 0.13 and 0.79 ± 0.11, for Cys_{99} and Cys_{124}, respectively). The FRET difference (ΔE) between the two peaks in a single mutant (0.28 for $aSyn^{Cys99+Cys124}$ and 0.43 for $aSyn^{Cys78+Cys124}$) further supports to ability of FRET X to detect the difference of two cysteine in a single aSyn protein. (c) We further increased complexity by constructing an aSyn substrate with three cysteines ($aSyn^{Cys78+Cys99+Cys124}$) and observed three clearly separated peaks. The detected FRET efficiencies of the triple aSyn construct (0.33 ± 0.15, 0.49 ± 0.11 and 0.80 ± 0.27) are similar to the FRET efficiencies observed for the single mutants (**Figure 5.1**) and double mutants (panel a and b). The FRET efficiencies are reported as the mean ± FWHM of the Gaussian fits.

The distance between the reference point and the first cysteine (Cys_{124}) is kept the same for both constructs, while the distance to the second cysteine differs (Cys_{78} for **Figure 5.2**a, and Cys_{99} for **Figure 5.2**b). Upon performing our FRET X fingerprinting assay, we observed two distinct FRET populations for both mutants (**Figure 5.2**a and b, and **Figure S5.3**). We observed a high FRET peak reporting on the relative position of Cys_{124} (0.80 for **Figure 5.2**a and 0.79 for **Figure 5.2**b). As expected, the FRET efficiency of the second cysteine differs between Cys_{78} (0.37, **Figure 5.2**a) and Cys_{99} (0.51, **Figure 5.2**b). This is further supported by the FRET difference (ΔE) between the two peaks in a single mutant, which is 0.28 for $aSyn^{Cys99+Cys124}$ and 0.43 for $aSyn^{Cys78+Cys124}$.

Next , we prepared an aSyn construct that contains all three cysteines that occur in the former two variants. On this variant we observed three FRET populations, in line with the number FRET pairs (**Figure 5.2**c). Furthermore, the FRET efficiencies for the double and triple cysteine mutants are similar to the FRET efficiencies found in our experiments with the single cysteine mutants (**Figure 5.1** and **Figure S5.2**).

5.3.2 FRET X fingerprinting of globular proteins

So far, we have used the intrinsically disordered protein aSyn as a model substrate for our experiments. However, to effectively fingerprint cellular proteins on the single-molecule level, our FRET X platform should be able to cope with the globular protein structure that most cellular proteins have.

To demonstrate the ability of FRET X for single-molecule fingerprinting of globular proteins, we expressed and purified the human apoptosis regulator Bcl-2-like protein 1 (Bcl) isoform Bcl$_{XL}$. Similar to the aSyn constructs, biotinylated DNA was conjugated to the C-terminus for immobilization and as a reference point. The Bcl-X$_L$

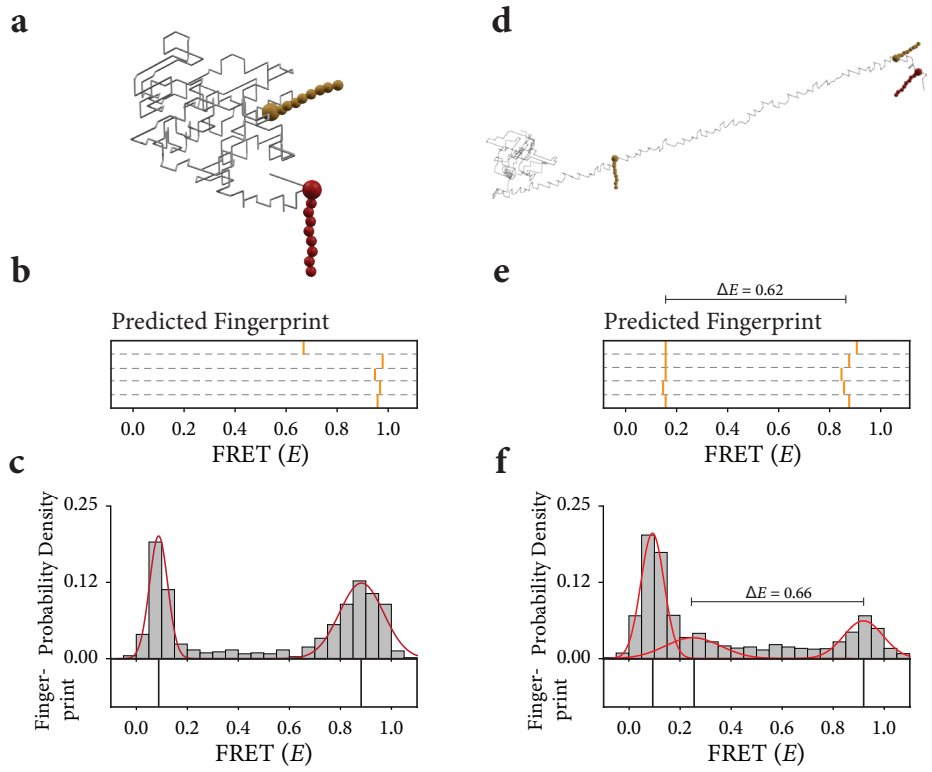


Figure 5.3: Single-molecule fingerprinting of globular proteins.
(**a** and **d**) Energy-optimized lattice model structures of model proteins Bcl-X$_L$ (panel a) and Homer-1 (panel d) with DNA-docking strands attached to cysteines (orange). The reference acceptor docking strand (red) is added to the C-terminus of the proteins. (**b** and **e**) We have simulated the protein fingerprint using our previously described fingerprint prediction tool[17], and predicted a single high FRET peak for Bcl-X$_L$ (panel b) and two FRET peaks for Homer-1 (panel e). (**c**) The experimental determined fingerprint for Bcl-X$_L$. As expected, a single high FRET peak (0.88 ± 0.20) is observed using our FRET X fingerprinting approach. (**f**) Next, we determined the FRET efficiency for Homer-1, and observed two FRET peaks (0.25 ± 0.25 and 0.91 ± 0.18) reporting on the location of the two cysteines to the acceptor strand. The ΔE is similar for the predicted (panel e, $\Delta E_{prediction}$ = 0.62) and experimental values (panel f, $\Delta E_{experimental}$ = 0.66).

protein has a single cysteine that is located close to the C-terminus of the protein (**Figure 5.3**a). Next, we used our previously developed FRET X fingerprint prediction tool[17] to obtain an expected fingerprint. For the prediction, we ran five individual simulations to estimate the fingerprint for our model protein (**Figure 5.3**b). For the Bcl$_{XL}$ model protein we predicted that the fingerprint should consist of a single high FRET peak (**Figure 5.3**b). Indeed, also experimentally we observed a distinct high FRET peak (**Figure 5.3**c and **Figure S5.4**a,c). Next, we took the postsynaptic density scaffolding protein Homer analogue 1 that contains two cysteine (**Figure 5.3**d). Our fingerprinting lattice models predicted a fingerprint that consists of a low and a high FRET peak for the Homer-1 protein (**Figure 5.3**e). After labeling and immobilization of the homer-1 protein, we flushed the imaging buffer containing all of the imager strands and observed two FRET peaks (**Figure 5.3**f and **Figure S5.4**b,d). The ΔE is similar for the predicted ($\Delta E_{prediction}$ = 0.62) and experimental values ($\Delta E_{experimental}$ = 0.66). Altogether, these results show that our FRET X fingerprinting approach is capable of obtaining reproducible fingerprints for both intrinsically disordered and globular human proteins.

5.4 Discussion

We introduced a fingerprinting approach that relies on the determination of the location of amino acids to a reference point using FRET X. By using short fluorescently labeled DNA strands and their transient binding, we can determine the relative position of multiple amino acid residues in a single molecule. FRET X allows repeated examination of an amino acid residue in a single protein, increasing the accuracy of the location determination of each residue and in turn the overall accuracy of the protein fingerprint.[17,18] We show that FRET X is capable of fingerprinting full length proteins, such as intrinsically disordered protein alpha-synuclein or globular proteins such as Bcl$_{XL}$ or Homer-1. This avoids the need for additional sample preparation steps like trypsin digestion or protein linearization and translocation that are often required for other single molecule protein identification approaches.[11,12]

We have demonstrated that FRET X can obtain unique protein fingerprints based on cysteine labeling. By using protein enrichment strategies, the here presented FRET X fingerprinting approach would be a novel approach for targeted single-molecule proteomics using cysteine fingerprint. Labeling additional residues (e.g. lysines, arginines and methionines) will improve the uniqueness of protein fingerprints[13–17] and thereby expand the dynamic range of our FRET X approach.

One of the main challenges for high throughput single molecule proteomics lies in the abundance of which different protein species are present in the cell, which can span several orders of magnitudes[2,21], due to which low abundant species can easily get masked by more abundant ones. To ensure that we can detect low abundance proteins, we must analyze a large number of proteins. Our FRET X fingerprinting approach relies on immobilization of single proteins and repetitive interrogation of each protein using transient binding probes, allowing us to fingerprint several hundreds of single molecules in a single field of view. By using more advanced TIRF--microscopes that include automated acquisition and scanning stages, the throughput

of our immobilization based single molecule protein fingerprinting approach can reach millions of protein molecules in a single run.[22] We envision that FRET X will benefit the same scaling ability and reach a similar throughput allowing for a first peak in the dark corner of the proteome.

FRET X will find applications in the medical realm, such as on-site medical diagnosis and early-stage disease diagnosis. Pharmaceutical companies could implement FRET X in their research and development pipeline for highly sensitive quality control and biomarker discovery.

5.5 Materials and Methods

5.5.1 Protein expression and purification

All proteins were codon optimized for *E. coli* BL21 (DE3) and inserted into a pET52b (+) for alpha-synuclein or pET15b for BCL2L1-XL and Homer-1 (see **Table S5.1** for full list of protein sequences). All proteins were engineered to contain an C terminal aldehyde encoding sequence. The cysteine in this motif is converted in vivo into formylglycine by co-expressed formylglycine-generating enzyme.[19,20] The plasmids and protein encoding genes were synthesized and prepared at GenScript.

The proteins were expressed in *E. coli* BL21(DE3) cells. Cultures were grown at 37 °C in LB medium supplemented with 50 µg/mL kanamycin and 50 µg/mL ampicillin until and OD_{600} of 0.5 was reached. The expression of FGE was induced with 1 % L-arabinose at 37 °C, after 30 minutes the expression of the model proteins was induced by 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG). The cultures were transferred to 26 °C to allow for expression of the proteins for 5 hours, after which the cells were harvested at 4,000 xg for minutes. The cells were lysed by resuspending the pellet in 10 mL lysis buffer (50 mM HEPES-KOH pH 7.5, 500 mM NaCl, 0.5 % Titon X-100). For the alpha-synuclein proteins, the cells were lysed by boiling the cell suspension for 15 minutes. The cells containing BCL2L1-XL and Homer-1 proteins were lysed by tumbling the cell suspension for 2 hours at room temperature, followed by sonication on ice during 6 cycles of 30 s ON and 1 min OFF at 30 % amplitude. Next, the cell lysates of each model protein were centrifuged at 30,000 xg for 30 minutes at 4 °C. The proteins were purified from the cell-free extract using HisPurTM Ni-NTA resin according to the manufacturer manual and buffer exchanged into storage buffer (50 mM HEPES-KOH pH 7.5, 150 mM NaCl, 10 % glycerol, 25 mM TCEP) using 10 kDa Amicon Ultra centrifugal filters. All proteins were aliquoted and stored at -80 °C.

5.5.2 Biomarker labeling

After purification, cysteine residues were reduced with 40-fold molar excess Tris(-2-carboethyl)phosphine (TCEP) for 30 minutes and then labeled with 25-fold molar excess monoreactive maleimide-Dibenzocyclooctyne (DBCO) (Sigma Aldrich) in 50 mM HEPES pH 7.5 150 mM NaCl 1% Triton X-100 buffer overnight at room temperature. Excess maleimide-DBCO and TCEP was removed with ZebaTM Spin desalting columns 7kDa MWCO (ThermoFisher) and the reaction buffer was changed into 50 mM HEPES pH 6.9, 150 mM NaCl, 1% Triton X-100. Then monoreactive Azidobenzoate-(5') functionalized DNA was added in 10-fold molar excess (ratio 1:10, cysteine to linker) and incubated overnight at room temperature. The formylglycine residues were acceptor-labeled with 10-fold excess biotinylated and hydrazide-functionalized DNA for 96 hours at 4 °C in a rotary shaker. Free hydrazide-DNA-biotin was removed with Ni-NTA Magnetic Agarose Beads (Qiagen) according to manufacturer's protocol. See **Table S5.2** for the full list of substrates.

5.5.3 Single-molecule setup

All experiments were performed on a custom-built microscope setup. An inverted microscope (IX73, Olympus) with prism-based total internal reflection was used. In combination with a 532 nm diode-pumped solid-state laser (Compass 215M/50mW, Coherent). A 60x water immersion objective (UPLSAPO60XW, Olympus) was used for the collection of photons from the Cy3 and Cy5 dyes on the surface, after which a 532 nm long pass filter (LDP01-532RU-25, Semrock) blocks the excitation light. A dichroic mirror (635 dcxr, Chroma) separates the fluorescence signal which is then projected onto an EM-CCD camera (iXon Ultra, DU-897U-CS0-# BV, Andor Technology). Our pixel size is 107 x 107 nm and the complete field of view is 512x256 pixels (54.8 μm x 27.4 μm) and contains ± 200 molecules. A series of EM-CDD images was recorded using custom-made program in Visual C++ (Microsoft).

5.5.4 Single-molecule data acquisition

Single-molecule flow cells were prepared as previously described.[23,24] In brief, to avoid non-specific binding, quartz slides (G. Finkerbeiner Inc) were acidic piranha etched and passivated twice with polyethylene glycol (PEG). The first round of PEGylation was performed with mPEG-SVA (Laysan Bio) and PEG-biotin (Laysan Bio), followed by a second round of PEGylation with MS(PEG)4 (ThermoFisher). After assembly of a microfluidic chamber, the slides were incubated with 20 μL of 0.1 mg/mL streptavidin (Thermofisher) for 2 minutes. Excess streptavidin was removed with 100 μL T50 (50mM Tris-HCl, pH 8.0, 50 mM NaCl). Next, 50 μL of 75 pM DNA-labeled protein was added to the microfluidic chamber. After 2 minutes of incubation, unbound protein was washed away with 200 μL T50. Then, 50 μL of 10 nM donor labeled imager strands and 50 nM acceptor labeled imager strands in imaging buffer (50 mM TrisHCl, pH 8.0, 500 mM NaCl, 0.8 % glucose, 0.5 mg/mL glucose oxidase (Sigma), 85 ug/mL catalase (Merck) and 1 mM 6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid (Trolox) (Sigma)) was injected. All single-molecule FRET experiments were performed at room temperature (23 ± 2 °C). See Table S.2 for the full list of docking and imager strands.

5.5.5 Data analysis

Fluorescence signals are collected at 0.1-s exposure time unless otherwise specified. Timetraces were subsequently extracted through IDL software using a custom script. Through a mapping file, the script collects the individual intensity hotspots in the acceptor channel and pairs them with intensity hotspots in the donor channel, after which the time traces are extracted. During the acquisition of the movie, the green laser is used to excite the Cy3 donor fluorophores. For automated detection of individual fluorescence imager strand binding events, we used a custom Python code (Python 3.7) utilizing a two-state K-means clustering algorithm on the sum of the donor and acceptor fluorescence intensities of individual molecules to identify

the frames with high intensities.[25] To avoid false positive detections, only binding events that lasted for more than three consecutive frames were selected for further analysis. FRET efficiencies for each imager strand binding event were calculated and used to build the FRET kymograph and histogram. Populations in the FRET histogram are automatically classified by Gaussian mixture modeling. The automated analysis code in Python is freely available at: https://github.com/kahutia/transient_FRET_analyzer2.

5.5.6 Protein fingerprinting simulation

A protein folding simulation was implemented to incorporate DNA-tags attached to certain residues and account for their effect on the protein structure.[17] Lattice models were used because of the far lower computational power needed for folding simulations compared to fully atomistic models allowing unrestricted movement, which is attained by reducing each amino acid to a pseudo-atom and restricting its possible positions to the vertices of a lattice. Such models have previously been used in applications where low computational requirements were essential.[26–31] The procedure starts with a fully atomistic native structure, which is converted to a lattice structure with tagged residues marked. This structure is assigned a modeling energy E_{tot}, based on interactions between pseudo-atoms located on adjacent vertices, the presence of native secondary structures and steric hindrance between pseudo-atoms and DNA-tags:

$$E_{tot} = E_{AA} + E_{sol} + E_{ss} + E_{tag} + E_{reg}$$

Here E_{AA} and E_{sol} represent the sums of pairwise residue interaction energies and residue-solvent interaction energies respectively.[32] E_{ss} rewards formation of native secondary structures and E_{dsb} rewards disulfide bridges. E_{tag} penalizes steric hindrance between DNA-tags and other pseudo-atoms. Finally E_{reg} penalizes large single-step changes in structure to better retain overall structure. E_{tot} is then minimized using a Markov chain Monte Carlo (MCMC) process, by repeatedly applying random perturbations to the structure and accepting or rejecting them based on the incurred change in the model energy. Further MCMC iterations are used to generate hundreds of slightly different structures, from which distances between donor and acceptor dye positions are deduced. These values are then translated to FRET efficiencies E_{FRET} as follows:

$$E_{FRET} = \frac{1}{1 + (R / R_0)^6}$$

Here R is the modeled inter-dye distance and R_0 is the Förster radius, which characterizes the used FRET dye pair (R_0 assumed constant at 54 Å for the Cy3-Cy5 FRET.[33]

5.6 Supporting Information

5.6.1 Supporting Figures

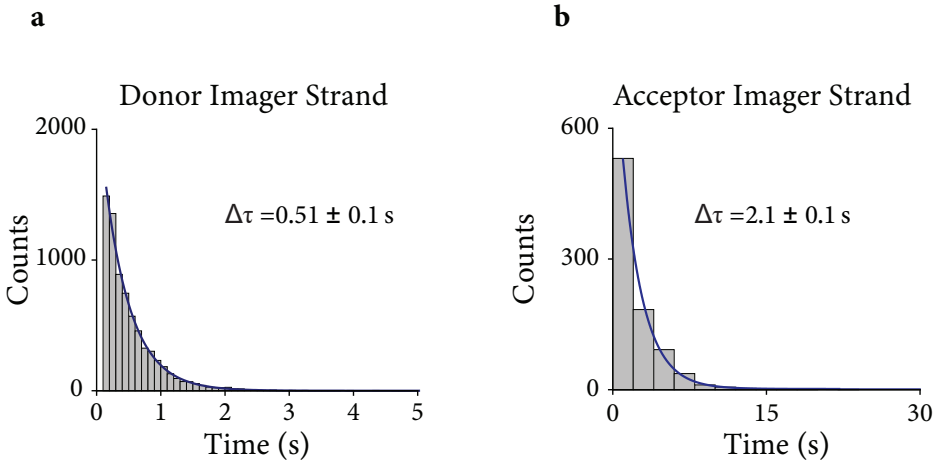


Figure S5.1: Single-molecule binding kinetics of FRET X imager strands.
(a) Dwell-time histogram for the FRET X donor imager strand (**Table S5.2**) fitted with a maximum likelihood estimation for a single exponential distribution (blue line). Average ± standard deviation of four different estimations gives: 0.51 ± 0.1 s. The number of datapoints for this distribution: n = 4288 and the protein substrate was $aSyn^{Cys78}$. **(b)** Dwell-time histogram for the FRET X acceptor imager strand (**Table S5.2**) fitted with a maximum likelihood estimation for a single exponential distribution (blue line). Average ± standard deviation of four different estimations gives: 2.1 ± 0.1 s. The number of datapoints for this distribution: n = 870 and the protein substrate was $aSyn^{Cys78}$.

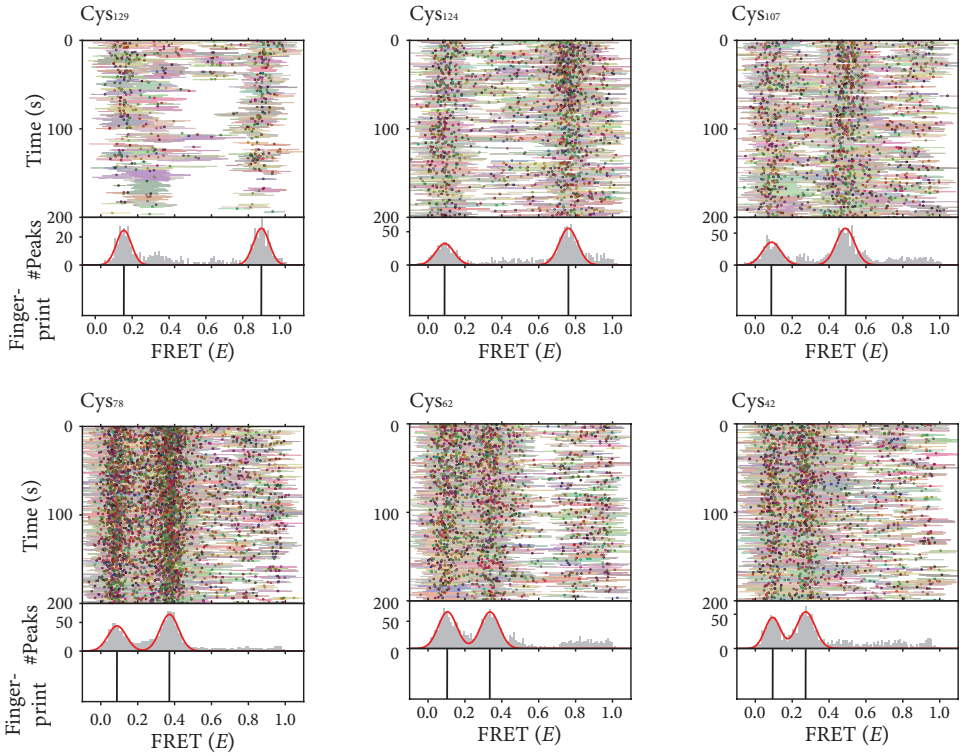


Figure S5.2: Single-molecule FRET kymographs for single cysteine mutants of alpha-synuclein. Ensemble FRET kymographs obtained for each of the different aSyn mutants measured in **Figure 5.1**. We observed FRET efficiencies of 0.88 ± 0.12 for Cys129, 0.76 ± 0.16 for Cys124, 0.49 ± 0.16 for Cys107, 0.37 ± 0.13 for Cys78, 0.32 ± 0.13 for Cys62, and 0.27 ± 0.14 for Cys42. The FRET efficiencies are reported as the mean ± FWHM of the Gaussian fits.

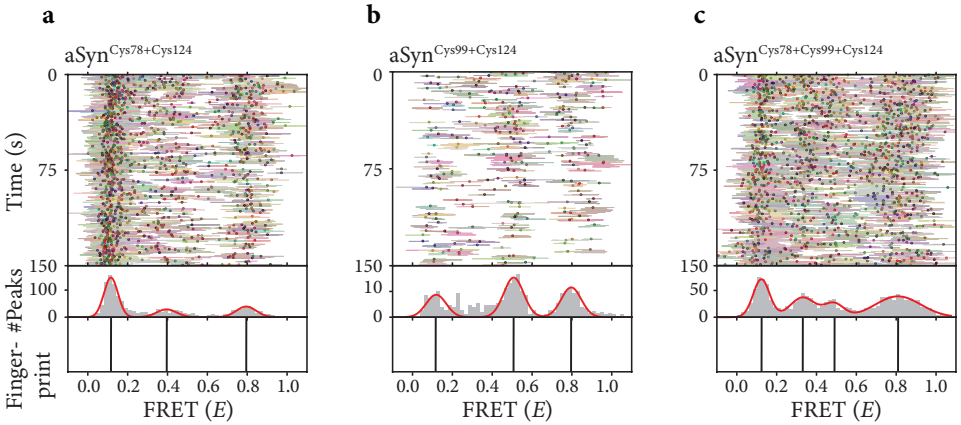


Figure S5.3: Single-molecule FRET kymographs for double and triple cysteine alpha-synuclein constructs.

(**a-c**) Ensemble FRET kymographs obtained for each of the different aSyn mutants measured in**Figure 5.2**. (**a**) We observed FRET efficiencies of 0.37 ± 0.16 and 0.80 ± 0.16 for aSyn$^{Cys78+Cys124}$. (**b**) Next, we designed an construct in which the cysteines were closer together, and observed FRET efficiencies of 0.50 ± 0.13 and 0.79 ± 0.11 for aSyn$^{Cys99+Cys124}$. (**c**) For the aSyn construct with all three cysteines, aSyn$^{Cys78+Cys99+Cys124}$, we observed three distinct FRET population with efficiencies of 0.33 ± 0.15, 0.49 ± 0.11, and 0.81 ± 0.27, reporting on the distance of each of the cysteines in respect to the C-terminal acceptor strand. The FRET efficiencies are reported as the mean ± FWHM of the Gaussian fits.

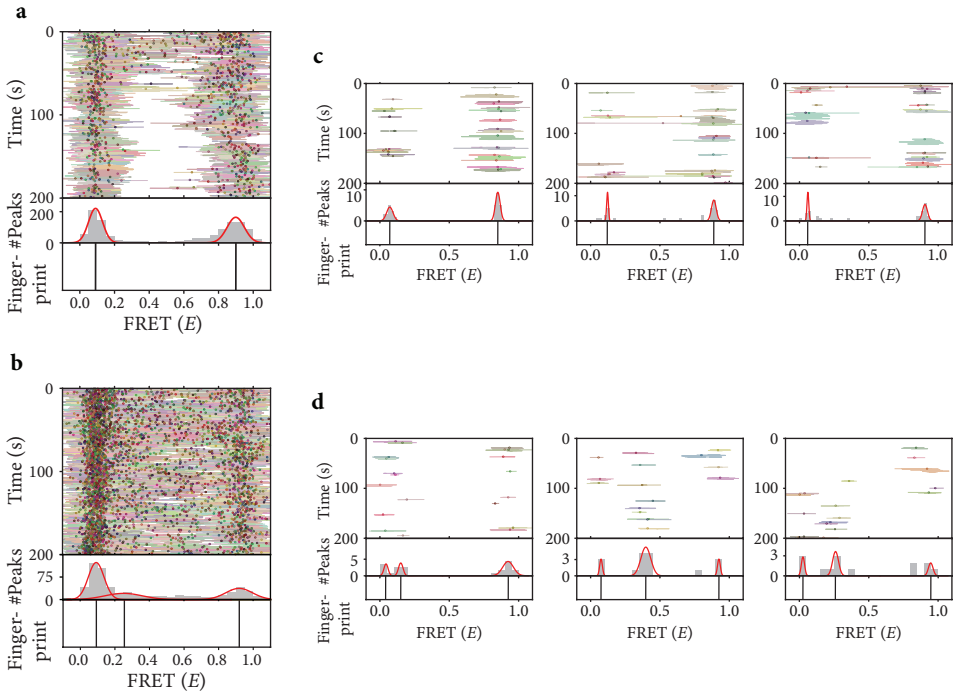


Figure S5.4: Single-molecule FRET X analysis of globular protein substrates.
(**a,b**) Ensemble FRET kymographs obtained for globular model protein Bcl-Xʟ and Homer-1. For Bcl-Xʟ we observed a single high FRET peak with an efficiency of 0.88 ± 0.20 (panel a). The ensemble FRET kymograph for Homer-1 shows two FRET populations, with FRET efficiencies of 0.25 ± 0.26 and 0.91 ± 0.18 (panel b). (**c,d**) Representative FRET kymographs obtained from single Bcl-Xʟ (panel c) and Homer-1 (panel d) proteins. The kymographs show repetitive interogation of the FRET X fingerprint on the single molecule level. The distribution of each individual molecule can be fitted with high precision (s.d. ≤ 0.03 for each distribution).

5.6.2 Supporting Tables

Table S5.1: Amino acid sequences of model proteins.

Construct	Amino Acid Sequence (N → C)	Supplier
Alpha-Synuclein	MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHG-VATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEE-GAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEALCTPSRYQDPVQVDAAAEL-ALVPRGSSAHHHHHHHHHH	Genscript
Bcl-XL	MGSSHHHHHHSSGLVPRGSHMSQSNRELVVDFLSYKLSQKGYSWSQFSDVEEN-RTEAPEGTESEMETPSAINGNPSWHLADSPAVNGATGHSSSLDAREVIP-MAAVKQALREAGDEFELRYRRAFSDLTSQLHITPGTAYQSFEQVVNELFRDGV-NWGRIVAFFSFGGALCVESVDKEMQVLVSRIAAWMATYLNDHLEPWIQENG-GWDTFVELYGNNAAAESRKGQERFNRWFLTGMTVAGVVLLGSLFSRKLCTPSR	Genscript
Homer-1	MGSSHHHHHHSSGLVPRGSHMGEQPIFSTRAHVFQIDPNTKKNWVPTSKHAVT-VSYFYDSTRNVYRIISLDGSKAIINSTITPNMTFTKTSQKFGQWADSRANT-VYGLGFSSEHHLSKFAEKFQEFKEAARLAKEKSQEKMELTSTPSQESAGGD-LQSPLTPESINGTDDERTPDVTQNSEPRAEPTQNALPFSHSSAISKHWEAELAT-LKGNNAKLTAALLESTANVKQWKQQLAAYQEEAERLHKRVTELECVSSQAN-AVHTHKTELNQTIQELEETLKLKEEEIERLKQEIDNARELQEQRDSLTQKLQEV EIRNKDLEGQLSDLEQRLEKSQNEQEAFRNNLKTLLEILDGKIFELTELRDN-LAKLLECSEFELRRQAGLCTPSR	Genscript

Table S5.2: DNA Constructs.

DNA Strand	Nucleotide Sequence (5' → 3')	Modification	Supplier
Donor imager strand	CTCCTC	3' Cy3	Ella Biotech (GmbH)
Acceptor imager strand	TAATGAAGA	3' Cy5	Ella Biotech (GmbH)
Donor docking sequence	TATACATCTAT	5' Azidobenzoate	Biomers.net (GmbH)
Acceptor docking sequence	TTCTTCATTACT	5' Hydrazide	Biomers.net (GmbH)

5.7 References

1 Smith, L. M. et al. Proteoform: a single term describing protein complexity. Nat. Methods 10, 186–187 (2013).

2 Aebersold, R. et al. How many human proteoforms are there? Nat. Chem. Biol. 14, 206–214 (2018).

3 Kim, H. K. et al. Alternative splicing isoforms in health and disease. doi:10.1007/s00424-018-2136-x.

4 Paronetto, M. P., Passacantilli, I. & Sette, C. Alternative splicing and cell survival : from tissue homeostasis to disease. 23, 1919–1929 (2016).

5 Lin, H. & Caroll, K. S. Introduction: Posttranslational Protein Modification. Chem. Rev. 118, 887–888 (2018).

6 Leney, A. C. & Heck, A. J. R. Native Mass Spectrometry: What is in the Name? J. Am. Soc. Mass Spectrom. 28, 5–13 (2017).

7 Mehmood, S., Allison, T. M. & Robinson, C. V. Mass spectrometry of protein complexes: from origins to applications. Annu. Rev. Phys. Chem 66, 453–474 (2015).

8 Tighe, P. J., Ryder, R. R., Todd, I. & Fairclough, L. C. ELISA in the multiplex era: Potentials and pitfalls. Proteomics - Clin. Appl. 9, 406–422 (2015).

9 Zubarev, R. a. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 13, 723–726 (2013).

10 Pagel, O., Loroch, S., Sickmann, A. & Zahedi, R. P. Current strategies and findings in clinically relevant post-translational modification-specific proteomics. Expert Rev. Proteomics 12, 235–253 (2015).

11 Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. Nat. Nanotechnol. 13, 786–796 (2018).

12 Alfaro, J. A. et al. The emerging landscape of single-molecule protein sequencing technologies. Nat. Methods 18, 604–617 (2021).

13 Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A Theoretical Justification for Single Molecule Peptide Sequencing. PLoS Comput. Biol. 11, 1–17 (2015).

14 Yao, Y., Docter, M., Van Ginkel, J., De Ridder, D. & Joo, C. Single-molecule protein sequencing through fingerprinting: Computational assessment. Phys. Biol. 12, 10–16 (2015).

15 Ohayon, S., Girsault, A., Nasser, M., Shen-Orr, S. & Meller, A. Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. PLoS Comput. Biol. 15, 1–21 (2019).

16 Rodriques, S. G., Marblestone, A. H. & Boyden, E. S. A theoretical analysis of single molecule protein sequencing via weak binding spectra. PLoS One 14, e0212868 (2019).

17 Lannoy, C. de, Filius, M., Wee, R. van, Joo, C. & Ridder, D. de. Evaluation of FRET X for Single-Molecule Protein Fingerprinting. bioRxiv 2021.06.30.450512 (2021) doi:10.1101/2021.06.30.450512.

18 Filius, M., Kim, S. H., Severins, I. & Joo, C. High-Resolution Single-Molecule FRET via DNA eXchange (FRET X). Nano Lett. 21, 3295–3301 (2021).

19 Carrico, I. S., Carlson, B. L. & Bertozzi, C. R. Introducing genetically encoded aldehydes into proteins. Nat. Chem. Biol. 3, 321–322 (2007).

20 Shi, X. et al. Quantitative fluorescence labeling of aldehyde-tagged proteins for single-molecule imaging. Nat. Methods 9, 499–503 (2012).

21 Anderson, N. L. & Anderson, N. G. The Human Plasma Proteome HISTORY, CHARACTER, AND DIAGNOSTIC PROSPECTS. Mol. Cell. Proteomics 1, 845–867 (2002).

22 Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. Nat. Biotechnol. 36, 1076–1082 (2018).

23 Chandradoss, S. D. et al. Surface passivation for single-molecule protein studies. J. Vis. Exp. 4–11 (2014) doi:10.3791/50549.

24 Filius, M. et al. High-Speed Super-Resolution Imaging Using Protein-Assisted DNA-PAINT. Nano Lett. (2020) doi:10.1021/acs.nanolett.9b04277.

25 Kim, S. H., Kim, H., Jeong, H. & Yoon, T.-Y. Encoding Multiple Virtual Signals in DNA Barcodes with Single-Molecule FRET. Nano Lett. 21, 1694–1701 (2021).

26 Kolinski, A. & Skolnick, J. Reduced models of proteins and their applications. doi:10.1016/j.polymer.2003.10.064.

27 Abeln, S., Vendruscolo, M., Dobson, C. M. & Frenkel, D. A simple lattice model that captures protein folding, aggregation and amyloid formation. PLoS One 9, 85185 (2014).

28 Coluzza, I., Muller, H. G. & Frenkel, D. Designing refoldable model molecules. Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top. 68, 046703 (2003).

29 Bianco, V., Pagès-Gelabert, N., Coluzza, I. & Franzese, G. How the stability of a folded protein depends on interfacial water properties and residue-residue interactions. J. Mol. Liq. 245, 129–139 (2017).

30 Dijkstra, M. J. J., Fokkink, W. J., Heringa, J., van Dijk, E. & Abeln, S. The characteristics of molten globule states and folding pathways strongly depend on the sequence of a protein. Mol. Phys. 116, 3173–3180 (2018).

31 Van Gils, J. H. M. et al. The hydrophobic effect characterises the thermodynamic signature of amyloid fibril growth. PLoS Comput. Biol. 16, e1007767 (2020).

32 Miyazawa, S. & Jernigan, R. L. Self-Consistent Estimation of Inter-Residue Protein Contact Energies Based on an Equilibrium Mixture Approximation of Residues. doi:10.1002/(SICI)1097-0134(19990101)34:1.

33 Lerner, E. et al. Toward dynamic structural biology: Two decades of single-molecule förster resonance energy transfer. Science (80-.). 359, (2018).

5

6

High-Speed Super-Resolution Imaging Using Protein-Assisted DNA-PAINT

Mike Filius*, Tao Ju Cui*, Adithya N. Ananth, Margreet W. Docter, Jorrit W. Hegge, John van der Oost, and Chirlmin Joo

* These Authors have contributed equally to this work

Nano Letters
Nano Lett. 2020, 20, 4, 2264–2270
DOI: 10.1021/acs.nanolett.9b04277



6.1 Abstract

Super-resolution imaging allows for visualization of cellular structures on a nanoscale level. DNA-PAINT (DNA Point Accumulation In Nanoscale Topology) is a super-resolution method that depends on the binding and unbinding of DNA imager strands. The current DNA-PAINT technique suffers from slow acquisition due to the low binding rate of the imager strands. Here we report on a method where imager strands are loaded into a protein, Argonaute (Ago), that allows for faster binding. Ago pre-orders the DNA imager strand into a helical conformation, allowing for 10 times faster target binding. Using a 2D DNA origami structure, we demonstrate that Ago-assisted DNA-PAINT (Ago-PAINT) can speed up the current DNA-PAINT technique by an order of magnitude while maintaining the high spatial resolution. We envision this tool to be useful for super-resolution imaging and other techniques that rely on nucleic-acid interactions.

6.2 Introduction

Single-molecule localization microscopy techniques allow researchers to image cellular structures that are not visible through diffraction-limited microscopy methods. Most single-molecule localization techniques rely on the stochastic blinking of fluorescent signal, by using photoswitchable fluorophores as in photoactivated-localization microscopy (PALM)[1] and (direct) stochastic optical reconstruction microscopy ((d)STORM).[2] An alternative approach to achieve stochastic blinking is through fluorescent probes that transiently bind their target, as in point accumulation in nanoscale topography (PAINT).[3–5]

In DNA-PAINT, a fluorophore is attached to a short DNA oligonucleotide, namely an imager strand that specifically binds to a complementary target DNA sequence, namely a docking strand.[6] The stochastic blinking of signals is achieved through binding and unbinding of the incoming imager strands to the docking strands and is imaged using total internal reflection fluorescence (TIRF). By changing the length and sequence of an imager strand, one can tune the on- and off-rates of the imager and adjust the specificity. This allows for high multiplexing capabilities since the number of probes is only limited by the number of orthogonal DNA sequences. Furthermore, compared to conventional super-resolution techniques, DNA-PAINT comes with the advantage that imager strands are continuously replenished from the solution and thus photobleaching is circumvented during the imaging process.

A critical limitation of DNA-PAINT, however, is the low binding rate of DNA, which is typically in the order of 10^6 M^{-1} s^{-1}. Given this binding rate, obtaining images with high spatial resolution (5 nm) usually takes several hours.[7–9] Shorter acquisition times can be achieved by increasing concentration of the imager strand. However, single-molecule binding events become unresolvable from the background of unbound imager strands, even when TIRF is used. To reduce this acquisition time, DNA-PAINT was recently combined with single-molecule Förster Resonance Energy Transfer (smFRET).[10,11] This, however, comes at a cost of reduced spatial resolution due to reduced energy transfer efficiency and due to limited choice of dyes. Here we describe an alternative approach, in which protein-assisted delivery of imager strands is demonstrated to speed up the acquisition time 10-fold and only to require a single fluorescence channel.

Argonaute proteins (Agos) are a class of enzymes that utilize a DNA or RNA guide to find a complementary target, either to inactivate or to cleave it. In eukaryotes, an RNA guide directs Ago to complementary RNA targets for post-transcriptional regulation.[12] Ago proteins initially bind their target through base pairing with the seed segment of the guide (nucleotides 2-7 for human Ago).[13–15] Crystal structures have revealed that Ago pre-orders this seed segment into a helical conformation, allowing for the formation of a double helix between guide and target, and hence effectively pre-paying the entropic cost of target binding.[16,17] This results in binding rates that are near-diffusion limited (~10^7 M^{-1} s^{-1}).[18–21] In prokaryotes, there is a broad diversity of Agos with respect to the identity of their guide (RNA/DNA) and their target (RNA/DNA).[22,23] Some well-characterized prokaryotic Ago nucleases (TtAgo, CbAgo) use DNA guides to target single-stranded (ss)DNA.[24,25]

Here we describe a new DNA-PAINT method based on protein-assisted delivery of DNA imager strands, which allows for faster acquisition of super-resolved nanostructures. We use a wild type Ago protein from the bacterium Clostridium butyricum (CbAgo) to speed up the kinetic binding of DNA imager strands. CbAgo reshapes the binding landscape of the imager strand, resulting in a 10-fold higher binding rate compared to conventional DNA-PAINT. Ago-PAINT can be implemented without additional complexity whilst retaining the programmability and specificity of DNA-PAINT, due to the favourable targeting feature of CbAgo.[25,26] We determine the spatial resolution of Ago-PAINT through the use of 2D DNA origami structures and show that Ago-PAINT generates super-resolution images of diffraction limited structures at least 10-fold faster than conventional DNA-PAINT.

6.3 Results

For high-quality super-resolution images, a PAINT-based method requires more than five transient binding events per localization spot[7], each with a dwell time of at least several hundreds of milliseconds.[7–9] A typical 8-nt DNA-PAINT imager strand exhibits an on-rate of ~ 10^6 M^{-1} s^{-1} and a dwell time (= 1/off-rate) of ~1 s.[9] DNA-PAINT experiments use an imager strand concentration between 1-10 nM. This range is chosen to be high enough to obtain a sufficient number of binding events during the acquisition time, but not too high to avoid cross-talk localization between structures.[7]

We determined the on- and off-rates of Ago-PAINT imager strands and compared these to the on- and off-rate of conventional DNA-PAINT with the same imager strands using a smFRET assay (**Figure 6.1**). Acceptor (Cy5)-labelled ssDNA targets were immobilized through biotin-streptavidin conjugation on a PEGylated quartz slide. Next, either donor (Cy3)-labelled 8-nt DNA-PAINT imager strands or Ago-PAINT imager strands (CbAgo loaded with a Cy3-labelled guide) were injected, and their interactions with the immobilized target strand were probed using TIRF microscopy (**Figure 6.1**a). The assay was designed to give a high-FRET signal upon specific binding of either DNA imager strand or Ago-guide complex to the complementary target (**Figure 6.1**b and c). The Cy3 position was picked the same as in previous studies with CbAgo, to prevent any photophysical artefacts from occurring.[26,27] The time between introduction of the imager strands and the first binding event is the arrival time (which is the inverse of the on-rate, kon). The duration of the FRET binding events is the dwell time (**Figure 6.1**b).

For a comparison between Ago-PAINT and DNA-PAINT, we designed an 8-nt DNA-PAINT imager strand (**Figure 6.1**c) and found that under our experimental conditions the average dwell time of this imager strand is 1.1 ± 0.2 s (**Figure 6.1**d). Next, we sought to find an Ago-PAINT guide with a similar dwell time. The first nucleotide of an Ago guide is embedded within the protein structure (**Figure S6.1**a).[16,17] Therefore, we determined the dwell time of Ago-guide complexes with different numbers (N) of base pairing with the target starting from the second nucleotide onwards (**Figure S6.1**b). A guide with N=5 (nt 2-6) base-pairing to the target exhibited a comparable dwell time of 1.2 ± 0.2 s (**Figure 6.1**e). We observed that for Ago-PAINT the apparent binding rate is influenced by the number of base pairs that

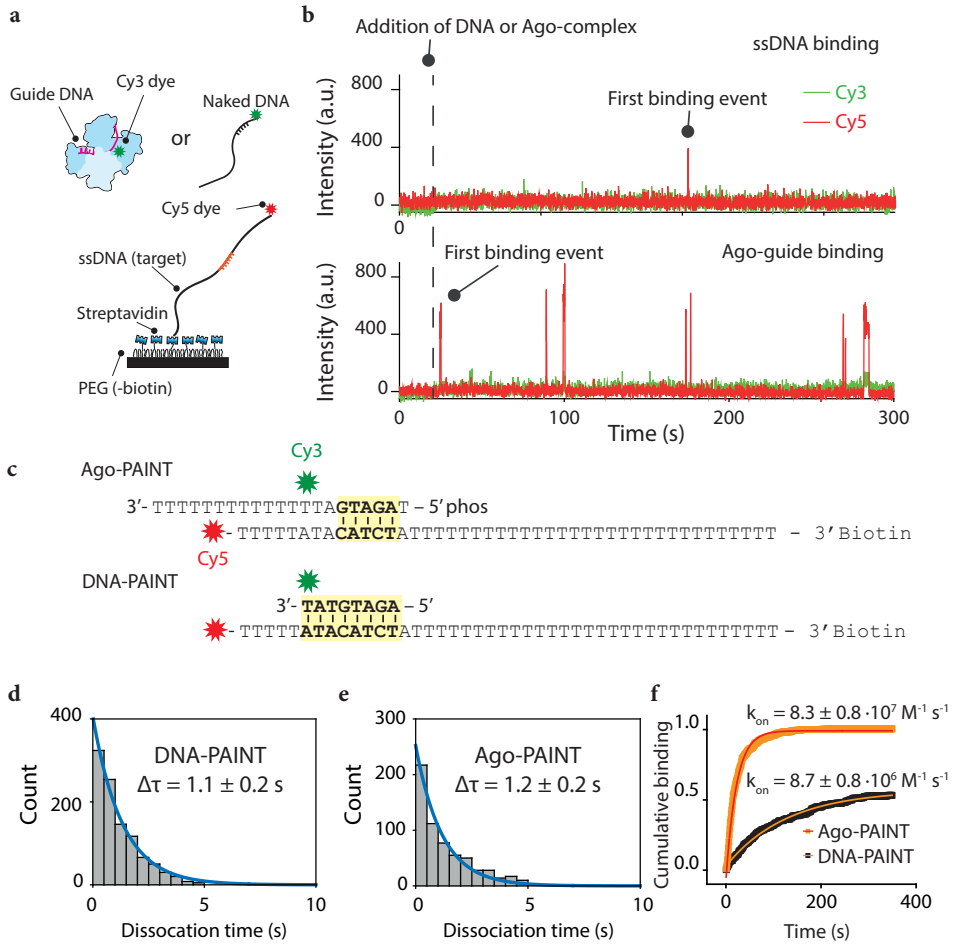


Figure 6.1: Single-molecule FRET assay to quantify binding kinetics Ago vs DNA-PAINT

(a) A schematic of the single-molecule FRET assay with the target strand immobilized on a PEGylated surface through biotin-streptavidin conjugation. The green and red stars indicate the Cy3 and Cy5 dye respectively. Binding of Ago-guide complex or ssDNA probe to the ssDNA target results in high FRET signal. **(b)** Representative traces of ssDNA binding (top) and Ago-complex binding (bottom). The dashed line indicates the timepoint at which Ago-guide or DNA is introduced inside the microfluidic chamber. **(c)** A schematic of the sequences used for Ago-PAINT and DNA-PAINT. Upon binding, both constructs will give rise to a high FRET signal. **(d)** Dwell-time histogram ($\Delta\tau$) of ssDNA (sequence shown in panel c). Maximum likelihood estimation (MLE) gives 1.1 ± 0.2 s as the parameter for a single-exponential distribution (blue line). Number of data-points: 1029. **(e)** Dwell-time histogram ($\Delta\tau$) of Ago (sequence shown in panel a). MLE fitting gives 1.2 ± 0.2 s as the parameter for a single-exponential distribution (blue line). Number of data-points: 696. **(f)** Cumulative binding event plots of DNA-PAINT (Black) and Ago-PAINT (Orange) vs time. A single-exponential fit is used for DNA-PAINT (red line) and Ago-PAINT (orange line). Errors in (d), (e) and (f) are determined by taking the 95% confidence interval of 105 bootstraps.

are formed between the guide and its target. For N=5 or larger, the on-rate reaches a saturated value (kon = 0.6-1.0 · 10^8 M^{-1} s^{-1}) (**Figure S6.1**c). Those values are 10 times higher than the typical on-rates for an 8-nt DNA-PAINT imager strand, 8.7 ± 0.8 ·10^6 M^{-1} s^{-1} (**Figure 6.1**f).

To demonstrate the use of Ago-PAINT for super-resolution imaging, we designed a rectangular 2-dimensional DNA origami structure of 76 nm x 80 nm (**Figure 6.2** and **Figure S6.2**). The DNA origami structure has four docking sites that are spaced 61 nm x 68 nm apart (**Figure 6.2**a). To achieve optimal Ago binding to the DNA origami docking strands, we introduced a polyT linker between the target sequence of Ago and the DNA origami structure (**Figure 6.2**a, right panel). As our previous

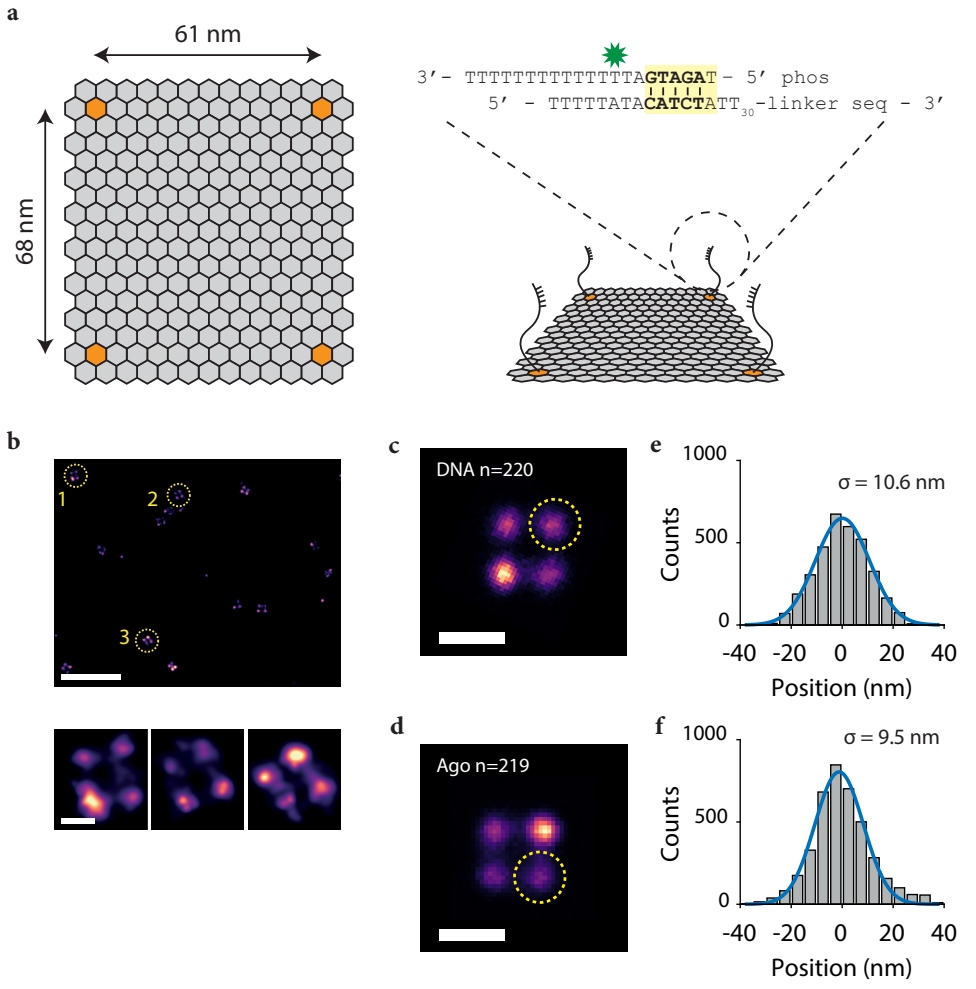


Figure 6.2: Ago-PAINT enables the same localization precision as conventional DNA-PAINT
(a) Left: A schematic design of the 2D-DNA origami structure. The orange honeycombs indicate the approximate locations of binding sites. Right: 3D representation of the imaging scheme with the used docking strand sequence. The green star indicates the position of the Cy3 dye labelled on the backbone of an amino-modified thymine. **(b)** A representative super-resolution image showcasing

binding sites of the 2D-DNA origami structures using Ago-PAINT. Bottom: Super-resolution reconstruction of the four-corner origami structures of the top panel. **(c)** A summed image of 220 origami structures visualized through the use of DNA-PAINT. **(d)** A summed image of 219 origami structures made through the use of Ago-PAINT. The concentration of imager strand was 1 nM for both conventional DNA-PAINT and Ago-PAINT. **(e)** Fitting of a cross-sectional intensity histogram from the yellow encircled area in panl C to a Gaussian (blue line) shows that a localisation precision of 10.6 nm can be achieved, similar to Ago-PAINT. **(f)** Fitting of a cross-sectional intensity histogram from the yellow encircled area in Figure D to a Gaussian (blue line) shows that a localisation precision of 9.5 nm is possible under these conditions. Scale bars in (b) indicate 500 nm (top) and 50 nm (bottom three). Scale bars in (c) and (d) indicate 100 nm.

observation with CbAgo[26] suggested that the protein occupies a footprint of around 20 nt, we made the length of polyT to be 30 nt.

Next, we sought to compare the localization precision of Ago-PAINT and DNA-PAINT. We tested our Ago-PAINT approach by injecting guide-loaded Ago into our flow cell in which DNA origami structures were immobilized. A super-resolution image could be reconstructed from the Ago-PAINT data which revealed four detectable spots on the origami structures as expected from our assay design (**Figure 6.2**b). We determined the localization precision by selecting 220 origami structures for DNA-PAINT and 219 structures for Ago-PAINT and created a sum image using the Picasso analysis software[7] (**Figure 6.2**c and d). The localization precision was determined by plotting the cross-sectional histogram of one of the four binding sites of the summed DNA origami structure. For DNA-PAINT this resulted in a localization precision of 10.6 nm (**Figure 6.2**e) and for Ago-PAINT we found a localization precision of 9.5 nm (**Figure 6.2**f). The histogram demonstrates that Ago-PAINT delivers the same quality of localization precision when compared to the DNA-PAINT approach. Nearest neighbour analysis[28] reconfirms that a localization precision is similar for both Ago-PAINT and DNA-PAINT (**Figure S6.3**). Additionally, we probed the possibility to use different linker lengths for Ago-PAINT imaging. When we tested DNA origami structures with longer linkers (50 thymines or 100 thymine nucleotides), we found that this did not affect the localization precision of Ago-PAINT (**Figure S6.4** and **Figure S6.5**), showing that Ago-PAINT is compatible with various linker lengths (≥ T30).

Finally, we compared the speed of super-resolution imaging through Ago-PAINT with the conventional DNA-PAINT approach using the 2D DNA origami structures as a testing platform. We evaluated the quality of a super-resolution image after each time point for both Ago-PAINT and DNA-PAINT (**Figure 6.3**a). The overall resolution of a single-molecule localization microscopy image is dependent on the number of localizations per docking strand. Therefore, to quantify the speed of imaging we plot the standard error of the localization precision as a function of frame number (**Figure 6.3**b) where we took the sigma values from **Figure 6.2**e and f as the localization precision. We observed that the standard error of the localization precision for Ago-PAINT is smaller than that of DNA-PAINT at each time point, indicating that super-resolved images of identical resolution will be obtained 10x faster through Ago-PAINT compared to DNA-PAINT. This result is further supported by the intensity vs time traces, which shows that our Ago-PAINT method results in

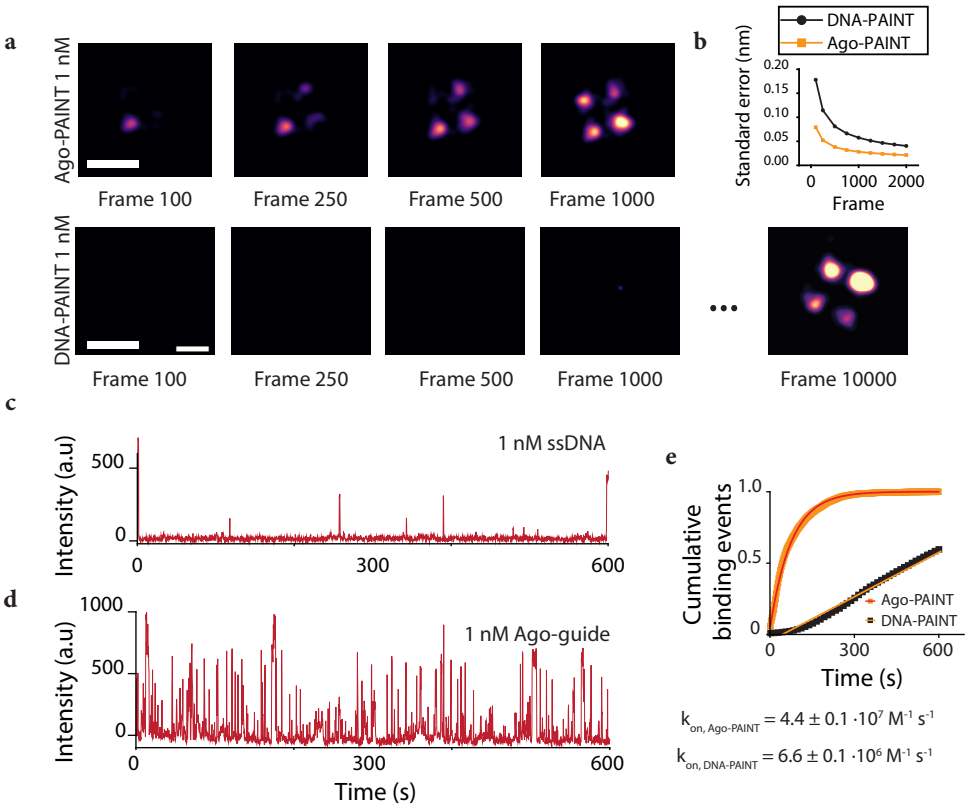


Figure 6.3: Ago-PAINT enables fast imaging of super-resolved structures.
(a) Snapshots in time for Ago-PAINT (top) and DNA-PAINT (bottom) showing super-resolution images being formed over time. Exposure time: 0.3 s. The same color scale is used for the intensity in all images. **(b)** Standard error of Ago-PAINT vs DNA-PAINT plotted versus frame number. **(c)** Representative intensity vs time data trace of DNA-PAINT at 1 nM DNA concentration shows few binding events occurring within 600 s. The raw data trace is taken from a single origami plate. **(d)** Representative intensity vs time data trace of 1 nM Ago-guide complex shows binding events occurring frequently within 600 s. The raw data trace is taken from a single origami plate. **(e)** Normalized cumulative distribution of dark times (the time between binding events) for DNA-PAINT (black, n = 4870) and Ago-PAINT (orange, n = 5793). A single-exponential growth curve (red for DNA-PAINT, orange for Ago-PAINT) is used to estimate the binding rate. Scale bars in (A) indicate 100 nm.

more binding events compared to DNA-PAINT approach, under similar conditions with DNA concentrations of 1 nM (**Figure 6.3**c-e and **Figure S6.6**). The on-rates for both Ago-PAINT (kon = 4.4 ± 0.1 ·10^7 M^{-1} s^{-1}) and DNA-PAINT (kon = 6.6 ± 0.1 ·10^6 M^{-1} s^{-1}) on our DNA-origami structure (**Figure 6.3**e) are similar to the on-rates that we found in our single-molecule experiments (**Figure 6.1**f).

6.4 Discussion

Here we presented a proof-of-concept of Ago-PAINT that allows for rapid super--resolution imaging. We demonstrated that fast Ago-PAINT recording can be used to acquire super-resolution images of nanostructures while retaining the programmability and predictability of DNA-PAINT.

For the visualization of several complex cellular components in a single cell, multiplexing super-resolution is highly anticipated. Recent developments allow for temporal[29] and spectral[30] multiplexing of DNA-PAINT, and we believe that these methods can be integrated with Ago-PAINT. And in our previous work, we showed that different guide sequences resulted in distinctly different binding kinetics.[25] This kinetic fingerprinting will allow for additional freedom when designing multiplexing Ago-PAINT.[29,31] Furthermore, optimization of the imager sequence and imaging conditions allowed for further increase in acquisition time for DNA-PAINT.[32] Although the binding kinetics or Ago-PAINT are near diffusion limited, we expect that optimization of the guide sequence could further improve the kinetics Ago-PAINT.

In this study, Ago-PAINT experiments are performed with the wild-type CbAgo protein which substantially increases the probe size compared to conventional DNA-PAINT. However, successful applications of Argonaute proteins for in vivo gene silencing[33,34] hint that our Ago-PAINT approach could be used in cellular super-resolution imaging. While targeting complex cellular structures in cells could be an issue with a full size CbAgo, it is possible to use truncated versions of Ago. Some truncated versions of approximately half the size (short Agos) exist in nature.[22] We speculate that it will be possible to truncate them further as Ago-PAINT only relies on the property of pre-forming the helix structure of the imager strand. For example, an Ago variant from Kluyveromyces polysporus that contains only the C-lobe was reported to retain almost all the binding properties of the untruncated version.[35] Furthermore, as the imager strand is loaded and protected inside the protein, degradation of the imager strand is less likely to occur over time, unlike oligos that are rapidly digested.[36]

In this paper we demonstrated the use of CbAgo for super-resolution microscopy. While this CbAgo targets ssDNA, Agos from other species can target RNA.[22] For example, the Ago from Marinitoga piezophila (MpAgo)[37,38] targets RNA and one could harness the property of a high association rate for other single molecule imaging applications such as RNA sensing. Recently, dTtAgo has been combined with FISH[39] to allow for labelling of genomic loci in fixed cells. We anticipate the use of RNA guided Agos for a significant speed-up in similar applications for RNA FISH. Lastly, complementary approaches such as DNA-based STED imaging[40], qPAINT[41] or crosslinking on single-molecule target using Action-PAINT[42] could be combined with our Ago-PAINT approach. We envision the use of Ago-PAINT as a general toolkit to speed up many current existing applications that rely on base-pairing interactions.

6.5 Materials And Methods

6.5.1 Expression and purification of CbAgo

The CbAgo gene was codon harmonized for *E. coli* Bl21(DE3) and inserted into a pET-His6 MBP TEV cloning vector (Addgene plasmid #29656) using ligation independent cloning. The CbAgo protein was expressed in Bl21(DE3) Rosetta™ 2 (Novagen). Cultures were grown at 37 °C in LB medium containing 50 μg ml^{-1} kanamycin and 34 μg ml^{-1} chloramphenicol till an OD_{600} nm of 0.7 was reached. CbAgo expression was induced by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.1 mM. During the expression cells were incubated at 18 °C for 16 h with continues shaking. Cells were harvested by centrifugation and lysed, through sonication (Bandelin, Sonopuls. 30% power, 1 s on/2 s off for 5 min) in lysis buffer containing 20 mM Tris-HCl pH 7.5, 250 mM NaCl, 5 mM imidazole, supplemented with a EDTA free protease inhibitor cocktail tablet (Roche). The soluble fraction of the lysate was loaded on a nickel column (HisTrap Hp, GE healthcare). The column was extensively washed with wash buffer containing 20 mM Tris-HCl pH 7.5, 250mM NaCl and 30 mM imidazole. Bound protein was eluted by increasing the concentration of imidazole in the wash buffer to 250 mM. The eluted protein was dialysed at 4 °C overnight against 20 mM HEPES pH 7.5, 250 mM KCl, and 1 mM dithiothreitol (DTT) in the presence of 1 mg TEV protease (expressed and purified according to Tropea et al.[43]) to cleave of the His_6-MBP tag. Next the cleaved protein was diluted in 20 mM HEPES pH 7.5 to lower the final salt concentration to 125 mM KCl. The diluted protein was applied to a heparin column (HiTrapHeparin HP, GE Healthcare), washed with 20 mM HEPES pH 7.5, 125 mM KCland eluted with a linear gradient of 0.125–2 M KCl. Next, the eluted protein was loaded onto a size exclusion column (Superdex 200 16/600 column, GE Healthcare) and eluted with 20 mM HEPES pH 7.5, 500 mM KCl and 1 mM DTT. Purified CbAgo protein was diluted in size exclusion buffer to a final concentration of 5 μM. Aliquots were flash frozen in liquid nitrogen and stored at −80 °C.

6.5.2 Single-molecule setup

All experiments were performed on a custom-built microscope setup. An inverted microscope (IX73, Olympus) with prism-based total internal reflection is used. In combination with a 532 nm diode laser (Compass 215M/50mW, Coherent). A 60x water immersion objective (UPLSAPO60XW, Olympus) was used for the collection of photons from the Cy3 and Cy5 dyes on the surface, after which a 532 nm long pass filter (LDP01-532RU-25, Semrock) blocks the excitation light. A dichroic mirror (635 dcxr, Chroma) separates the fluorescence signal which is then projected onto an EM-CCD camera (iXon Ultra, DU-897U-CS0-#BV, Andor Technology). A series of EM-CDD images was recorded using custom-made program in Visual C++ (Microsoft). Time traces were extracted from the EM-CDD images using IDL (ITT Visual Information Solution) and further analyzed with Matlab (Mathworks) and Origin (Origin Lab).

6.5.3 Single-molecule data acquisition

To avoid non-specific binding of CbAgo protein to the surface, quartz slides were PEGylated as previously described (Chandradoss 2014). Briefly, acidic piranha etched quartz slides (Finkenbeiner) were passivated twice with polyethylene glycol (PEG). The first round PEGylation was performed with mPEG-SVA (Laysan) and PEG-biotin (Laysan), followed by a second round of PEGylation with $MS(PEG)_4$ (ThermoFisher). After assembly of a microfluidic chamber, the slides were incubated with 1 % Tween-20 for 15 minutes. Excess Tween-20 was washed away with 100 µL T50 (50 mM Tris-HCl, pH 8.0, 50 mM NaCl) followed by a 2 min incubation of 20 µL streptavidin (0.1 mg/mL, ThermoFisher). Excess streptavidin was removed with 100 µL T50. Next, for single-molecule experiments we immobilized 50 µL of 100 pM Cy5 labelled target DNA for 2 minutes, unbound DNA was washed with 100 µL T50, followed by 100 µL of origami-buffer (50 mM Tris-HCl, pH 8.0, 50 mM NaCl, 1 mM $MnCl_2$, 5 mM $MgCl_2$). The Ago-guide complex was formed by incubating 10 nM CbAgo with 1 nM of Cy3 labelled DNA guide for 20 minutes at 37 °C in the origami-buffer. For single-molecule experiments, we injected 50 µL of 1 nM Ago-guide complex or 50 µL of 1 nM DNA-PAINT imager strand in imaging buffer (50 mM Tris-HCl, pH 8.0, 50 mM NaCl, 1 mM $MnCl_2$, 5 mM $MgCl_2$, 0.8 % glucose, 0.5 mg/mL glucose oxidase (Sigma), 85 ug/mL catalase (Merck) and 1 mM Trolox (Sigma)). The single-molecule FRET experiments for Figure 1 were performed at room temperature (23 ± 2 °C). For super-resolution DNA origami experiments, we flushed 50 µL of ~200 pM DNA origami structures in a streptavidin coated channel and incubated for 3 minutes to allow for specific immobilization. Unbound DNA-origami was washed with origami--buffer. Next, 50 µL of 100 pM of Ago-guide complex or 1 nM DNA-PAINT imager strand was injected in imaging buffer.

6.5.4 Assembly of DNA oligo plate

The 2D rectangular DNA origami structure was designed by using CaDNAno software based on square lattice.[44] The 2D rectangular DNA origami structure was twist corrected and structural behaviour of the origami plate was checked by coarse-grained simulations in CanDo.[45,46] The parameters used for simulations are axial rise per base-pair = 0.34 nm, helix diameter = 2.25 nm, crossover spacing = 10.5 bp, axial stiffness = 1100 pN, bending stiffness = 230 pN nm^2, torsional stiffness = 460 pN nm2, nick stiffness factor = 0.01. The 2D rectangular DNA origami structure self-assembled in a total reaction volume of 100 µL containing 10 nM of p8064 scaffold strand (Tilibit nanosystems), 100 nM core staples (Integrated DNA Technologies), 100 nM Ago-PAINT handles and 100 nM biotin handles in 1x TE folding buffer (Tilibit nanosystems) supplemented with 11 mM $MgCl_2$. The origami structures were annealed using a thermocycler. First, the reaction mixture was heated for 10 minutes at 65 °C, then a temperature gradient was applied from 60 °C to 40 °C with a rate of 1 °C/hour. After self-assembly, the origami structures were purified using Amicon spin filter (100K MWCO) and stored in T50 buffer containing 11 mM $MgCl_2$. The purified DNA origami structures were analysed on a 2 % agarose

gel (Tris-borate-EDTA, 11 mM $MgCl_2$). The gel was run at 90 V for 2 hours in ice. After staining the gel with ethidium bromide, the samples were imaged to verify the quality of the folding procedure (**Figure S6.2**A). Next, the purified origami sample were checked for rectangular structure by atomic force microscopy (AFM) on mica surface according to AFM imaging procedures. Briefly, 0.01 % (w/v) polylysine was incubated 1 min on a freshly cleaved 3 mm (1/8 inch) diameter mica disk. The mica surface was gently washed with MQ water and blow dried with N2. Next, 5 µL of 500 pM DNA origami samples was incubated onto a mica disk for 5 minutes. The mica disk was washed gently with 1 mL (3x) of folding buffer with 11 mM $MgCl_2$ to remove any unbound DNA origami structures, then quickly rinsed with MQ water and blow dried with N_2. Dry AFM images were acquired in Bruker Multimode 8 AFM. Sharp AFM tips were used for AFM measurements (Bruker PeakForce HIRS-F-B) with 0.12 N/m nominal spring constant. AFM images were acquired in tapping mode. Example images of AFM images can be found in **Figure S6.2**B and C.

6.5.5 Super-resolution data analysis

CCD movies were acquired through custom-written program. The resulting files were converted to .raw file format using a custom-written script in Matlab (Mathworks). Super-resolution reconstruction, drift-correction and alignment were performed using the Picasso software package,[7] for both Ago-PAINT and DNA-PAINT.

6.6 Supporting Information

6.6.1 Supporting Figures

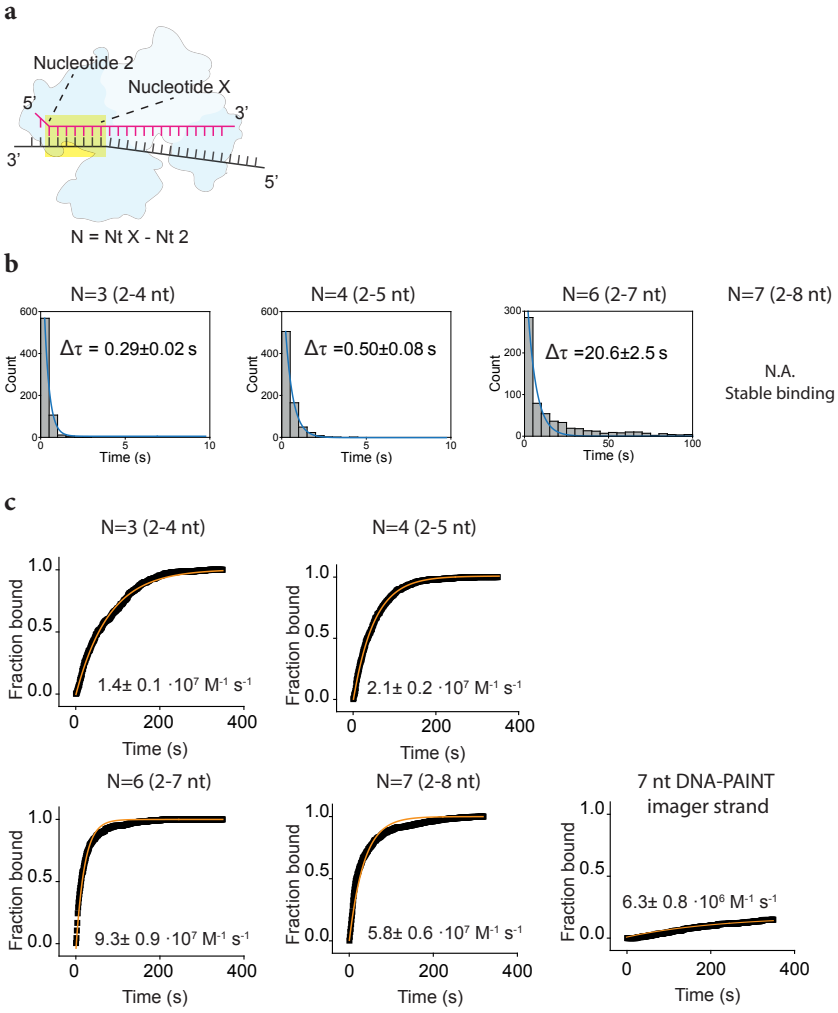


Figure S6.1: Single-molecule binding and unbinding kinetics.
(a) Top left: A cartoon figure indicating the base indexes and the definition of N (the number of base pairs). Since the first nucleotide of the imager strand is embedded in Ago, base pairing starts from the second nucleotide.**(b)** Dwell time histograms for DNA-PAINT for 2-4, 2-5, 2-7 and 2-8 nt base pairing. For 2-8 nt base-pairing, accurate measurements of dwell times were limited by photobleaching. **(c)** Fractional binding curve for CbAgo-siDNA for 2-4, 2-5, 2-7 and 2-8 nucleotide base pairing with the target sequence. Additionally, a fractional binding curve is shown for 7 nt base pairing with DNA-PAINT. Data was taken on two different days. A single-exponential fit was performed on the data (orange line). Error bars are given by the 95% confidence interval acquired from 105 bootstraps.

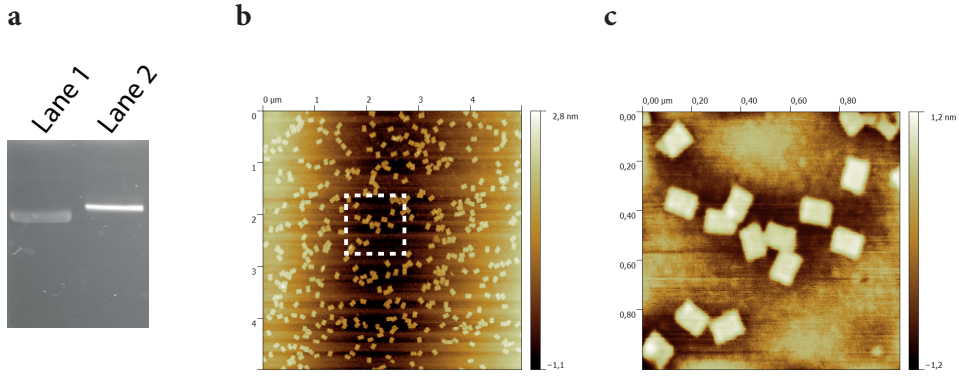
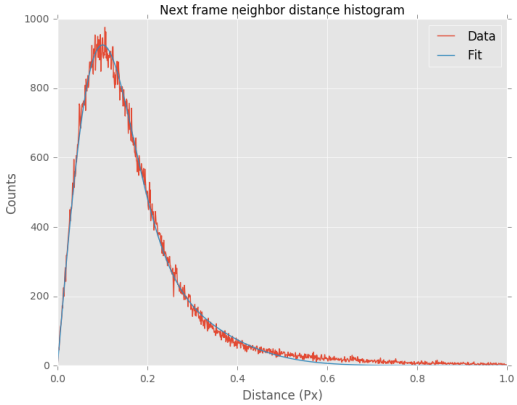


Figure S6.2: Quality control of origami plate assembly.
(a) Agarose gel image showing formation of the DNA origami structure. Lane 1: single-stranded M13mp18 p8064 scaffold. Lane 2: annealed origami mixture in 1x TE folding buffer. DNA origami and DNA scaffold were run in a 0.5x TBE + 11 mM $MgCl_2$ buffered 2 % agarose gel. **(b)** An AFM image of 500 pM DNA origami plates deposited on a polylysine treated mica disc. **(c)** A zoom-in from the white striped square region from (a) shows the rectangular form of the respective plates.

a

Ago-PAINT



b

DNA-PAINT

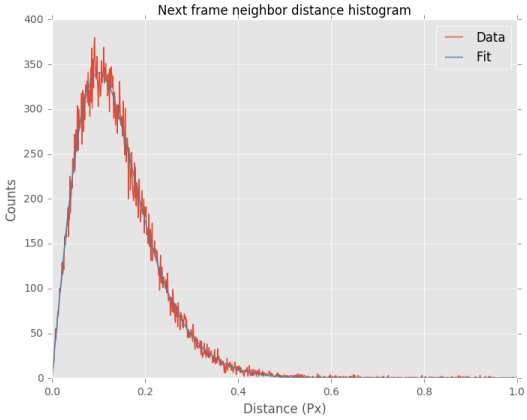


Figure S6.3: Comparison of localization precision between Ago-PAINT and DNA-PAINT through nearest neighbour analysis for a 30x thymine linker docking strand.

(a) A nearest neighbour in adjacent frame histogram of super-resolution binding sites made through Ago-PAINT. The pairwise displacement fit is given by the blue curve with a NeNA precision 0.09 pixel = 9.7 nm. **(b)** A nearest neighbour in adjacent frame histogram of super-resolution binding sites made through DNA-PAINT. The pairwise displacement fit is given by the blue curve with a NeNA precision 0.09 pixel = 9.7 nm.

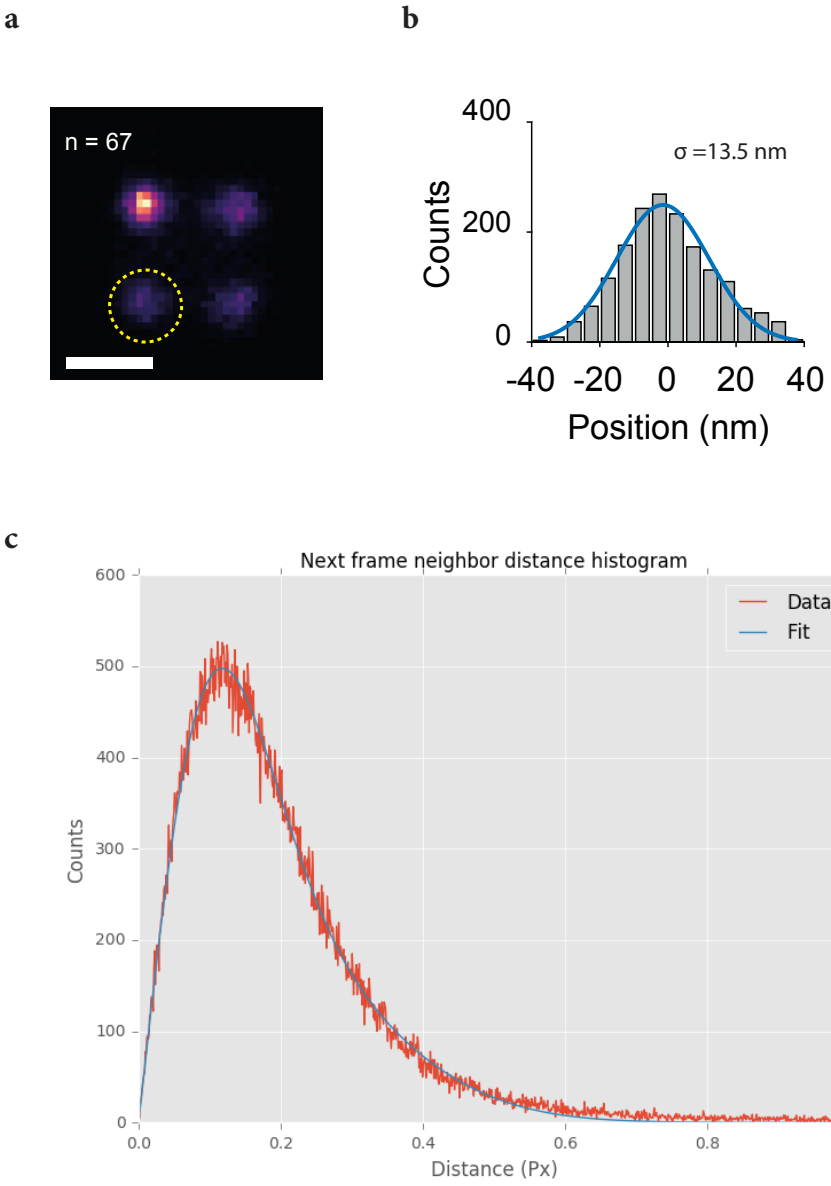


Figure S6.4: Localization precision of 50x thymine linker docking strand.
(a) A summed image of 67 origami structures made through the use of Ago-PAINT. Scale bar indicates 50 nm. **(b)** A cross-sectional histogram taken from the yellow encircled area in panel (a). The standard deviation or localization uncertainty is given by σ = 13.5 nm. **(c)** A nearest neighbour in adjacent frame histogram. The pairwise displacement fit is given by the blue curve with a NeNA precision of 0.1 pixel = 11.3 nm.

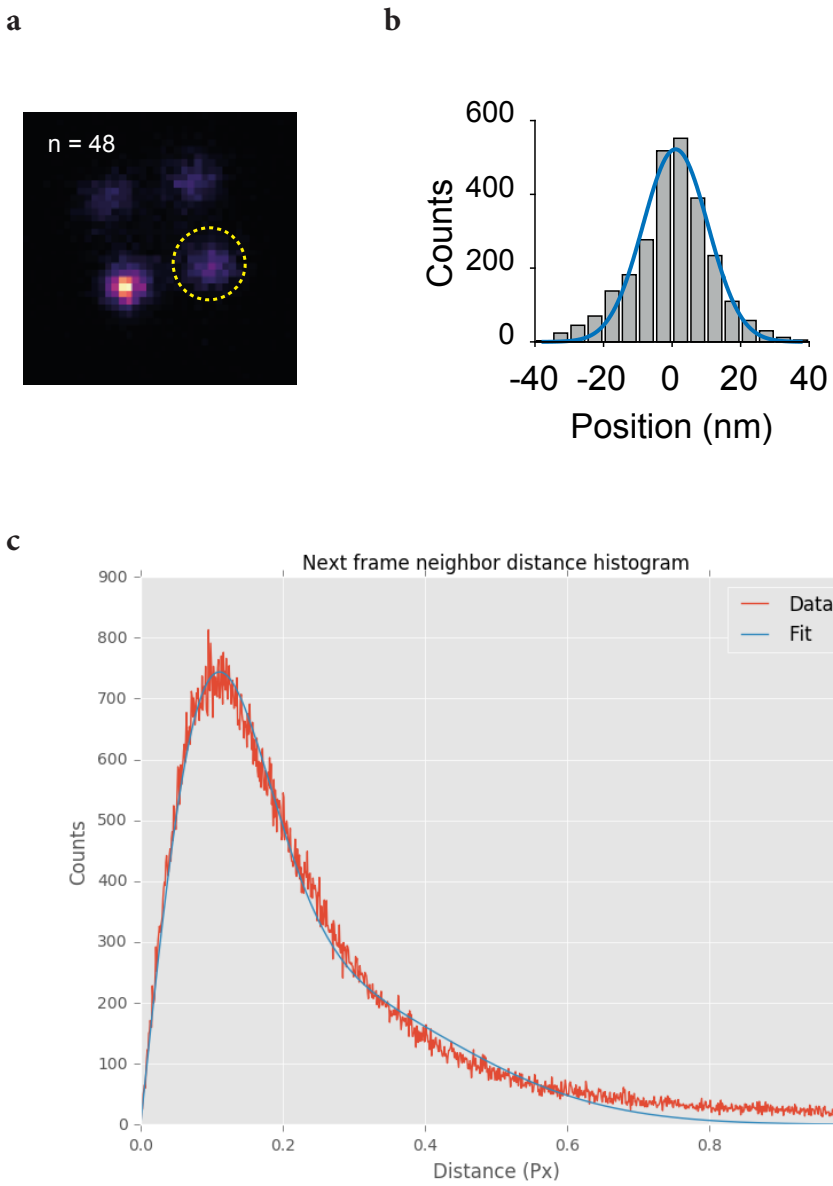


Figure S6.5: Localization precision of 100x thymine linker docking strand.
(a) A summed image of 48 origami structures made through the use of Ago-PAINT. Scale bar indicates 50 nm. **(b)** A cross-sectional histogram taken from the yellow encircled area in panel (a). The standard deviation or localization uncertainty is given by σ = 9.8 nm. **(c)** A nearest neighbour in adjacent frame histogram. The pairwise displacement fit is given by the blue curve with a NeNA precision of 0.1 pixel = 10.8 nm.

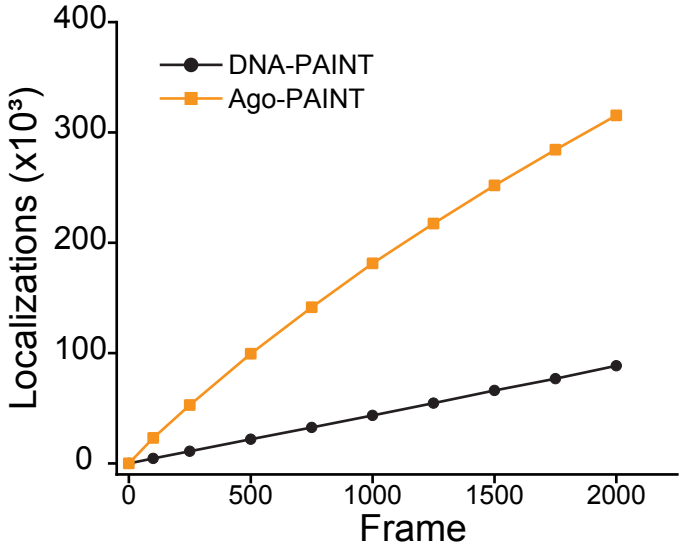


Figure S6.6: Number of Localizations versus the number of frames.

6.6.2 Supporting Tables

Table S6.1: Single-Molecule DNA constructs

DNA Strand	Nucleotide Sequence (5' → 3')	Modification
Single-Molecule DNA target strand	TTTTTATACATCTATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	Cy3-5' end labelled, biotin-3' end labelled
DNA PAINT imager strand (7nt)	GATGTAT	Cy3- 3' end labelled
DNA PAINT imager strand (8nt)	AGATGTAT	Cy3- 3' end labelled
Ago_PAINT 2-3	TAGTACTT**X**TTTTTTTTTTTT	X= Cy3-dT
Ago_PAINT 2-4	TAGACTTT**X**TTTTTTTTTTTT	X= Cy3-dT
Ago_PAINT 2-5	TAGATTTT**X**TTTTTTTTTTTT	X= Cy3-dT
Ago_PAINT 2-6	TAGATGAT**X**TTTTTTTTTTTT	X= Cy3-dT
Ago_PAINT 2-7	TAGATGTT**X**TTTTTTTTTTTT	X= Cy3-dT
Ago_PAINT 2-8	TAGATGTAA**X**TTTTTTTTTTT	X= Cy3-dT

Table S6.1: Sequences of origami extensions

DNA Strand	Nucleotide Sequence (5' → 3')
5xT_1	TTTTTATACATCTATTTTTTTGACCTTATTACCTTATGCGATTCGTTGGGAA
5xT_2	TTTTTATACATCTATTTTTTCCAGTACGCGGGGTTTTGCTCAGTAAGAGGCT
5xT_3	TTTTTATACATCTATTTTTTCGTAATCCCTGTCGTGCCAGCTGGGCGGTTTG
5xT_4	TTTTTATACATCTATTTTTTGGCGGTCTTACATTGGCAGATTCACCTACATT

DNA Strand	Nucleotide Sequence (5' → 3')
30xT_1	TTT TTA TAC ATC TAT TTT TTT TTT TTT TTT TTT TTT TTT TTT GAC CTT ATT ACC TTA TGC GAT TCG TTG GGA A
30xT_2	TTT TTA TAC ATC TAT TTT TTT TTT TTT TTT TTT TTT TTT TTC CAG TAC GCG GGG TTT TGC TCA GTA AGA GGC T
30xT_3	TTT TTA TAC ATC TAT TTT TTT TTT TTT TTT TTT TTT TTT TTC GTA ATC CCT GTC GTG CCA GCT GGG CGG TTT G
30xT_4	TTT TTA TAC ATC TAT TTT TTT TTT TTT TTT TTT TTT TTT TTG GCG GTC TTA CAT TGG CAG ATT CAC CTA CAT T
50xT_1	TTT TTA TAC ATC TA TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT T GAC CTT ATT ACC TTA TGC GAT TCG TTG GGA A
50xT_2	TTT TTA TAC ATC TA TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT C CAG TAC GCG GGG TTT TGC TCA GTA AGA GGC T
50xT_3	TTT TTA TAC ATC TA TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT C GTA ATC CCT GTC GTG CCA GCT GGG CGG TTT G
50xT_4	TTT TTA TAC ATC TA TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT G GCG GTC TTA CAT TGG CAG ATT CAC CTA CAT T
100xT_1	TTT TTA TAC ATC TA TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT

6.7 References

1 Betzig, E. et al. Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. Science (80-.). 313, 1642–1645 (2006).

2 Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat. Methods 3, 793–796 (2006).

3 Giannone, G. et al. Dynamic Superresolution Imaging of Endogenous Proteins on Living Cells at Ultra-High Density. Biophys. J. 99, 1303–1310 (2010).

4 Schoen, I., Ries, J., Klotzsch, E., Ewers, H. & Vogel, V. Binding-activated localization microscopy of DNA l. Nano Lett. 11, 4008–4011 (2011).

5 Sharonov, A. & Hochstrasser, R. M. Wide-field subdiffraction imaging by accumulated binding of diffusing probes. Proc. Natl. Acad. Sci. 103, 18911–18916 (2006).

6 Jungmann, R. et al. Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. Nat. Methods 11, 313–318 (2014).

7 Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with DNA-PAINT. Nat. Protoc. 12, 1198–1228 (2017).

8 Dai, M., Jungmann, R. & Yin, P. Optical imaging of individual biomolecules in densely packed clusters. Nat. Nanotechnol. 11, 798–807 (2016).

9 Jungmann, R. et al. Single-Molecule Kinetics and Super-Resolution Microscopy by Fluorescence Imaging of Transient Binding on DNA Origami. Nano Lett. 10, 4756–4761 (2010).

10 Auer, A., Strauss, M. T., Schlichthaerle, T. & Jungmann, R. Fast, Background-Free DNA-PAINT Imaging Using FRET-Based Probes. Nano Lett. 17, 6428–6434 (2017).

11 Lee, J., Park, S., Kang, W. & Hohng, S. Accelerated super-resolution imaging with FRET-PAINT. Mol. Brain 10, 63 (2017).

12 Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. Cell 136, 215–233 (2009).

13 Baek, D. et al. The impact of microRNAs on protein output. Nature 455, 64–71 (2008).

14 Selbach, M. et al. Widespread changes in protein synthesis induced by microRNAs. Nature 455, 58 (2008).

15 Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466, 835–840 (2010).

16 Wang, Y. et al. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. Nature 456, 921–926 (2008).

17 Schirle, N. T. & MacRae, I. J. The Crystal Structure of Human Argonaute2. Science (80-.). 336, 1037–1040 (2012).

18 Chandradoss, S. D., Schirle, N. T., Szczepaniak, M., Macrae, I. J. & Joo, C. A Dynamic Search Process Underlies MicroRNA Targeting. Cell 162, 96–107 (2015).

19 Salomon, W. E. et al. Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. Cell 162, 84–95 (2015).

20 Jo, M. H. et al. Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs. Mol. Cell 59, 117–124 (2015).

21 Yao, C., Sasaki, H. M., Ueda, T., Tomari, Y. & Tadakuma, H. Single-Molecule Analysis of the Target Cleavage Reaction by the Drosophila RNAi Enzyme Complex. Mol. Cell 59, 125–132 (2015).

22 Swarts, D. C. et al. The evolutionary journey of Argonaute proteins. Nat. Struct. Mol. Biol. 21, 743–53 (2014).

23 Hegge, J. W., Swarts, D. C. & Van Der Oost, J. Prokaryotic argonaute proteins: Novel genome--editing tools? Nat. Rev. Microbiol. 16, 5–11 (2018).

24 Swarts, D. C. et al. DNA-guided DNA interference by a prokaryotic Argonaute. Nature 507, 258–61 (2014).

25 Hegge, J. W. et al. DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute. Nucleic Acids Res. 47, 5809–5821 (2019).

26 Cui, T. J. et al. Argonaute bypasses cellular obstacles without hindrance during target search. Nat. Commun. 10, 4390 (2019).

27 Hegge, J. W. et al. DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute. Nucleic Acids Res. 47, 5809–5821 (2019).

28 Endesfelder, U., Malkusch, S., Fricke, F. & Heilemann, M. A simple method to estimate the average localization precision of a single-molecule localization microscopy experiment. Histochem. Cell Biol. 141, 629–638 (2014).

29 Wade, O. K. et al. 124-Color Super-resolution Imaging by Engineering DNA-PAINT Blinking Kinetics. Nano Lett. 19, 2641–2646 (2019).

30 Deußner-Helfmann, N. S. et al. Correlative Single-Molecule FRET and DNA-PAINT Imaging. Nano Lett. 18, 4626–4630 (2018).

31 Shah, S., Dubey, A. K. & Reif, J. Programming Temporal DNA Barcodes for Single-Molecule Fingerprinting. Nano Lett. 19, 2668–2673 (2019).

32 Schueder, F. et al. An order of magnitude faster DNA-PAINT imaging by optimized sequence design and buffer conditions. Nat. Methods 16, 1101–1104 (2019).

6

33 Wittrup, A. & Lieberman, J. Knocking down disease: a progress report on siRNA therapeutics. Nat. Rev. Genet. 16, 543–552 (2015).

34 Setten, R. L., Rossi, J. J. & Han, S. The current state and future directions of RNAi-based therapeutics. Nat. Rev. Drug Discov. 18, 421–446 (2019).

35 Dayeh, D. M., Cantara, W. A., Kitzrow, J. P., Musier-Forsyth, K. & Nakanishi, K. Argonaute-based programmable RNase as a tool for cleavage of highly-structured RNA. Nucleic Acids Res. 46, 1–13 (2018).

36 Tyagi, S. Imaging intracellular RNA distribution and dynamics in living cells. Nat. Methods 6, 331–338 (2009).

37 Kaya, E. et al. A bacterial Argonaute with noncanonical guide RNA specificity. Proc. Natl. Acad. Sci. 113, 4057–4062 (2016).

38 Lapinaite, A., Doudna, J. A. & Cate, J. H. D. D. Programmable RNA recognition using a CRISPR-associated Argonaute. Proc. Natl. Acad. Sci. 115, 3368–3373 (2018).

39 Chang, L. et al. AgoFISH: cost-effective in situ labelling of genomic loci based on DNA-guided dTtAgo protein. Nanoscale Horizons 4, 918–923 (2019).

40 Spahn, C. et al. Protein-Specific, Multicolor and 3D STED Imaging in Cells with DNA-Labeled Antibodies. Angew. Chemie Int. Ed. 58, 18835–18838 (2019).

41 Jungmann, R. et al. Quantitative super-resolution imaging with qPAINT. Nat. Methods 13, 439–442 (2016).

42 Liu, N., Dai, M., Saka, S. K. & Yin, P. Super-resolution labelling with Action-PAINT. Nat. Chem. 11, 1001–1008 (2019).

7

Towards Single-Molecule Protein Sequencing: Outlook and Concluding Remarks

Mike Filius

7.1 Proteins are the workhorses of the cell

Proteins are the end products of life's decoding process that starts with the information in the genetic code within the cell. The primary structure of a protein – its amino acid sequence – drives the folding and the intramolecular interactions of polypeptide chain, and results in a unique 3D structure. Unlike DNA – that has only 4 different bases – the primary protein sequence is built from 20 different amino acids, making protein analysis and identification a complicated task. Furthermore, while the number of protein encoding genes is estimated to be around 20 000, the proteome is estimated to contain over a million different proteins. For example, due to processes such as alternative splicing, the use of different promotors or translational start sites, single amino acid polymorphisms, and post translational modifications (PTMs), it is estimated that there are >100 different protein isoforms derived from each protein encoding gene (**Figure 7.1**).[1] Each individual molecular form that is expressed from a protein encoding gene is referred to as a proteoform.[2]

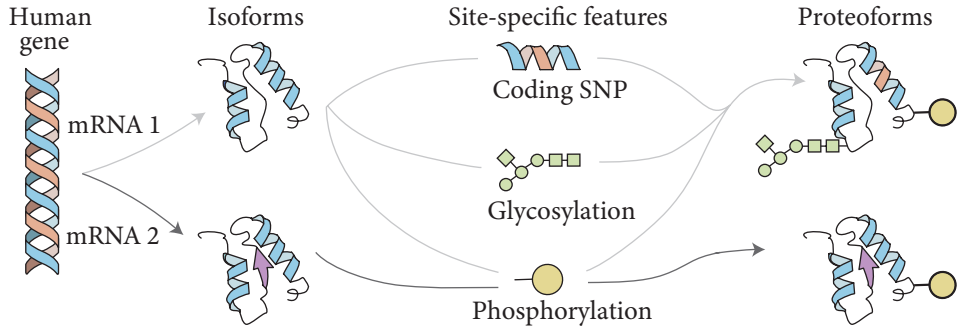


Figure 7.1: Illustration depicting various sources for proteoform diversity.
A single human gene can produce different mRNA molecules. These different mRNA molecules give rise to different protein primary sequences. The increase in different mRNA transcripts commonly arise from alternative splicing of RNA and from the use of different promotors or translational start sites. Protein isoform variation is complemented by site-specific changes to generate human proteoforms (at right); three examples of site-specific changes include single-nucleotide polymorphisms (SNPs) and co- or post-translational modifications like N-glycosylation or phosphorylation, respectively. Figure adopted from Aebersold *et al.*[1]

Since much regulation occurs beyond the genome and transcriptome, we must decipher the proteome to fully understand human health and disease at the molecular level. To achieve this ambitious goal, there is a demand for expanding the technological capabilities to study the proteome at the individual proteoform level. The discovery and identification of novel proteoforms can be divided into two main challenges. First there are many sources of variation that collectively cause a large pool of RNA isoforms from a protein encoding gene (**Figure 7.1**), translation of these isoforms would result in alterations in the amino acid sequence of proteins, for this a technology is required that is specialized in analysis of the amino acid sequence. Secondly the most daunting task will be the analysis of PTM profiles of each proteoform.

Single-molecule technologies are a valuable addition to the arsenal of proteomics tools as they provide ultimate sensitivity for the detection of low abundant proteoforms (see **Chapter 1**). Especially in the absence of a "protein polymerase" enzyme that can amplify protein molecules, ultimate sensitivity is required. Furthermore, single-molecule proteoform analysis may shed light on proteoform heterogeneity below the population level.

The work described in this thesis mainly focusses on the exploration of fluorescence-based approaches for single-molecule protein identification and analysis. While many different approaches for single-molecule protein sequencing are being explored (see overview in **Chapter 1**), each of these techniques has its unique benefits and drawbacks. For example, to fully understand and identify all proteoforms, each protein has to be sequenced at both the amino acid and PTM level. Furthermore, the dynamic range of protein copy numbers within the cell require an approach that can analyze millions of protein molecules in a short period of time, although single-molecule approaches have the sensitivity to analyze each protein molecule within a sample, current methods lack the throughput for complex samples, such as single-cell. In this chapter, I discuss some of the most common challenges and contemplate about future directions for the field of single-molecule proteins sequencing.

7.2 Amino acid analysis of proteoforms

The first layer of proteoform diversity arises from the many different RNA molecules that are produced by a single protein encoding gene. For example, alternative splicing events or the occurrence of alternative transcriptional start sites greatly increase the number of RNA molecules that can be transcribed from protein encoding genes. The translation of these differently transcribed RNA molecules will result in alterations in the primary sequence of the protein. Thus, to distinguish these group of proteoforms, the amino acid sequence must be analyzed.

Fluorescence approaches have been one of the main drivers in the development for next generation DNA sequencing devices. For example, most next generation sequencing devices employ the incorporation of fluorescently labeled nucleotides during DNA replication to determine the sequence. The development of a fluorescence based single-molecule protein/proteoform sequencing device comes with great challenges, including the lack of chemistry for all 20 different amino acids, as well as the inability to discriminate all 20 amino acids with fluorescent probes.

To simplify the task for protein identification using single-molecule fluorescence, our group and others have introduced the concept of protein fingerprinting.[3–6] Protein fingerprinting typically relies on the determination of the number and position of a small number of amino acid residues within the full length amino acid sequence. Then, by comparing the fingerprint to a reference database, one is able to identify >90 % of the human proteome.[3–6]

Despite the dream of full proteome analysis, single-molecule protein fingerprinting approaches are likely to find their first applications in the realm of targeted proteomics. For biomedical applications, the highly sensitive detection of disease biomarkers will facilitate early diagnosis of the disease, as well as inform on its progression. In **Chapter 4** we have shown that our fingerprinting approach was able to identify

the different splicing forms of Bcl-2 proteins. Alternative splicing of the apoptosis regulator Bcl-2 produces either in Bcl-XL (anti-apoptotic regulator) or Bcl-Xs and Bcl-Xb (both pro-apoptotic regulators).[7,8] The ratio between these regulators is of utmost importance for cell fate and its dysregulation may result in cancer invasion and metastasis. Assays for targeted protein identification can be further explored for other biomarkers. Besides Bcl-2, many other genes undergo alternative splicing and likewise the occurrence of these splicing events can have a major effect on health, suggesting that the proteoforms may serve as biomarkers.[9–11]

Notwithstanding the example of fingerprinting for spliceoform analysis as described above, not all biomarkers can be analyzed and identified by measuring a subset of amino acids. Furthermore, there are more site-specific modifications, such as PTMs (discussed in the next section) or coding single-nucleotide polymorphisms (**Figure 7.1**). The latter would result in single amino acid substitutions that can have impact on health and disease. The detection of these single substitutions requires a technology that goes beyond fingerprinting schemes and is capable of analysing every amino acid within the primary structure of proteins.

Nanopore based research is moving fast in the direction of protein analysis. With low cost and ease of use, nanopore sequencing holds great promise for single-molecule protein sequencing and could be commercialized within the next decade. Recently, nanopore based protein analysis has reached the level of which a current signal could be obtained for all 20 different amino acids.[12] Furthermore, by attaching a short DNA linker to a peptide, and controlling the translocation with a helicase, individual peptides can be measured multiple times resulting in extremely low error rates.[13] However, despite the tremendous advancement in the detection of single amino acids within a model peptide, the question still remain how nanopore based devices will cope with the complexity of full length proteins. As mentioned above, the proteoform dictionary goes beyond the 20 amino acid code due to the many PTMs. Furthermore, heterogeneous charge of the polypeptide chain, complex 3D structure and translocation speed are among the most challenging hurdles to move beyond the simplified peptide model systems to the analysis of full-length proteins.

7.3 Deciphering proteoforms at the PTM level

In the previous section, we have discussed the ability and potential use of using single-molecule fingerprinting schemes for the detection of alternative splicing and the biomarkers produced thereof. However, what makes proteoform analysis even more challenging are the PTMs. PTMs, including phosphorylation, glycosylation and many other modifications, expand the proteoform diversity exponentially. For example, recent data have shown that ~ 60% of all proteins are glycosylated.[14] It comes as no surprise that dysregulation of PTMs has been implicated in a myriad of diseases (**Figure 7.2**).

Single-molecule proteoforms analysis is a promising approach to identify and quantify the number of PTMs on individual proteins. This has been proposed and validated using model peptides with different PTMs in a biological nanopore. In a proof-of-concept study, Restrepo-Perez *et al*. were able to differentiate between phosp-

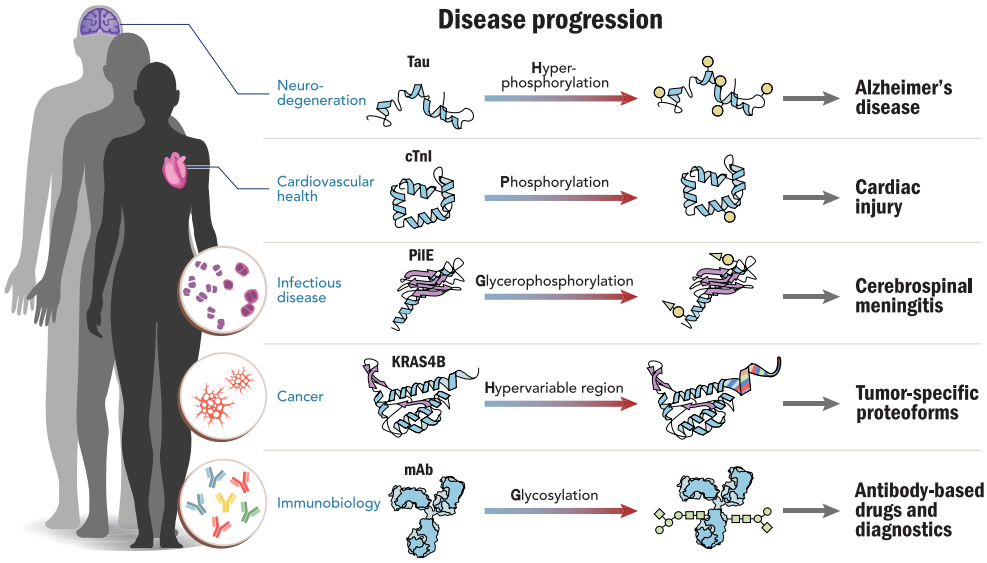


Figure 7.2: Proteoforms in human disease. Five important clinical areas are depicted and serve as examples where proteoforms have been identified and linked to the progression of human disease. Figure adapted from Smith *et al.*[15]

horylated and O-glycosylated peptides at the single-molecule level.[16] The detection of the PTMs did not require labeling of the PTMs themselves, thereby keeping the sample preparation relatively simple. However, a simplification of the assay was that the peptide backbone was specifically designed to promote directional translocation of the peptide through the nanopore. Furthermore, the label free analysis of PTMs in human proteoforms will be challenging due to the signal that arises from multiple amino acids residues that reside within the nanopore, as well as the many different other PTMs that might be present on a particular peptide or protein molecule. Therefore, early applications for single-molecule PTM profiling may involve the enrichment of proteins containing a particular PTM and the enhancement of the PTM signal in the assay.

Labeling or derivatization of PTMs is a common technique that was used in the early days of PTM profiling using MS.[17] However, the analysis of protein phosphorylation using MS was difficult due to the labile phosphorylation modification that often got lost during the ionization step in the MS procedure. One of most common phosphoprotein enrichment and derivatization approaches is the combination of beta-elimination and Michael addition (BEMA).[18] This allows for site-specific modification of phosphorylated residues. Recently, the BEMA reaction has been adapted and modified for the attachment of fluorescent probes[19] or short DNA oligos[20] for single-molecule profiling of protein phosphorylation. The attachment of short DNA oligos to phosphorylation sites can be easily implemented in our FRET X approach. In **Chapter 5** we have shown that we can determine the relative position of multiple cysteines to a reference point in full-length proteins. We envision that our assay can

be further developed to enable profiling and site mapping of phosphorylation on individual full-length proteins.

An alternative approach for the labeling of PTMs to enhance their signal for different single-molecule technologies can be the use of metabolic labeling strategies. For example, with metabolic glycan labeling unnatural sugars are attached to glycoproteins within the cell. These unnatural sugars are often modified to contain a chemical tag (azide, alkyne, etc.) that can later be used for the attachment of fluorescent probes or short DNA tags via copper free click-chemistry between azide and dibenzocyclooctyne (DBCO). This metabolic labeling can provide an interesting intermediate step towards single-molecule analysis particular PTMs within entire cells.

7.4 Affinity-based approaches for single-molecule proteoform analysis.

Affinity-based approaches have been extremely successful in detecting and identifying proteins without the need for covalent attachment of chemical probes to proteins. The well-established ELISA (enzyme-linked immune sorbent assay) has revolutionized biomarker detection with its ability to detect single proteins in patient blood samples.[21] Several affinity based single-molecule sequencing approaches have been proposed.[22] For example, one scheme aims to measure the amino acid at the N terminus of a protein. After its recognition, the N-terminal amino acid is cleaved using an Edman degradation cycle and the subsequent amino acid is identified.[23,24] This approach requires the development of twenty affinity based probes, each providing high selectivity and affinity to any of the amino acids. Although this approach holds great promise for the identification of small peptides, the measurement time rapidly increases due to multitude of washing and cleaning steps. Furthermore, this approach is hampered by the fact that cleavage of the of the N terminal amino acid is not always successful, which leads to errors in the sequence determination.

Therefore our group and others[25] are exploring a different use of affinity-based probes for single-molecule protein identification or proteoform analysis. In this approach, each affinity probe provides information about the target proteoform as a whole, and not just its N-terminal amino acid.

7.5 Towards single cell proteomics.

Despite tremendous effort and technological advances in the single-molecule protein analysis methods, most of the described technologies are still in their 'proof--of-principle' phase and are demonstrated on simple model systems such as synthetic peptides or recombinant proteins. For single-molecule technologies to become the new standard for protein identification, we have to start thinking about ways to deal with the complexity of cells.

One of the main hurdles for cell analysis is the large dynamic range of protein copy numbers within the cell. Protein copy number have been reported to be as low

as 10-100 protein copies per cell, and can go up to 10^{11} protein copies for a single human cell (**Figure 7.3**).[26] The detection of the low abundant proteins is difficult for two main reasons; (1) the intrinsic small quantity of the sample, (2) for conventional technologies the bottleneck is the limited dynamic range. Together, these factors cause the dark corner of the proteome to get masked by the most abundant proteins and go undetected.[27] For example, the most abundant protein within cells is albumin and is present at 30-50mg/mL in normal conditions, whereas the lowest abundance protein (such as cytokines) are only present in low pg/mL concentration. Additionally, the 20

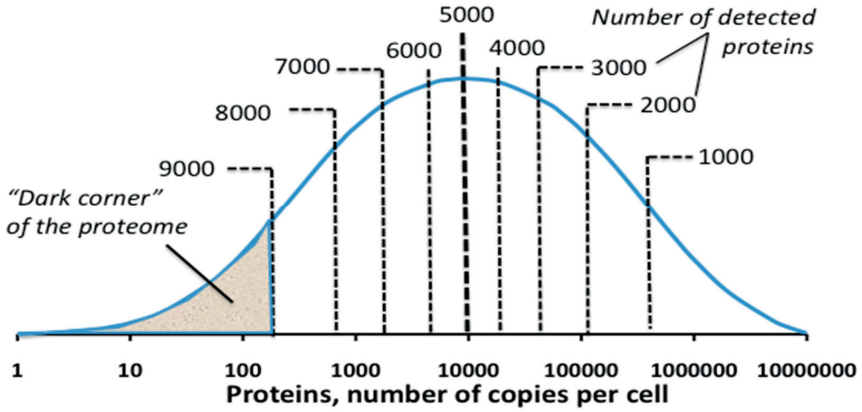


Figure 7.3: The dark corner of the proteome. Distribution of protein abundances is a bell-shape curve on a logarithmic copy number scale. Conventional proteomics analysis detects highly abundant proteins, and stretches for about four orders of magnitude. Deeper proteome analysis requires much larger sample size. The "dark corner" rep- esents the most challenging for detection part of the proteome, approximately 1000 least abundant proteins. Figure adopted from Zubarev.[27]

most abundant proteins within cells account for 97 % of the total plasma proteins.[28]

Single-molecule approaches have the required sensitivity to analyze small samples thereby mitigating the low sample issue. However, to also overcome the masking issue, a full proteome study would require measuring millions of protein molecules from a cell.

Current proteomics focus on the analysis of the low-copy-number proteins. These low-copy-number proteins (fewer than 1,000 molecules per cell) are involved in crucial functions such as gene expression, cellular metabolism and cell signalling. As such, the expression level of low-copy-number proteins of individual cells provides key information for the in-depth understanding of biological processes and diseases. However, analysis of these low-copy-number proteins requires targeted sample preparation via the removal of the most abundant proteins. Currently, within the MS community two main approaches are used to do so, from which we can learn: immunodepletion and fractionation by chromatography.[29]

In principle, both approaches could be applied to remove the most abundant proteins. However, immunodepletion is preferred because it is compatible with full length, intact proteins and does not require the use of organic solvents or denaturating conditions. Additionally, since single-cell analysis at the genomic and transcriptomic

level has become increasingly popular, the number of published protocols is growing rapidly. The single-molecule proteomics community can learn from many of the progress made by the single-cell genetics field and adopt and modify their sequencing workflow. Many of the single-cell genetics/transcriptomics workflows start by cell sorting using FACS, followed by a sample amplification and library preparation step. Amplification with of the DNA with polymerases is crucial, as it allows for the analysis of low quantity DNA samples that are obtained from single-cells.

Unfortunately, single-molecule protein sequencing technologies cannot exploit an amplification step, since there is no enzyme that can perform a polymerase chain reaction-like reaction on protein substrates. The lack of a protein amplification method makes the analysis of low quantity proteins (e.g. from a single cell) challenging. However, despite the lack of such an approach, several groups have developed methods that can deal with small protein samples that are extracted from cells. For example, a solid-phase capture-release protocol has been described in which peptides are covalently attached to a resin.[30] This facilitates purification of the protein/peptides from any other cellular material with high yield and purity. The authors have demonstrated single-molecule identification of 40 000 peptides that were obtained from the lysate of a HEK293T cell. Additionally, the covalent attachment of peptides to a resin allows chemical derivatization of peptides (and perhaps later full proteins). The latter is especially interesting for many of the proposed single-molecule protein sequencing techniques as most of them require the modification of the protein by fluorophores, DNA oligos, or short peptides to ensure directional translocation through a nanopore (see **Chapter 1**).

Alternatively, affinity-based approaches are being explored for targeted proteomics at the single-molecule level. Recently, a single-molecule capture device has been designed for the detection of several biomarkers from blood samples.[21] The authors demonstrated that by using the appropriate antibodies they could detect a well-established prostate cancer biomarker, PSA, at the single-molecule level. Moreover, Liu et al. demonstrated the capability to probe low-copy-number proteins in single living cells by designing a single-cell plasmonic immuno sandwich assay.[31] By using a particular affinity ligand, the method has been extended to for the diagnostic detection of biomarkers.

The field of single-molecule proteoform analysis is still in the embryonic phase, where many scientists are still proposing, designing and validating new ideas. It is no surprise that in contrast to the increasingly large number of publications on protein sensing and identification, there is only a handful attempts for single-cell proteomics.[32] This mostly has to do with the relatively low throughput that most single-molecule approaches have. Therefore, I envision that single-molecule proteins sequencing or proteoform profiling technologies will find their first applications in targeted approaches and will slowly move towards more holistic assays and complex systems, such as cells.

7.6 Industry opportunities for single-molecule protein sequencing

The human genome project was a tremendous success and paved the way for unravelling the blue print of life. It remains the world's largest collaborative biological projects to date and stands as a signature scientific achievement. The project was an incredible success transforming and accelerating biological and medical research while converting a ~ $4 billion public investment into a major industry nowadays worth over $700 billion US dollar.[33]

The total cost of sequencing the first draft of the entire human genome is estimated to be several hundreds of millions of dollars (**Figure 7.4**). However, the help of industry allowed for the rapid development and advancement of the sequencing technology, drastically bringing down the costs. Illumina introduced its next generation sequencing platform in 2007, which was a game changer and sparked an enormous reduction in sequencing costs. Furthermore, companies such as Oxford Nanopore and Pacific Biosciences (PacBio) have introduced devices that can sequence DNA at the single-molecule level. The single-molecule sequencing devices have further

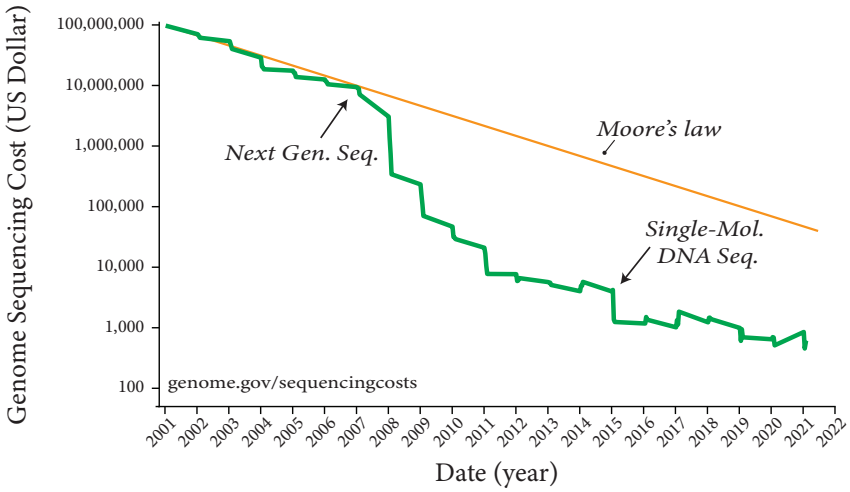


Figure 7.4: The cost for human genome sequencing. To illustrate the nature of the reductions in DNA sequencing costs, the graph also shows data reflecting Moore's Law (orange line), which describes a long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years. Technology improvements that 'keep up' with Moore's Law are widely regarded to be doing exceedingly well, making it useful for comparison. In this graph, note: (1) the use a logarithmic scale on the Y axis; and (2) the sudden and profound out-pacing of Moore's Law in the fall of 2007 (green line). The latter represents the time when the sequencing centers transitioned from Sanger-based sequencing to next generation (or second generation) DNA sequencing technologies. Furthermore, the next noticeable reduction in sequencing costs occurred in 2015 when single-molecule DNA sequencing platforms such as Pac Bio and Oxoford Nanopore's MinIon become commercially available. Figure adopted from National Human Genome Research Institute.

reduced the costs of genome sequencing. Oxford Nanopore's MinIon is a portable USB sequencing device that can be used anywhere without the need for a high-tech laboratory.

It is the dream of many scientists that proteomics will undergo a similar trend in cost reduction as DNA sequencing. Equally important is to make protein sequencing accessible to all by improving the ease of use. For this, the consortium for Top-Down Proteomics, led by the pioneer Prof. Neil Kelleher from Northwestern University launched the Human Proteoform Project.[15] The main objective of this ambitious project is defining the human proteome, that is, to generate a definitive reference of the proteoforms produced from the genome. Based on the huge economic output of the human genome project, several companies are hoping for a similar output with the human proteome project

The landscape of emerging proteomics technologies is and exciting and upcoming field, resulting in many excellent papers describing different approaches (see overviews in refs[34,35]). Additionally, several start-up companies have secured industry investments of several hundred million US dollars to develop single-molecule protein sequencing.[36,37] Earlier, industry investment has been shown to greatly increase technological development and cost reduction for genome sequencing and a similar trend will be required for proteoform analysis for the human proteoform project to be successful.

7.7 Concluding remarks

The field of single-molecule protein sequencing has seen tremendous progress and rsulted several dozens of high impact papers (see these reviews for an overview of the methods[34,35]). The realization of a single-molecule protein sequencer is technically challenging. However, when realized, it would revolutionize proteomic research by allowing for the analysis of tiny samples such as single cells. Furthermore, single-molecule sequencing may lead to the development of compact devices (as demonstrated by the advancement in DNA sequencing) which allow increase the ease of use and open up the possibility for on-site single-molecule protein sequencing.

The short-term goals for many of the emerging technologies will be optimizing the sensitivity, proteome coverage (fraction of whole proteins in the sample covered), sequence coverage (average fraction of protein sequences covered), read length (mean number of amino acids in a single read), accuracy (error in calling an amino acid or in identifying a whole protein), cost and throughput. Furthermore, the field of single-molecule protein sequencing would benefit from the development of multiple complementary technologies, each tailored to a specific niche of proteoform analysis. For example, in this chapter I have discussed the need for proteoforms analysis at the amino acid level and a specific method tailored to this challenge may be developed. Other technologies will instead be specialized in deciphering proteoform complexity at the PTM level.

The human genome project has demonstrated that technological development can be boosted by bringing together the academic community with the industry. With the start of the Human Proteoform project, I expect that attention from the industry will be drawn for the development of single-molecule protein sequencing devices. By combining forces, the road to realization of protein sequencing and proteoform analysis at the single-molecule level will be shortened and the dream of understanding the blue prints of life at the genome, transcriptome and proteome level will finally become reality!

7.8 References

1 Aebersold, R. et al. How many human proteoforms are there? Nat. Chem. Biol. 14, 206–214 (2018).

2 Smith, L. M. et al. Proteoform: a single term describing protein complexity. Nat. Methods 10, 186–187 (2013).

3 Lannoy, C. V. de, Filius, M., Wee, R. van, Joo, C. & Ridder, D. de. Evaluation of FRET X for single-molecule protein fingerprinting. iScience 24, 103239 (2021).

4 Yao, Y., Docter, M., Van Ginkel, J., De Ridder, D. & Joo, C. Single-molecule protein sequencing through fingerprinting: Computational assessment. Phys. Biol. 12, 10–16 (2015).

5 Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A Theoretical Justification for Single Molecule Peptide Sequencing. PLoS Comput. Biol. 11, 1–17 (2015).

6 Ohayon, S., Girsault, A., Nasser, M., Shen-Orr, S. & Meller, A. Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. PLoS Comput. Biol. 15, 1–21 (2019).

7 Kale, J., Osterlund, E. J. & Andrews, D. W. BCL-2 family proteins : changing partners in the dance towards death. Cell Death Differ. 25, 65–80 (2017).

8 Shiraiwa, N. et al. An additional form of rat Bcl-x, Bcl-xβ, generated by an unspliced RNA, promotes apoptosis in promyeloid cells. J. Biol. Chem. 271, 13258–13265 (1996).

9 Brinkman, B. M. N. Splice variants as cancer biomarkers. 37, 584–594 (2004).

10 Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. Nat. Publ. Gr. 18, 437–451 (2017).

11 Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. Nat. Publ. Gr. 17, 19–32 (2015).

12 Ouldali, H. et al. Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. Nat. Biotechnol. 38, 176–181 (2020).

13 Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A. & Dekker, C. Multiple rereads of single proteins at single–amino acid resolution using nanopores. Science. (2021)

14 Apweiler, R., Hermjakob, H. & Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database.

15 Smith, L. M. et al. The Human Proteoform Project: Defining the human proteome. Sci. Adv. 7, 734 (2021).

16 Restrepo-Pérez, L., Wong, C. H., Maglia, G., Dekker, C. & Joo, C. Label-free detection of post--translational modifications with a nanopore. Nano Lett. 19, 7957–7964 (2019).

17 Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. Nat. Biotechnol. 2003 213 21, 255–261 (2003).

18 Oda, Y., Nagasu, T. & Chait, B. T. Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. Nat. Biotechnol. 19, 379–382 (2001).

19 Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. Nat. Biotechnol. 36, 1076–1082 (2018).

20 Shrestha, P. et al. Single-molecule mechanical fingerprinting with DNA nanoswitch calipers. Nat. Nanotechnol. 2021 1–9 (2021).

21 Mao, C. P. et al. Protein detection in blood with single-molecule imaging. Sci. Adv. 7, (2021).

22 Tullman, J., Marino, J. P. & Kelman, Z. Leveraging nature's biomolecular designs in next-generation protein sequencing reagent development. Appl. Microbiol. Biotechnol. 1–11 (2020).

23 Rodriques, S. G., Marblestone, A. H. & Boyden, E. S. A theoretical analysis of single molecule protein sequencing via weak binding spectra. PLoS One 14, e0212868 (2019).

24 Tullman, J., Callahan, N., Ellington, B., Kelman, Z. & Marino, J. P. Engineering ClpS for selective and enhanced N-terminal amino acid binding. Appl. Microbiol. Biotechnol. 103, 2621–2633 (2019).

25 Egertson, J. D. et al. A theoretical framework for proteome-scale single-molecule protein identification using multi-affinity protein binding reagents. bioRxiv 2021.10.11.463967 (2021) doi:10.1101/2021.10.11.463967.

26 Ponomarenko, E. A. et al. The Size of the Human Proteome: The Width and Depth. Int. J. Anal. Chem. 2016, (2016).

27 Zubarev, R. a. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 13, 723–726 (2013).

28 Anderson, N. L. & Anderson, N. G. The Human Plasma Proteome History, Character, and Diagnostic Prospects. Mol. Cell. Proteomics 1, 845–867 (2002).

29 Roche, S. et al. Depletion of one, six, twelve or twenty major blood proteins before proteomic analysis: The more the better? (2009).

30 Howard, C. J. et al. Solid-Phase Peptide Capture and Release for Bulk and Single-Molecule Proteomics. bioRxiv (2020).

31 Liu, J., He, H., Xie, D., Wen, Y. & Liu, Z. Probing low-copy-number proteins in single living cells using single-cell plasmonic immunosandwich assays. Nat. Protoc. 16, 3522–3546 (2021).

32 Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. Nat Rev Genet advance on, (2017).

33 Drake, N. What is the human genome worth? Nature (2011).

34 Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. Nat. Nanotechnol. 13, 786–796 (2018).

35 Alfaro, J. A. et al. The emerging landscape of single-molecule protein sequencing technologies. Nat. Methods 18, 604–617 (2021).

36 Seattle Times. Seattle biotech startup Nautilus to get $350 million, stock listing in blank-check deal. The Seattle Times (2021).

37 Reuters. Protein sequencing firm Quantum-Si to go public via $1.46 billion SPAC merger. Reuters (2021).

Summary

Proteins are the workhorses of the cell, as such, they form the basis of all living systems. In order to fully understand biological processes, the ability to identify and quantify the proteins in cells is crucial. Identification can be achieved by determining the amino acid sequence of proteins, since this sequence is unique for each protein. However, protein sequencing remains an enormous challenge. The dynamic range at which proteins can occur spans several orders of magnitude, and the need to identify all 20 different amino acids are only a few of the challenges that are currently preventing us from sequencing proteins. However, when realized, single-molecule protein sequencing will create the opportunity for single-cell proteomics and screening for on-site medical diagnostics. It will lead to a revolution in biophysics, biotechnology, and healthcare.

Fluorescence techniques belong to one of the most commonly used techniques in biophysics and have brought about a deeper understanding of biological processes at the single-cell or single-molecule level. Furthermore, the field of DNA sequencing has demonstrated that the use of fluorescence approaches has enabled one of the main breakthroughs in DNA sequencing: the development of next-generation sequencers (NGS). These NGS greatly reduced the sequencing costs per human genome. It comes as no surprise that fluorescence approaches are being explored for protein sequencing as well. In this thesis, we pioneer with single-molecule FRET (fluorescence resonance energy transfer) and DNA nanotechnology approaches for the development of a protein identification platform.

Several technical challenges are limiting the pace at which single-molecule protein identification techniques can be developed. For example, the large difference in protein abundance and the ability to distinguish 20 different amino acids belong to the most challenging hurdles to overcome. Therefore, many single-molecule identification approaches are being explored that circumvent these challenges. In **Chapter 1**, we present an overview of these emerging technologies and summarize the relevant developments in the field of single-molecule protein sequencing.

For fluorescence-based protein identification, the main challenges are the lack of organic fluorophores for the detection of 20 different amino acids without substantial signal crosstalk and the absence of orthogonal chemistry to label each of the 20 amino acids. In recent years, our group and others proposed a simpler idea: protein fingerprinting, in which proteins can be identified by probing only a subset of all amino acids. In addition to the challenges associated with the primary protein structure, a single-molecule protein fingerprinter must be able to obtain fingerprints from full-length folded protein substrate. For this, our first developed approach utilized FRET and ClpXP to scan proteins. ClpXP is a naturally occurring enzyme that can unfold and translocate protein substrates in a controlled manner. In **Chapter 2** we present the first proof-of-concept of a single-molecule peptide fingerprinter and demonstrated that the order of amino acids can be read in (acceptor labeled) peptide substrates using our (donor labeled) ClpXP fingerprinter.

Notwithstanding the exciting proof-of-concept data obtained with our ClpXP

fingerprinting, several challenges remain, including enzyme activity and complicated FRET fingerprints. This inspired us to explore alternative approaches for fluorescence-based protein identification. In **Chapter 3** we introduced a new high-resolution technique that combines single-molecule FRET and DNA nanotechnology. We used programmable, transient binding between short (donor or acceptor labeled) DNA strands to resolve the FRET efficiency of multiple fluorophore pairs in a single molecule. By ensuring only a single FRET pair is formed at a time, we could determine the pairwise distance with sub-nanometer precision. We coined this approach FRET X for FRET via DNA eXchange. While **Chapter 3** focused on the development and validation of the high-resolution FRET X approach, in **Chapter 4** we evaluated the use of our FRET X approach for protein fingerprinting. This new protein fingerprinting approach relies on localizing a subset of amino acids within the 3D protein structure. Our FRET X protein fingerprinting approach requires the attachment of small DNA strands to a subset of amino acids (e.g. cysteines and lysines) as well as conjugation of a reference sequence to either of the protein termini. By flushing in complementary fluorescently labeled DNA imager strands, we can determine the FRET efficiency (thus location) of cysteines and lysines to the reference point. In **Chapter 4**, we simulated fingerprints for hundreds of proteins using a coarse-grained lattice model and experimentally demonstrated FRET X fingerprinting on model peptides. In a simulated complex mixture of >300 human proteins our approach was able to identify constituents with 95% accuracy when considering only the cysteines, lysines, and arginines.

The results presented in **Chapter 5** demonstrate the use of FRET X for fingerprinting full-length protein substrates. We validated the precision of FRET X for different cysteines within a biomedically relevant protein: alpha-synuclein. Furthermore, we demonstrated that FRET X is capable of detecting multiple cysteines within a single protein. We further increased the complexity of the protein substrates and localized the cysteines within globular biomarkers Bcl_{xL} and Homer-1. We observed reproducible FRET X fingerprints for both proteins, demonstrating that FRET X can reliably fingerprint protein with a globular structure.

Our FRET X protein fingerprinting approach utilizes the predictable and programmable features of DNA hybridization, which allows for customization and a broad range of applications for our technology. However, one of the drawbacks of DNA is the slow binding rate, which necessitates long acquisition times to obtain high-resolution data. The results presented in **Chapter 6** aim to overcome this slow binding rate by preloading the DNA imager strands into an Argonaute protein (Ago). Ago preforms the imager strand, which allows the Ago-loaded imager strand to rapidly bind to the target. In this manner, the binding rate achieves a near-diffusion limited speed and is an order of magnitude higher compared to naked imager strands. In **Chapter 6** we utilized the fast binding property of Ago and demonstrated that it can speed up the existing DNA-based super-resolution technique: DNA-PAINT. For our proof-of-concept experiments, we assembled a DNA origami plate with four binding sites and show that Ago-assisted DNA-PAINT can obtain super-resolved images 10x faster compared to conventional DNA-PAINT. We envision the use of Ago-PAINT as a general strategy to accelerate existing applications that rely on base-pairing interactions, including FRET X protein fingerprinting.

The field of single-molecule protein sequencing has seen tremendous progress and produced several dozens of high impact papers. However, the realization of a single-molecule protein sequencer is technically challenging. In **Chapter 7**, I discuss some of the most significant challenges and contemplate about future directions for the field of single-molecule proteins sequencing.

Samenvatting

Eiwitten zijn de werkpaarden van de cel en vormen zodanig de basis van alle levende systemen. Om biologische processen volledig te begrijpen, is het vermogen om de eiwitten in cellen te identificeren en te kwantificeren cruciaal. Identificatie van eiwitten kan worden bereikt door de aminozuurvolgorde van eiwitten te bepalen, aangezien deze volgorde voor elk eiwit uniek is. Het sequencen van eiwitten blijft echter een enorme uitdaging. Het dynamische bereik waarin eiwitten kunnen voorkomen omvat verschillende ordes van grootte en de noodzaak om alle 20 verschillende aminozuren te identificeren zijn slechts enkele van de uitdagingen die ons momenteel belemmeren om eiwitten te sequencen. Wanneer dit echter wordt gerealiseerd, zal sequentiebepaling van één eiwitmolecuul per keer de mogelijkheid creëren voor de proteomica van een enkele cel en voor medische diagnostiek welke ter plaatse kan worden uitgevoerd. Het zal leiden tot een revolutie in de biofysica, biotechnologie en gezondheidszorg.

Fluorescentietechnieken behoren tot een van de meest gebruikte technieken in de biofysica en hebben geleid tot een dieper begrip van biologische processen op zowel cellulair als moleculair niveau. Bovendien heeft het veld van DNA-sequencing aangetoond dat het gebruik van deze fluorescentietechnieken een van de belangrijkste doorbraken in DNA-sequencing mogelijk heeft gemaakt: de ontwikkeling van 'next-generation sequencers' (NGS). Deze NGS hebben de sequencing kosten, per menselijk genoom, aanzienlijk verlaagd. Het is geen verrassing dat de toepassing van fluorescentietechnieken ook wordt onderzocht voor eiwit sequencing. In dit proefschrift pionieren we met enkel molecuul FRET (fluorescentie resonantie energieoverdracht) en DNA nanotechnologie voor de ontwikkeling van een eiwit identificatieplatform.

Verschillende technische uitdagingen beperken het tempo waarin technieken voor de identificatie van eiwitten, één molecuul per keer, kunnen worden ontwikkeld. Het grote verschil in eiwitconcentratie per cel en het vermogen om 20 verschillende aminozuren te onderscheiden, behoren bijvoorbeeld tot de meest uitdagende hindernissen voor de realisatie voor deze techniek. Daarom worden er veel benaderingen voor identificatie van één eiwitmolecuul onderzocht. In **Hoofdstuk 1** presenteren we een overzicht van deze opkomende technologieën en vatten we de relevante ontwikkelingen op het gebied van sequencing op een enkel molecuul niveau samen.

Voor de op fluorescentie gebaseerde eiwitidentificatie zijn de belangrijkste uitdagingen het ontbreken van organische fluoroforen voor de detectie van 20 verschillende aminozuren zonder substantiële signaal overlap. Een andere uitdaging is de afwezigheid van orthogonale chemie om elk van de 20 aminozuren te labelen. In de afgelopen jaren hebben onze groep en anderen een eenvoudiger idee voorgesteld: eiwit-vingerafdrukken, waarbij eiwitten kunnen worden geïdentificeerd door het detecteren van slechts een subset van alle aminozuren. Naast de uitdagingen die gepaard gaan met de primaire eiwitstructuur, moet een eiwitvingerafdruk scanner in staat zijn om vingerafdrukken te verkrijgen van volledig gevouwen eiwitten. Voor

onze eerste eiwitvingerafdruk scanner hebben we gebruikt gemaakt van FRET en ClpXP om eiwitten te scannen. ClpXP is een natuurlijk voorkomend enzym dat op een gecontroleerde manier eiwitsubstraten kan ontvouwen en verplaatsen. In **Hoofdstuk 2** presenteren we de eerste proof-of-concept van een eiwitvingerafdruk scanner die individuele moleculen kan scannen. Ook demonstreren we dat de volgorde van aminozuren kan worden gelezen in (acceptor-gelabelde) peptidesubstraten met behulp van onze (donor-gelabelde) ClpXP-scanner.

Ondanks de veelbelovende proof-of-concept data die zijn verkregen met onze ClpXP-vingerafdruk scanner, heeft deze techniek een aantal praktische nadelen, waaronder enzymactiviteit en gecompliceerde FRET-vingerafdrukken. Dit inspireerde ons om alternatieve benaderingen te onderzoeken voor eiwitidentificatie op basis van fluorescentie. In **Hoofdstuk 3** hebben we een nieuwe hoge resolutie techniek geïntroduceerd die FRET en DNA-nanotechnologie combineert. We gebruikten programmeerbare, tijdelijke bindingen tussen korte (donor- of acceptor-gelabelde) DNA-strengen om de FRET-efficiëntie van meerdere fluorofoor-paren in een enkel molecuul op te lossen. Door ervoor te zorgen dat er slechts één FRET-paar tegelijk wordt gevormd, kunnen we de paarsgewijze afstand bepalen met een precisie van minder dan een nanometer. We hebben deze benadering FRET X genoemd: FRET via DNA eXchange.

In **Hoofdstuk 3** beschreven we de ontwikkeling en validatie van de FRET X-techniek met hoge resolutie, evalueerden we in **Hoofdstuk 4** het gebruik van onze FRET X-techniek voor eiwitvingerafdrukken. Deze nieuwe benadering van eiwitvingerafdrukken is gebaseerd op het lokaliseren van een subset van aminozuren binnen de 3D-eiwitstructuur. Onze FRET X eiwitvingerafdruk aanpak vereist de hechting van kleine DNA-strengen aan een subset van aminozuren (bijv. cysteïnes en lysines) evenals conjugatie van een referentiesequentie aan een van de eiwittermini. Door complementaire fluorescent gelabelde DNA imager strengen toe te voegen, kunnen we de FRET efficiëntie (dus locatie) van cysteïnes en lysines naar het referentiepunt bepalen. In **Hoofdstuk 4** hebben we vingerafdrukken voor honderden eiwitten gesimuleerd met behulp van een 'coarse-grained lattice model' en hebben we experimenteel laten zien dat we FRET X vingerafdrukken van model peptiden kunnen meten. In een gesimuleerd complex mengsel van >300 menselijke eiwitten was onze techniek in staat om eiwitten met 95% nauwkeurigheid te identificeren wanneer de vingerafdruk van alleen de cysteïnes, lysines en arginines werd gemeten.

De resultaten gepresenteerd in **Hoofdstuk 5** demonstreren het gebruik van FRET X voor het bepalen van de vingerafdrukken van eiwitten van volledige lengte. We hebben de precisie van FRET X gevalideerd door verschillende cysteïnes binnen een biomedisch relevant eiwit (alfa-synucleïne) te lokaliseren. Verder hebben we aangetoond dat FRET X in staat is om meerdere cysteïnes binnen een enkel eiwit te detecteren. We hebben de complexiteit van de eiwitsubstraten verder verhoogd en de cysteïnes gelokaliseerd in de gevouwen biomarkers Bcl_{xL} en Homer-1. We hebben reproduceerbare FRET X vingerafdrukken waargenomen voor beide eiwitten, wat aantoont dat FRET X betrouwbaar vingerafdrukken kan maken van een enkel eiwit met een gevouwen structuur.

Onze FRET X vingerafdruk techniek maakt gebruik van de voorspelbare en programmeerbare kenmerken van DNA hybridisatie, wat maatwerk en een breed

scala aan toepassingen voor onze technologie mogelijk maakt. Een van de nadelen van DNA is echter de langzame bindingssnelheid, waardoor lange acquisitietijden nodig zijn om data met een hoge resolutie te verkrijgen. De resultaten gepresenteerd in **Hoofdstuk 6** zijn bedoeld om deze langzame bindingssnelheid te overwinnen door de DNA imager strengen vooraf te laden in een Argonaute eiwit (Ago). Ago vormt de imager streng voor, waardoor de met Ago geladen imager streng snel aan het doelwit kan binden. Op deze manier bereikt de bindingssnelheid een bijna diffusie-begrensde snelheid en is deze een orde van grootte hoger in vergelijking met niet-geladen imager strengen. In **Hoofdstuk 6** hebben we de snelle bindingseigenschap van Ago gebruikt en aangetoond dat het een bestaande, op DNA gebaseerde, superresolutietechniek (DNA-PAINT) kan versnellen. Voor onze proof-of-concept experimenten hebben we een DNA-origamiplaat met vier bindingsplaatsen samengesteld. Op deze samengestelde plaat laten we zien dat Ago-assisted DNA-PAINT (Ago-PAINT) tien keer snellere super-resolutie beelden kan verkrijgen in vergelijking met conventionele DNA-PAINT. We zien het gebruik van Ago-PAINT als een algemene strategie om bestaande toepassingen te versnellen die afhankelijk zijn van basenparende interacties, waaronder onze FRET X vingerafdruk techniek.

Binnen het gebied van het sequencen van een enkel eiwitmolecuul is een enorme vooruitgang geboekt en zijn er tientallen publicaties uitgebracht die een grote bijdrage hebben geleverd. De realisatie van een eiwit-sequencer, met een gevoeligheid van één enkel molecuul, is technisch zeer uitdagend. In **Hoofdstuk 7** bespreek ik enkele van de belangrijkste uitdagingen en blik ik vooruit op de toekomstige richtingen op het gebied van het sequencen van eiwitten.

Acknowledgments

During my PhD I got to learn many different things, but the most important thing is that this journey would not have been possible without all the amazing people that I got to meet and work with. But also the people with whom I shared many laughs, dinners, borrels, and many other fun activities outside the lab. You made the past years truly an amazing experience that I will remember for the rest of my life! This makes it very strange, but at the same time very rewarding to write this chapter. This is for you!

First of all, I would like to start by expressing my gratitude to my mentor and promotor, **Chirlmin**. I want to thank you for believing in me and inviting me to your lab for my PhD study. I appreciate and admire the way you manage the lab. The way you treat each lab member the same way, together with your brilliant and creative scientific knowledge makes you the perfect supervisor. I want to thank you for teaching and pushing me to think outside the box, because the crazier the idea, the better they are. I have always felt your full support throughout my scientific adventures and your encouragement to pursue my own ideas has been key to making my PhD such a success. I think your lab is a great and inspiring environment where people are very happy to work and can pursue their scientific ambitions. Joining your lab for my PhD was the best decision I made 5 years ago. Thank you!

Cees, thank you for also being my promotor. Despite our limited interactions, I always enjoyed the lively discussions we had. We discussed old school videogames such as Prince of Persia, to what the best approach for protein sequencing would be. I admire your energy and enthusiasm for everything you do. With this, you have been able to create an excellent scientific environment in your lab, but also within the entire department. An achievement you can be proud of.

I would also like to express my gratitude to my defence committee, Prof. **Giovanni Maglia**, Prof. **Marthe Walvoort**, Prof. **Peter Steeneken**, Prof. **Tae-Young Yoon**, and Prof. **Marileen Dogterom**. Thank you for taking the time to evaluate my PhD thesis and for taking part in the defence ceremony.

Apart from the great supervisors that I worked with over the past few years, I want to take this moment to thank two of my very first mentors, **Edward Drost** and **Jeroen Ouwehand**. I was fortunate to have Edward as my mentor during the last two years of my VMBO high school years, and he inspired me to continue working within the subjects that I liked the most: Chemistry, Mathematics, and Physics. I decided to follow his advice and joined Zadkine, where I started to take courses for becoming a technician in Biotechnology. For the last two years of this study, I was taking classes from Jeroen to dive into some of the core principles of biochemistry and biotechnology, and he always said that sky is the limit. Jeroen suggested to think outside the box when looking for my first research internship. With your help , I decided to join the Grant Booker lab, at the University of Adelaide in Australia. This was the place where I was inspired to continue studying and pursue a PhD study.

A special thanks to my friend from down under, **Tatiana**. When we just met I was a complete newbie in science but am truly grateful that you were always there to help me, both inside and outside the lab. You are the wise Oma that I could ask anything ;). I am happy that I have met you and am always looking forward to meeting up every 3 years to go for a 'broodje kroket' at Van Dobben in Amsterdam. Maybe someday, when both of us are tired of science, we can open our small food place: Gezellig (with two g's). I hope to see you soon!.

This journey started way before the start of my PhD, 5 years ago in Delft. I was very fortunate to be surrounded by great people that made all these years of studying fun. I never got to thank you for the great time we had. It all started at Zadkine, I want to thank **Menno** and **Santoesha** for the great time we had in Adelaide. I also want to thank the people that I met in Breda at AVANS. Thanks, **Patrick**, **Natassia**, **Barbara**, and **Brechje** for accepting us, the people from above the rivers, in Breda. I had a great time there! The last stop before my PhD was The Vrije Universiteit in Amsterdam. I want to thank **Jana** and **Rosanne** for the great time! The dinners, drinks, and turbo doners we had were amazing. We always say that we should do this more often, I hope that we can do this soon!

A special thanks to **Sven Dekker**, I am great to have met you at Zadkine, and was very happy that we went to Breda and Amsterdam together. We spend many hours on the train discussing 'what the actual F we started, and why did we do this' but it was all worth it! Even the many Exam retakes were worth it, because we reminded ourselfs: 'you also pay for the retakes, so why not do them as well' :). I am happy that you found a nice lab to work in, and of course that you are a father now. I wish you all the best, and perhaps our scientific roads cross again in the future.

I was very fortunate to join the Joo Lab not once, but twice. My first time in the Joo lab was for my master's research project in 2015 and was a bit overwhelming. Where I was trained as a Biochemist, I had no idea what single-molecule biophysics was. However, the Joo lab members made me feel at home and were very patient to answer all of my questions. I want to thank **Luuk**, **Laura**, **Malwina**, **Pawel**, **Jetty**, **Mohamed**, **Stanley**, **Ivo**, and **Viktorija** for creating a nice atmosphere in the lab during my master's project. The great atmosphere in the lab was one of the main reasons for me to join the Joo Lab for my PhD.

The past 5 years in BN and the Joo lab were truly an amazing experience. When I just joined the lab I knew a few people that were still around from my masters' project, but also many new people. **Luuk**, I am very happy that you were still around in the Joo lab at the time when I started my PhD. I could come to you for answering my questions and for pro tips, thanks for that! I enjoyed our chats about what the best would be to throw half a cow on the BBQ. We made an excellent team in the kitchen, as we organized a few BBQs that I think were always lots of fun to do and a huge success! I find it unfortunate that there was little overlap between our stays in the lab because I think that we could've done some cool projects together as we have the same mindset in the lab. But who knows what the future holds! I wish you, Mierelle, and little Oscar all the best for the future, and am curious where your lab will be. **Laura**, Oma, my partner in crime in the protein sequencing project. The

idea was indeed very simple in itself, however, the experiments and all others were not so much... Although our sequencing approaches were very different, it was great chatting with you about the many challenges we had to overcome and sharing our love for cysteines! I learned a lot from you both in and outside the lab. It was great having you around and always gezellig to go out for bowling, drinks, and dinner. **Thijs**, thanks for answering all my physics and Matlab related questions. I think we had a great time in the Joo Lab where we often would brainstorm about friday afternoon projects. This resulted in the start of something completely new to our lab, super-resolution imaging. It was a lot of fun working on this together with you! I would also like to thank you for the nice conference trips we made together. **Viktorija**, V! It was great to have you around in the lab, sometimes you could complain a bit much ;), but I enjoyed all the Cokes we shared (I think you still owe me one for the time you forgot your card) talking about anything, except science! On Monday we would recap all the football games of the past weekend and once we even went to 'De Coolsingel' to celebrate the Feyenoord Championship. You were the guardian of the microscope and helped me with all questions on this. Thank you for this! **Sungchul**, my Korean friend and THE biologist of the lab. I know you will probably not believe anything that I wrote in this thesis but I truly enjoyed every conversation we had. I admire your creativity and appreciate that I always could come over for some advice on experiments. Apart from being an excellent researcher, your skills in the kitchen are perhaps even better. I thank you for the dinners you organized and the great food you prepared. I wish you all the best in running your lab! Your students will have a difficult time working to your standards. But I guess you just need to start believing! **Ilja**, the Oma of the lab :). I am so happy that you joined the sequencing team, your contribution to this thesis is enormous. It didn't matter which crazy protein we found in the literature, you would design some vector, place some order and most of the time successfully purify the proteins, thanks for this! It was also great being in the same office and sharing the love/hate relationship both of us have with Feyenoord. **Alessia** and **Cecilia**, I want to thank both of you for joining the sequencing team and for your help with protein purification. I enjoyed working together with both of you and I hope that you are happy with your new jobs as well! **Ivo**, I enjoyed the cups of coffee and chats we had in the office about wrapping up this PhD thing! I should also thank you for all the nice Xmas and birthday cakes you made, they were amazing! Your project is very challenging but I think you are very close to making it work. I wish you all the best with the 'laatste loodjes' of wrapping up your PhD! **Iasonas**, my Greek friend! What a joker you are! It was great to share an office with you. In the beginning, I wasn't sure what kind of person I was sitting next to, but soon I realized that you are one of the craziest people I met in BN. We would crack many jokes, making it always a lot of fun to go to the office (well at least when you were coming to the office ;)). I admire your theoretical knowledge and it was great getting tips and tricks on data analysis from you. **Adam**, it was great when you joined the project. You brought a lot of knowledge on the more chemistry side of the project. Thanks for always being available for small chats and troubleshooting on labeling problems or other challenges we had to overcome in the project. I wish you all the best with setting up your own lab! **Carolien**, I think you brought great momentum to the high-throughput team. I like your go-getter attitude which is needed for such

a challenging project. I want to thank you for organizing a nice group retreat and hope that we can organize some nice borrels/dinners again shortly. Before you know, you will be one of the older PhD students in the lab and people will come to you for all their questions ;). I wish you all the best for the rest of your PhD project!. **Sung Hyun**, it was great news that you joined our lab again. I want to thank you for the analysis code, which made the analysis much more easily, and also for the help with other data analysis! I think we should try and bring back our 15:00 home brew coffee moment again!. **Bhagyashree**, it was great to have you join the sequencing project. I admire the number of projects that you started and are currently working on. Thanks for organizing a great lab retreat! **Jack**, welcome to the protein sequencing project! It was great to discuss and start several PTM related projects together during the past months. Let's keep pushing and make them work ASAP! **Kijun** and **Koushik**, welcome to the lab. Thanks for the already great conversations we had so far, and I wish both of you all the best for your stay in the Joo Lab. **Margreet**, thank you for making sure that both the setups and software kept doing what they should be doing! **Jan**, thank you for helping with the ordering and making sure that all is organized well in the lab. I also enjoyed our chats about video games, perhaps we meet in Anno1800 someday! A special mention and thanks to the honorary members of the Joo Lab – Mischa and Mathia! **Mischa**, I am happy to have met you and it was a lot of fun visiting Chicago after BPS! **Mathia**, I am grateful to have met you. You were always very kind and positive! I am happy that you joined the Joo Lab drinks and BBQs. I remember that when I invited you to the first Joo Lab BBQ at my house, you came to me and asked if we would prepare some vegetarian dishes. Luuk and I were a bit hesitant but looked into it and prepared some greens on the grill. However, when all the food was prepared, I think you destroyed 2kg of spare ribs and half a chicken – I think you are the worst vegetarian that I know but I loved it :)! We miss you!

I would also like to thank the students that were brave enough to join the challenging protein sequencing project for their BEP or MEP. Thanks, **Anna**, **Nicola**, and **Isabell** for joining the team. It was great working together with you on this challenging project! I learned a lot from you, thank you for this! A special thank you to **Raman,** or RW. I think we made a great team together which resulted in great scientific output! Apart from the many scientific discussion we had, I enjoyed our chats about life and all the crazy things that are happening in the world right now! I appreciate your work ethic, scientific knowledge, and dedication and I wish you all the best for your PhD journey.

The great thing about science is the ability to work together with people from different disciplines. **Rienk**, thank you for your help, advice, and suggestions for many of the chemistry challenges during my PhD project. **Tobias**, I am happy that I could contribute to two of your projects. It was always a lot of fun having you around and doing some experiments together! **Peggy**, you work on a crazy project, but I think we are very close to finalizing the project. Goodluck with the final experiments! An important part of this thesis involved bioinformatic analysis and validation of our new fingerprinting approach, for this I want to thank Dick and Carlos. **Dick**, thank you for your time and valuable comments during the FOM meetings and on the

manuscript that we prepared together. **Carlos**, thank you for pressing 'enter' and doing the actual computational work! I am very happy to have worked with you and I learned a lot from you on the computational side of the projects. I can happily say that we made the Xmas deadline, again! :) Good luck with wrapping up your PhD and your post-doc in the Joo Lab, I can show you how the wet lab experiments work.

A big thank you to all the amazing people from BN. Thanks to **Sonja**, **Alessio**, **Daniel V**, **Jorine**, **Fede**, **Stephanie**, **Sandro**, **Henry**, **Eugene**, **Adi**, **Ghanji**, **Benjamin**, **Becca**, **Jochem**, **Sam**, and many others for making the department a very lively and enjoyable place! Special thanks to **Martin Depken**, for making sure all runs smooth in BN and for trying to bring back all the social events!

I obtained a lot of support from friends and family outside the walls of BN and TU Delft. I want to thank **Swen** and **Jorik** – the CoD team. There is nothing better than to play a few 'relaxing' games of CoD after a frustrating day in the lab. Despite many frustrating rounds, I think we always had a lot of fun playing those games. We just need to remember the most important rule in gaming, when the game says failed – you failed. **Alex**, **Vincent**, and **Jarno**, even though we don't see each other too often, I think we always have fun when meeting up and hope that we can do this more often in the future. **Mac** and **Marleen**, the pizza masters of the family, I always enjoy coming over for a BBQ or pizza party at your place. Although we haven't done this in a long time we should do so again next year! I want to thank **Kim** & **Joey**, **Emma** & **Igor**, **Lisa**, and **Claire** for the game nights, borrels, and dinners that I could join together with Naomi.

Daan and Coach Dave, thanks for opening the gym and making sure that the body also stays fit. **Daan**, it was great sharing some of the scientific frustration with an expert in the family! **Coach Dave**, thanks for training Sandy and me. I appreciate that you were always open to suggestions and adjustments for exercises that Sandy or myself had. **Sandy**, it was great to have you as my training partner in Dave's Gym, thanks for all the laughs! **Opa** and **Oma**, thank you for always showing interest in my work and sharing the papers with the rest of the family! **Ome Leen**, thanks for the nice boat rides in Amsterdam, and we should go and have dinner together at Restaurant Dynasty!

A special thanks to the Spekkies! Beau and Reinier, it is always great to have you over or visit you for a cup of coffee or dinner. **@Beaudines**, I am happy to have you as my 'schoonzus', I think we always have a lot of fun now and I hope that I can rely on your help and advice for picking gifts for many years to come :) **Reinier**, I admire your creativity in the kitchen and wish you all the best for the future of Frisj Mints! **Bert** and **Luus**, I am very happy to have you as my 'schoonouders'. The beginning of me joining your family was a little bit strange for me because I was dating the daughter of the friends of my parents that also happen to be the employer of my sister. But you made me feel at home right from the start! I admire the dedication and hard-working mindset that both of you have. I hope that now that you received the keys to your apartment, you can enjoy life at the coast of Spain more often!

Nikkie, I am very happy to have you by my side as my paranymph during the defence ceremony for two reasons: 1, you are likely to be more nervous than I am, and 2, you are my best friend. Ever since we were little, we were great together and I am super happy that we still are. I always enjoy our chats and complaints at home with a cup of coffee or over face time. Last August you shared the exciting news that the two of you will be are expecting a little girl in April next year. I am super proud of you, and I am sure that you will be the best mother that Maud can have! I can't wait to meet the little girl.

Folkert, we know each other for most of our lives. It all started at the baseball fields where we, together with Robert-Jan, belong to the try-hard players that had all the gadgets that looked amazing. But over the last 10 years you become part of the family for being together with Nikkie. I am proud to call you my 'zwager' and grateful that you make Nikkie happy! I am sure you will be a great father and wish you all the luck in the world together with your girls! I am happy to also have you by my side as a paranymph during my defence, if needed you would be able to answer some of the biology questions!

To **my parents**, it is difficult to put everything I feel into so few words but I am quite certain that without your support and love, this PhD journey would not have been the success that it is. I want to thank you for the support you always provided and the great environment that you created at home where I could pursue my dreams! **Dad**, you have been the most important role model in my life and you showed that with hard work and dedication anything is possible, a trait that I try to live by every day. I enjoyed our diving trips to Malta and hope that we can start doing this again soon! **Mum**, together with Wim, you are my number one fan and always supported the decisions I made. You always put others ahead of yourself, a trait that I admire and wish to have. You make sure that everybody is happy and ogranize dinners wehere everybody comes together, I like to think that I got this trait from you and this helped me make many friends both in and outside the scientific world! I am proud and jealous that the two of you are enjoying life to the fullest by traveling Europe for several months each year! I appreciate that you invite Nikkie, Folkert, Naomi, and myself (+ the future grandkids) for great weekends away and I hope that we can do this for many years to come! I don't say this enough, but I love both of you. This thesis is for you!

Naomi, I am so happy that I have met you. It was during a very crazy time because dating during lockdown is not the easiest thing but I am happy to have done this with you and can look back at a great time! We spend a great amount of time together already (also in quarantine), but could not have done this with anyone else. I am very happy with all the things we do togther, from baking cakes to making puzzles in record time. I want to thank you for your help in the lab, I think you are the Zeba master (the small tubes with the red caps). I am very proud of you for starting your new adventure after the Xmas break in 'groep 7'! You are an amazing woman and I am lucky to have you! I am ready and excited for the many new adventures that we will take together in our life. I love you!

Mike Filius
Delft, December 2021

Curriculum vitae

Mike Filius

14 November 1990	Born in Vlaardingen, The Netherlands
2003 - 2007	Intermediate Preparatory Vocational Education (VMBO-TL) Kastanje College, Maassluis, The Netherlands
2007 - 2011	Intermediate Vocational Education (MBO) in Biotechnology and Laboratory Research ROC Zadkine School of Laboratory Techniques, Rotterdam, The Netherlands.
2011 - 2014	B.Sc. in Molecular and Biotechnology Laboratory Research AVANS Hogeschool, Breda, The Netherlands
2014 - 2017	M.Sc. in Biomolecular Science Vrije Universiteit Amsterdam, The Netherlands
2017 - 2021	Ph.D. in Biophysics Title: "Next-Generation Protein Identification: Advancing Single-Molecule Fluorescence Approaches" Promotor: Prof.dr. C. Joo Promotor: Prof.dr. C. Dekker Department of Bionanoscience Delft University of Technology, The Netherlands

List of Publications

* These authors contributed equally

13. **Filius, M.**, van Wee, R., de Lannoy, C., Westerlaken, I., de Agrela Pinto, C., de Ridder, D., and Joo, C. "Single-Molecule Protein Identification Using FRET X." (*Manuscript in Preparation).*

12. **Filius, M.** and Joo, C. "High-Resolution Single-Molecule FRET X", Springer Book: "Single-Molecule Analyis: Methods and Protocols", (*Bookchapter In Preparation)*

11. de Lannoy, C*., **Filius, M.***, van Wee, R., Joo, C. & de Ridder, D. "Evaluation of FRET X for Single-Molecule Protein Fingerprinting". *iScience* 24, 103239 (2021).

10. van Wee, R.*, **Filius, M.***, and Joo, C. "Completing the canvas: advances and challenges for DNA-PAINT super-resolution imaging." *Trends in Biochemical Sciences* (2021).

9. Alfaro, J.A.*, Bohländer, P.*, Dai, M.*, **Filius, M.***, Howard, C.J.*, van Kooten, X.F.*, Ohayon, S.*, Pomorski, A.*, Schmid, S.*, [...], Meller, A., and Joo, C. "The emerging landscape of single-molecule protein sequencing technologies". *Nature Methods*. 18, 604–617 (2021).

8. Joo, C., **Filius, M.**, De Lannoy, C. & De Ridder, D. "Single-Molecule FRET For Protein Characterization." WO Patent 2021049940 (2021).

7. **Filius, M.**, Kim, S.H., Severins, I., and Joo, C. "High-Resolution Single-Molecule FRET via DNA eXchange (FRET X)." *Nano Letters*. 21, 3295–3301 (2021).

6. Brevé, T.G., **Filius, M.**, Weerdenburg, S., van der Griend, S.J., Groeneveld, T.P., Denkova, A.G., Eelkema, R. "Light sensitive phenacyl crosslinked dextran hydrogels for controlled delivery." Preprint at *ChemRxiv*. 2021.06.28. (2021).

List of publications

* These authors contributed equally

5. de Lannoy, C., **Filius, M.**, Kim, S.H., Joo, C., and de Ridder, D. "FRETboard: Semi-supervised classification of FRET traces." *Biophysical Journal.* 120, 1–8 (2021).

4. Brevé, T.G., **Filius, M.**, Araman, C., van der Helm, M.P., Hagedoorn, P.L., Joo, C., van Kasteren, S.I., Eelkema, R. "Conditional copper-catalyzed azide alkyne cycloaddition by catalyst encapsulation." *Angewandte Chemie Int. Ed.* 59, 9340–9344 (2020).

3. **Filius, M.*,** Cui, T.J.*, Ananth, A.N., Docter, M.W., Hegge, J.W., van der Oost, J., and Joo, C. "High-Speed Super-Resolution Imaging Using Protein-Assisted DNA-PAINT." *Nano Letters*. 20, 2264–2270 (2020).

2. Lageveen-Kammeijer, G.S.M.*, de Haan, N.*, Mohaupt, P., Wagt, S., **Filius, M.**, Nouta, J., Falck, D., and Wuhrer, M. "Highly sensitive CE-ESI-MS analysis of N-glycans from complex biological samples." *Nature Communications*. 10, (2019).

1. van Ginkel, J., **Filius, M.**, Szczepaniak, M., Tulinksi, P., Meyer, A.S., and Joo, C. "Single-molecule peptide fingerprinting." *Proc. Natl. Acad. Sci. U.S.A.* 115, 3338–3343 (2018).