FROM SQUIGGLE TO SEQUENCE

BIOINFORMATICS IN THE ERA OF SINGLE-MOLECULE BIOPOLYMER ANALYSIS

CARLOS VICTOR DE LANNOY

Propositions

1. Electric and fluorescent sensing fulfill complementary niches in the single-molecule identification of heteropolymers.
   (this thesis)

2. Eliminating the need for fragmentation of biopolymers is imperative to make analyses more routine, better quality and cheaper.
   (this thesis)

3. Scientific equipment manufacturers benefit from an open science approach through the ingenuity and labor of their users.

4. Educating scientists on popular science journalism practices is key to mitigate the spread of sensationalized news.

5. Anonymity induces disregard of societal norms in peer reviewers.

6. Cursing during coding sessions increases code quality by mitigating frustration.

Propositions belonging to the thesis, entitled

From squiggle to sequence: bioinformatics in the era of single-molecule biopolymer analysis

Carlos Victor de Lannoy
Wageningen, 8 November 2022

From squiggle to sequence: Bioinformatics in the era of single-molecule biopolymer analysis

Carlos Victor de Lannoy

Thesis committee

Promotor
Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Other members
Prof. Dr A.H. Velders, Wageningen University & Research
Dr T.E.P.M.F. Abeel, Delft University of Technology
Dr S. Schmid, Wageningen University & Research
Dr J.A. Alfaro, University of Gdànsk, Poland

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

From squiggle to sequence: Bioinformatics in the era of single-molecule biopolymer analysis

Carlos Victor de Lannoy

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University,
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 8 November 2022
at 4 p.m. in the Omnia Auditorium.

Carlos Victor de Lannoy
From squiggle to sequence: bioinformatics in the era of single-molecule biopolymer analysis

PhD thesis, Wageningen University, Wageningen, The Netherlands (2021)
With references, with summary in English

ISBN: 978-94-6447-408-4
DOI: https://doi.org/10.18174/577300

CONTENTS

CHAPTER 1

Introduction

1.1 The age of single-molecule analysis

In 2003, the human genome project (HGP) announced that it had drafted the human genome for the first time [1]. This blueprint, encoded in the ordering of three billion deoxynucleic acid (DNA) base pairs, opened previously unimaginable venues to the understanding of cell functioning and disease mechanisms. This first draft was far from complete however; polymorphic regions and large repetitive stretches, accounting for eight percent of the genome, had been left unsequenced [2]. The methods of the time produced relatively short reads, causing long sequences of repetitive nature to collapse into fewer repeats, while polymorphic regions remained difficult to analyse due to the dissonance they caused in this mosaic assembly of multiple individuals. Only in March of 2021 did the telomere-to-telomere consortium publish what can be more accurately called a "fully" sequenced genome of a single individual, leaving only 0.3 percent of the genome unsequenced [3]. Over the coming years, the newly added regions to the known genome will doubtlessly shed light on a range of essential cell functions such as ribosome biogenesis [4], nucleolus formation [5] and chromosomal instability [6], as well as on associated diseases.

Although many innovations in sequencing technology and methodology have led up to this milestone, the newly developed third generation of DNA sequencing (TGS) technology – single-molecule (SM) and focused on producing long reads – provided a pivotal contribution. While first and second generation methods produced reads too short to resolve repeats, TGS reads are able to span even long repetitive sequences, avoiding their collapse into fewer repeats. Moreover, compared to the bulk-based approach of previous generations, the SM nature of TGS allows straight-forward separation of heterozygous alleles and their separate assembly, which simplifies the analysis of polymorphic regions tremendously. This contribution to genome sequencing is just one way in which SM analysis is revolutionizing biological sciences; from molecular kinetics [7; 8] and structure [9] to detection of biologically relevant yet minute populations of biomolecules [10], investigations at the SM level have produced essential insights that were hitherto obscured in bulk analyses.

Although SM observations have been made since the 70's [11], the past decade has seen a surge in methodological and technical advances. In particular, applications for the analysis of the three biopolymers (BPs) at the heart of the central dogma of molecular biology – DNA, RNA and proteins – have garnered much attention. Arguably, nucleic acid (NA) analysis methods have matured the most, to the point that SM sequencing of DNA and RNA has become standard fare. However even in NA research, there is still plenty of room for growth; beyond SM readouts of ever longer NA chains, the detection of epigenetic modifications, strand structure and interaction kinetics with other biomolecules are areas of continued investigation. As proteins are structurally more complex than NA chains, SM analysis of proteins has been progressing less quickly, however newly proposed methods are finding their way around these difficulties [12–16].

1.2 Readout of properties

1.2.1 Electrical sensing

Although the landscape of SM analysis methods for BPs is diverse and growing rapidly, most can be grouped under one of three categories, based on the nature of the signal that is read out. First, a BP may be interrogated by its modulation of electrical conductance. In an approach referred to as "recognition tunneling", molecular affinity agents for the target BP are bound to electrodes separated by a small gap [17]. Interaction between the affinity agents and a BP induce fluctuations in electrical conductance, which may be indicative of the identity of the BP. A more versatile and popular approach involves confining the BP to a small volume and measuring its influence of electrical conductance through this volume. This is typically done by passing the BP through a nanopore – a nanometer-scale opening in an electrically insulating barrier. So-called solid-state nanopores are burned directly in the barrier using a focused electron beam. Biological nanopores, derived from a naturally occurring pore-forming protein complex, have more reproducible dimensions and have therefore seen more usage. An electrical potential is applied over the nanopore, which moves the BP away from one electrode and toward the other through a combination of electro-osmotic force (EOF) and electrophoretic force (EF). By monitoring either the potential or the current during the passage of the BP through the pore, its effect on the electrical conductance can be measured. Depending on the size, charge and shape of the BP, it may block more or less electrons, thus electrical conductance through the pore may be higher or lower, and stable or fluctuating. Generally, it is desirable to increase the dwell time – the time spent by the BP in the pore – as otherwise the BPs or BP elements may pass through the pore in microseconds, which is too fast to obtain meaningful measurements [18].

For some applications, such as kinetic analysis, the BP needs to retain its functional shape and enter the pore folded. In this scenario several mechanisms to increase dwell time can be employed, such as current modulation to balance the counter-acting EOF and EF [19], or blocking the exit of the pore using a DNA-origami structure [20]. To directly read out the sequence of a BP however, it must be linearized and fed through the pore single-file, from one end to the other. Depending on the nature of the BP, processive motor proteins may be available to slow down the BP and increase the dwell time. The most prominent example of electrical readout of BPs is found in the latter category, in nanopore sequencers for NAs produced by Oxford Nanopore Technologies (ONT). In these sequencers, an NA strand of arbitrary length is fed through a biological nanopore while a helicase ratchets the strand in single or half-nucleotide steps to regulate the processing speed. In recent years, NA nanopore sequencers have garnered attention due to the theoretically unlimited read length, low initial investment cost compared to other sequencing devices and small benchtop footprint.

For either application, the produced data takes the shape of long time series of current or potential measurements, colloquially named "squiggles", in which plateaus at specific current levels are associated with the presence of a single BP –

or a defined fragment of it – in a particular conformation. Data analysis software is tasked to separate these plateaus and assign the correct BP or BP fragment to each. Given the sequential nature of the data, logical choices include hidden Markov models (HMMs) [21–23] and neural networks (NNs) [24–26], although hashing [27] and dynamic time warping-based analysis [28] have also seen successful application. For NA sequencing devices, analysis software takes the shape of so-called basecallers. A basecaller may assign a sequence of length k – a k-mer – to each plateau in the squiggle and then find the best consensus between subsequent k-mers, however currently used neural network applications are able to assign a single base per plateau.

1.2.2 Fluorescence sensing

Fluorescence-based SM methods, the second readout category, probe the presence and relative positions of fluorescent BP moieties or dyes by energetically exciting them through illumination and measuring the returned light. Restricting the observed sample volume to reduce background noise is of major importance in this category and several approaches have been developed to accomplish this. Both zero-mode waveguides (ZMWs) [29] and plasmonic nanostructures [30] restrict illumination to very small volumes, thus allowing the excitation and observation of single molecules. However, straight-forward total internal reflection fluoresence (TIRF) microscopy on surface-immobilized molecules remains the most popular choice in early development stages of SM methods. In a sparsely populated field of view containing thousands of molecules, the TIRF microscope's magnification is powerful enough to discern single fluorescent molecules. Innate fluorescence of BP moieties may be probed, however the signal of these is typically weak and limited to few of the building blocks of BPs. Most attention has been focused on the usage of fluorescent dyes ligated to "recognizers" – molecules that target discriminating molecular properties of BPs. Although the fluorescence no longer directly reports on properties of the BP itself, the choice of recognizer can attune fluorescence-based SM methods to properties of the BP which are difficult to detect otherwise. Through the use of various labeling schemes spanning one or more differently colored dyes and the behavior of dyes over time, this principle allows observation of a broad range of properties at SM resolution.

In a straightforward application the occurrence of certain features such as target monomers or structures in a single BP molecule may be visualized using recognizer-bound dyes. If multiples of single-dye fluorescence intensity can be measured, multiple occurrences of the same feature on a single BP can also be quantified [12]. By making use of two differently colored dyes that are capable of Förster resonance energy transfer (FRET), sub-nanometer-scale distances may be estimated accurately from the measured fluorescence intensity. This, in turn, may be used for a variety of purposes, such as to visualize SM kinetics in real-time [31] or to identify a protein by a set of signature intramolecular distances.

Similarly to electrical readouts, optical readouts over time too consist of squiggles, in which plateaus must be detected and assigned to BP features. Hence the same data analysis methods are popular for fluorescence data; HMMs see broad

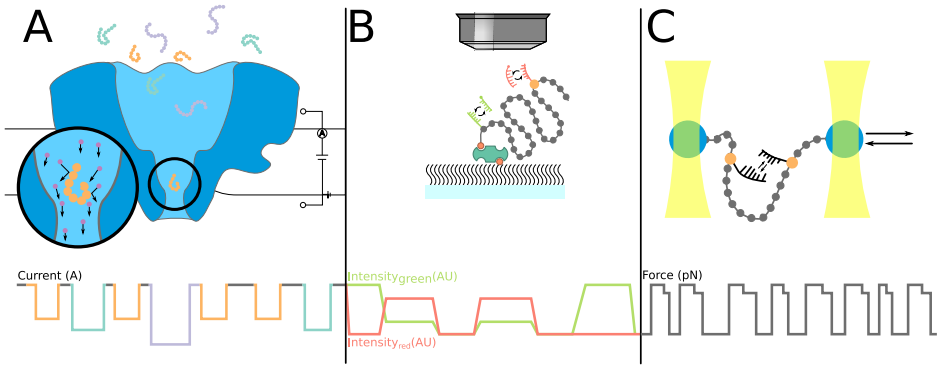


Figure 1.1: Three single-molecule readout approaches for analysis of biopolymers (BPs) and schematic example squiggles for each.**(A)** In electrical sensing, The effect of a BP or BP fragment on an electron current is measured to analyse them as single molecules. **(B)** In optical sensing, selected monomers are decorated with fluorescent dyes, of which emission intensity is read out using TIRF microscopy. **(C)** Finally, force-based sensing is done by decorating the BP with beads and physically manipulating them with tweezers (depicted are optical tweezers), while recording the extension length changes in response to the exerted force (as in [37]).

application [32–34], while neural networks have been used in select applications such as data filtering [35] and proof-of-concept simulation studies [13]. Again the widest usage for SM readout is found in NA sequencers. In single-molecule real-time (SMRT) sequencing, a TIRF setup is used to monitor a single polymerase as it copies a target NA strand using fluorescently labeled nucleotides. Each nucleotide is labeled with a differently colored dye, thus by noting the sequence of colors of incorporated nucleotides, the sequence of the target strand can be deduced. Sequencers produced by Pacific Biosciences (PacBio) make use of this principle and are popular due to their combination of high accuracy and long read length.

1.2.3 Mechanical force resistance

Thirdly, a BP may be probed by its resistance to mechanical force, as exerted by optical [36] or magnetic tweezers [37], or a centrifuge [38] for example. In such applications, beads are affixed to select points in an immobilized BP, which function as points of application of force. As a structure unfolds, its resistance to the applied force changes accordingly. As such, the force resistance profile provides information on the structure of the BP. If identification of the BP is the goal, this pattern may then be matched to known patterns. Although these patterns typically contain less information than electrical or optical readouts, several groups have demonstrated the usability of force-based measurements in SM protein identification [37] and structural analysis [36].

Lastly, several promising concepts have recently been proposed that rely on entirely different readouts, such as Dalton-level weight measurements using nano-electromechanical systems [39; 40] and Brownian motion behavior [41]. Although viable and promising in theory, they have not seen extensive application on BPs and are therefore not discussed in depth here.

1.3 SM protein analysis

1.3.1 Complications in protein analysis

Although each readout method can be adapted to interrogate different BP types at the SM level, it is clear that applications for NA analysis are the most mature compared to those for proteins. This difference is attributable to several differences between the two BP types. While NAs consist of four different monomers (i.e. nucleotides), each at least moderately distinct from the others, proteins are vastly more complex chains composed of twenty different monomers (i.e. amino acids), some of which are highly similar in structure. Moreover proteins are not uniformly charged like NAs, thus complicating manipulation of single molecules using electrical potentials, and they conform more strongly to a secondary structure, making linearization harder. Most importantly however, nature provides tools to process the sequence information in NA molecules, which can be repurposed for sequencing applications; virtually all existing NA sequencing methods make use of parts of this pre-existing machinery, such as helicases [42], polymerases [43] and Watson-Crick basepairing [44]. For proteins, which serve as machinery and building material rather than information storage, no such premade readout tools are available.

Considering these complications, it is understandable that the golden standard method for protein analysis is mass spectrometry (MS), a method which circumvents the usage of biological tools for its readout and reduces the complexity of proteins to a spectrum composed of mass-to-charge ratios (m/z). During its decades of dominance in the protein analysis field, MS has grown increasingly sensitive; proteins present in mere dozens of copies per sample can now be detected [45], while new databases [46] and interpretation methods [47] for MS spectra offer more certainty in the identification of proteins. Alas, it is unlikely that MS will ever be capable of single-molecule resolution in complex samples. Moreover, MS equipment will remain too unwieldy, costly and expert user-reliant for in-house analysis in small labs, most clinical settings and remote locations. As such there is a clear niche for SM protein analysis methods in biological research.

1.3.2 SM protein fingerprinting

Several promising methods targeting the niche of SM protein analysis are currently under development. Most of these deal with the complexity of proteins by reading out only a subset of the twenty proteinogenic residue types to construct a fingerprint, a signature that does not allow naive readout of the protein's composition but does contain sufficient information to identify it in a reference

database. The viability of fingerprinting approaches relies on the fact that only a very small fraction of all possible residue combinations is used in nature [48], with only a fraction of these occurring in a given context (e.g. a single organism or cell type). Therefore, the reduced information of a fingerprint combined with its context may be sufficient to recognize proteins, while additional information in protein databases or genome-derived gene predictions may even allow retrieval of the full sequence. Indeed computational simulations have shown that reduced sequences consisting of only three residue types are theoretically sufficient to uniquely identify almost all proteins in the human proteome [49].

In adapting electrical readout methods for protein fingerprinting, the lack of processive molecular motors as they exist for NAs and the variable charge of proteins are the most important hurdles. Lucas *et al.* [50] showed that charges inside protein pores may be tuned such that the charge-indiscriminate EOF dominates the EF, so that both positively and negatively charged peptides may be analysed at the same potential, albeit up to a certain charge. This proved sufficient for discrimination between a selection of short peptides, however usability for larger proteins has yet to be demonstrated. Both issues were also addressed effectively by Brinkhoff *et al.* by attaching a DNA tether to proteins; the tether is used to partially ratchet proteins through the pore using conventional DNA-processing motor proteins [51]. This allowed for the detection of single-residue differences in peptides of any charge, however the length of the peptide is limited by how far the tether can drag the peptide into the pore lumen before the motor protein has reached the end of the tether.

For fluorescence-based readout methods, it is the five-fold increase of monomer types that hampers the step from NA to protein analysis the most. For fluorescent dyes to remain discernible, their emission spectra need to be separable. Finding four such dyes, combining them with orthogonal labeling chemistries, and building four-color TIRF readout setups is already challenging, thus utilizing fingerprinting schemes to reduce the number of targeted residue types is an absolute necessity. Luckily, concept methods utilizing such reduced fingerprinting schemes have shown promise all the same. Swaminathan *et al.* developed an SM readout of the classic Edman degradation-based method for protein sequencing that allowed unique identification of 98.2% of the human proteome in simulations using a four-color readout, while two colors sufficed for prefractionated subsets of proteins [52].

1.4 The roles of bioinformatics in SM method development

1.4.1 Simulation

Over the past decade, computational biology and bioinformatics have played a major role in SM analysis method development for proteins. Proof-of-concept method papers are typically a combination of computational and experimental evidence; the experimental evidence supports the assumptions made in a computational model and the model indicates that a method should allow SM identification of a wide array of proteins. For example, Ohayon *et al.* performed physical simulations of sequential processing of proteins through a plasmonic

pore and the detection of three fluorescent dyes, to show that a neural network could classify 98% of proteins in the human proteome based on the simulated readout [13]. Although full *in vitro* sequence readouts have yet to be produced, the work did include experimental evidence showing that the simulated translocation speed is realistically achievable. Similarly, Hong *et al.* proposed that proteins can be translated into DNA barcodes *via* subsequently binding affinity agents, after which the barcodes can be sequenced to identify the protein [16]. Although their experimental data did not contain actual barcodes yet, simulation data already indicated the requirements that would need to be met by the affinity agents to identify the majority of human protein species. This demonstrates how bioinformatics may steer development of experimental methods by identifying which parameters must improve before a method, at least in theory, becomes able to analyse complex samples.

As with all computational simulations, it must be noted that those in the mentioned studies operate on incomplete information and partially supported assumptions. Fair warning is due for overly optimistic (or pessimistic) prospects as a result of necessary simplifications of reality in a model. As an experimental method is developed further, the model must be re-evaluated against new data and updated as necessary. To enable this, running the model should require as few computational resources as possible. Balancing required resources and model complexity is not trivial.

1.4.2 Data analysis

Equally important to the development of SM analysis methods are contributions from bioinformatics to data analysis. For a new protein analysis method to be adopted by users, it is imperative that it comes with easy to use and reliably accurate data analysis software tools.

As SM protein analysis is in an early state of development, simulated data are often used to develop analysis methods before the experimental method is able to produce the required quantities of data. Several of these investigations opt for a machine learning-based approach. This includes the previously mentioned plasmonic pore-based approach by Ohayon *et al.*, for which a feed-forward neural network was trained and evaluated [13]. Furthermore, Zhao *et al.* have demonstrated that a support vector machine can classify short peptides from fingerprints generated by electron tunneling measurements [53] and Brinkerhoff *et al.* showed that an HMM could classify electrical readouts of peptides [51]. Although not yet demonstrated, it is conceivable that fingerprints from electrical readout methods can be classified using other methods common in NA nanopore analysis as well, such as convolutional and recurrent neural networks [25; 26], and alignment forests [27]. Other SM method proof-of-concept studies have relied on the comparison of signatures to databases for identification. For instance, Egertson *et al.* presented a maximum likelihood-based method to find the most likely identity for simulated fingerprints from their affinity agent-based fingerprinting method [15], and Swaminathan *et al.* proposed a trie (i.e. a tree structure) to identify simulated fingerprints for their SM Edman degradation approach [52].

Upon widespread adoption of a method, bioinformatic contributions will need to be expanded even further, to benchmarking metrics and tools, as well as standardized quality measures akin to the PHRED score used broadly in NA sequencing [54; 55]. In this stage, large amounts of data will be generated, for which accessible databases will need to be constructed, which can be queried efficiently. This may require the development of new data representations, database structures and search methods, as is currently done for MS [47; 56; 46]

1.5 This thesis

SM fingerprinting and sequencing of NAs has already revolutionized biological research, with the coming years holding the promise of significant improvements on existing methods and the introduction of SM protein analysis. Bioinformatics can play a role in streamlining these future developments, by providing appropriate data analysis software for each stage of method development and *in silico* models that direct experimental design. In this thesis several examples of both these roles are presented.

First, **Chapter 2** delves deeper into electrical SM analysis of NAs and reviews software required for each step in the bioinformatic pipeline from sequencing device to high-quality genome assemble. Although the discussion of core principles in this chapter remains valid, the reader is advised to keep the time of writing (i.e. 2017) in mind for recommendations on specific analysis pipelines due to the fast pace of development in this field. The following three chapters describe tools that support SM analysis methods in stages of increasing maturity. The fluorescence data analysis tool FRETboard, the subject of **Chapter 3**, allows quick adaptation of classification algorithms to the variable experimental approaches encountered in early development. Once experimental design has been largely established, benchmarking of data analysis pipelines by users is of high importance to gauge the current state-of-art over iterative improvements. PoreTally, described in **Chapter 4**, streamlines benchmarking efforts and their publication to encourage this. For analysis methods in a further developed state, other considerations such as computational analysis cost and portability will play more significant roles. For instance, the baseLess tool described in **chapter 5** allows light-weight detection of a single NA sequence – e.g. in species detection – decreasing hardware cost and increasing portability of the analysis system by foregoing the full sequencing capability. The next two chapters focus on protein fingerprinting, in two hybrid *in silico*/laboratory feasibility studies of concept methods. Specifically, **chapter 6** considers a fluorescence-based method using FRET to measure intermolecular distances as a fingerprint. Similarly, **chapter 7** evaluates a nanopore-based approach, called chop-n-drop in which a protease-nanopore construct is used to lyse a target protein and interrogate the fragments. Finally, **chapter 8** discusses future prospects in the SM BP analysis field and provides recommendations to steer development toward practically applicable methods.

CHAPTER 2

The long reads ahead: *de novo* genome assembly using the MinION

This chapter has been published as:

Carlos de Lannoy, Dick de Ridder and Judith Risse. "The long reads ahead: *de novo* genome assembly using the MinION" *F1000* 6 (2017): 1083

Abstract

Nanopore technology provides a novel approach to DNA sequencing that yields long, label-free reads of constant quality. The first commercial implementation of this approach, the MinION, has shown promise in various sequencing applications. This review gives an up-to-date overview of the MinION's utility as a *de novo* sequencing device. It is argued that the MinION may allow for portable and affordable *de novo* sequencing of even complex genomes in the near future, despite the currently error-prone nature of its reads. Through continuous updates to the MinION hardware and the development of new assembly pipelines, both sequencing accuracy and assembly quality have already risen rapidly. However, this fast pace of development has also lead to a lack of overview of the expanding landscape of analysis tools, as performance evaluations are outdated quickly. As the MinION is approaching a state of maturity, its user community would benefit from a thorough comparative benchmarking effort of de novo assembly pipelines in the near future.

2.1 Introduction

The development of novel genome sequencing methods has been a major driving force behind the rapid advancements in genomics of the last decades. Notably, the advent of second generation sequencing (SGS) provided researchers with the required throughput and cost-efficiency to sequence many more genomes than was previously deemed feasible. Recent years saw the dawn of what can be considered a third generation; one that allows amplification-free reading of single DNA molecules in long consecutive stretches [57]. Currently, this new generation is dominated by two methods: nanopore sequencing and single-molecule real time (SMRT) sequencing, championed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), respectively.

Conceptually, nanopore sequencing is easier to explain than most other sequencing methods. An electrical potential is applied across an insulating membrane in which a single small pore is inserted. A DNA strand is pulled through the pore and the sequence is inferred from the characteristic way in which the passing base combinations influence the current. In 1989, David Deamer roughly sketched this concept as it is applied today, although it took more than two decades of key innovations to bring the concept to fruition [58]. Since the introduction of the first commercially available nanopore sequencing device, ONT's MinION, and the start of the MinION Access program (MAP) in 2014, the field of nanopore sequencing has been advancing at a rapid pace; both new applications and improvements to existing ones are published on a regular basis.

The advantages of the MinION over other sequencing devices are numerous. Both its size, roughly that of a cellphone, and its initial investment cost, a thousand dollars for a starter kit, are a mere fraction of that of competitors. Running the MinION is also reasonably time- and cost-effective; a 48-hour sequencing run

currently costs around 800 dollars[1] and yields up to 5 Gbases of raw sequenced data [59]. Furthermore, the technique does not rely on any labeling techniques to recognize different bases, while Sanger, second generation and SMRT sequencing methods do require some form of labeling of nucleotides. Amplification by PCR is optional for the MinION, while this step is mandatory for Sanger and SGS-methods. Not only does omitting these steps simplify sample preparation for MinION samples, it also helps to avoid errors and biases (e.g. the CG-bias for PCR) and allows detection of modified bases [60]. Finally, the maximum read length produced by the MinION is many times greater than that of both second-generation and Sanger sequencing and only paralleled by SMRT sequencing, which is highly advantageous in resolving repeat sequences.

The most prominent disadvantages of the MinION, with respect to its competitors, are the lower signal-to-noise ratio, stochasticity introduced by its biological components, and the resulting high error rate of basecalling. Indeed, the MinION is a product in development and the used materials (i.e. membranes, nanopores and buffers) are still being optimized. Furthermore, it is thought that significant improvements are still possible in the software pipelines that translate current signal to DNA sequence.

In this review, an up-to-date overview of *de novo* nanopore sequencing and assembly is provided. First, the physical sequencing process as it takes place inside the MinION is outlined. Then, the general structure of analysis pipelines is described, along with currently available software implemented in these pipelines and their respective strengths and weaknesses. It should be noted that nanopore sequencing is a rapidly advancing field. While some work discussed in this paper is considered cutting-edge at the moment of writing, the reader is advised to keep the publication date of said work in mind.

2.2 Physical basis of DNA sequencing using nanopores

The underlying principle of nanopore sequencing can be explained as follows: a microscopic opening wide enough to allow single-stranded DNA to pass - the nanopore - is introduced in an insulating membrane between two compartments filled with saline solution and an electric potential is applied across it. DNA strands are then added to one compartment and allowed to diffuse toward the nanopore, where they are captured by the electric field and threaded through the pore. While a strand is passed through, the characteristic way in which the bases influence the electric current through the nanopore is measured. These measurements can then be decoded to retrieve the sequence of the DNA strand (Figure 2.1).

In recent years, several key discoveries rapidly transformed nanopore sequencing into a usable DNA analysis method. In a step-by-step exploration of the sequencing process, these discoveries will be discussed next.

[1] Estimate based on a purchase of 24 flowcells and a $1D/1D^2$ sequencing kit, 13th of October 2017

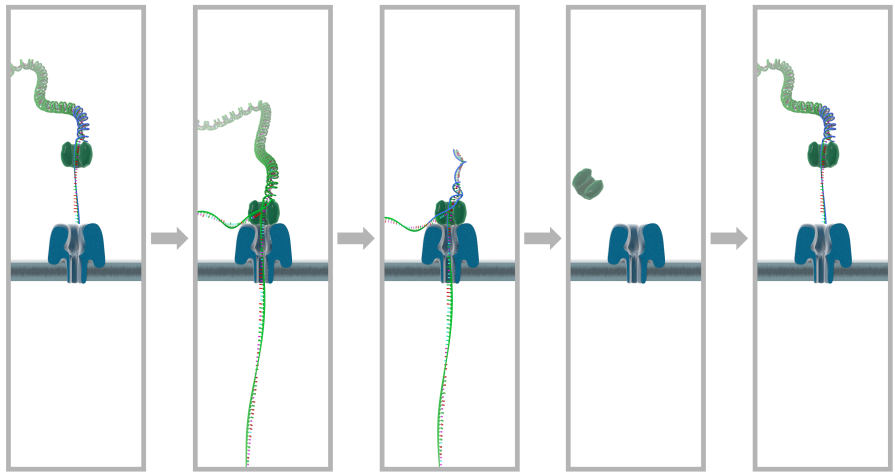


Figure 2.1: Sequencing of a DNA strand using nanopores. From left to right, double-stranded DNA with attached motor protein attaches to a pore protein in an insulating membrane. The applied potential pulls one strand through the pore, while the motor protein unzips the DNA in a step-wise fashion. After the DNA has been unzipped completely and one strand has passed through, the complex detaches from the pore entrance and the pore is ready to receive another strand. Image courtesy of Oxford Nanopore Technologies Ltd.

2.2.1 Choice of pore: Biological versus solid-state

Nanopore sequencing efforts are sub-categorized in two groups based on the choice of nanopore. Most current efforts implement biological nanopores, which are protein multimers derived from naturally occurring counterparts. Through genetic engineering, biological nanopores are modifiable in terms of dimensions and placement of electrical charge. These properties are also highly reproducible from one pore to the next. Functionality can be further modified by attaching compatible enzymes to the pore opening. Like their naturally occurring counterparts however, they need to be embedded in a lipid membrane, which is generally prone to disruption, particularly when exposed to varying electrical potentials. In the MinION, this was partly solved by constructing membranes out of a more stable single layer of polymers, rather than the traditional bilayer. Solid-state nanopores on the other hand, are made by burning openings in a synthetic membrane using a focused electron or ion beam [61]. Contrary to biological nanopores, solid-state nanopores are compatible with a wide range of strong and chemically stable materials with equally diverse properties. Pores are also more easily parallelized and integrated in electrical readout circuits. A major disadvantage at the moment is the irreproducibility of the pore dimensions. They also do not combine as easily with modifying enzymes. As a result, solid-state nanopores currently produce noisier and less easily interpretable signals than biological nanopores. In

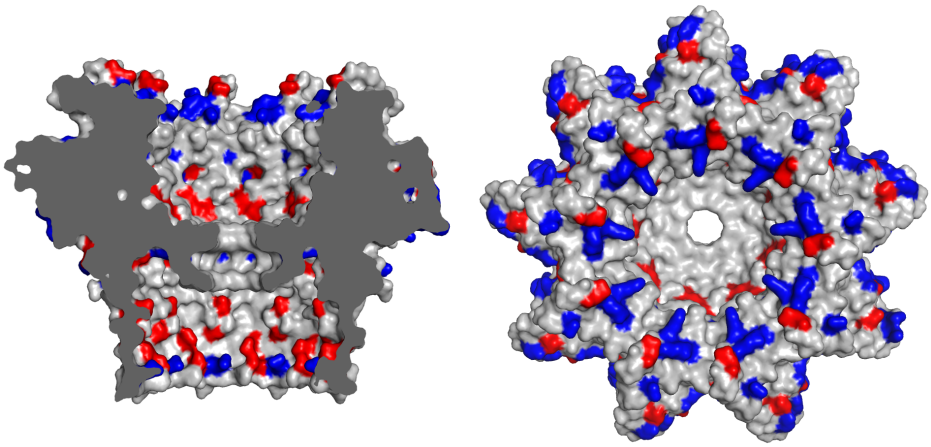


Figure 2.2: Protein structure of the CsgG pore protein complex, a variant of which is used in current generation MinION flow cells. Positive and negative residues are colored blue and red, respectively. Image generated by the authors using PyMOL v1.7.0.0. PDB ID: 4UV3 [62].

the following, the focus will lie on biological nanopore sequencing and the term nanopore will refer to the biological kind.

2.2.2 Structure and charge of the nanopore

One important structural property that makes a biological pore suitable for DNA sequencing is a constriction site at which the passing strand exerts the most influence on the electrical current. The length of the constricting passage largely determines how many bases simultaneously influence the electrical current and thus the number of bases that is "read" simultaneously at a given time. This number should be kept low enough to allow recognition of a signature current for each different combination of bases and high enough to allow for some overlap between subsequent base combinations, as this benefits basecalling accuracy by allowing every base to be read multiple times. Modified versions of both pore proteins that have seen application in the MinION, MspA (denoted by ONT with series numbers prefix "R7") and the currently used CsgG [62] (denoted with prefix "R9", Figure 2.2), have a constricted passage that allows detection of a manageable number of bases. For the 10Å-long constriction of the CsgG pore, basecalling models previously relied on the assumption that five nucleotides sufficiently influence the current at any given time to discern all different nucleotide combinations, and thus 5-mers were assigned to stretches of signal (Figure 2.3). Although this worked reasonably well, it was found that this assumption does not always hold, e.g. due to specific base sequences and the secondary structure of the molecule influencing the current differently. Newer basecalling models therefore

no longer make this assumption and assign a variable number of bases (see also section 2.3.1).

For sequencing to commence, a DNA strand first needs to diffuse towards one side of the pore, referred to as the cis-side, where it is captured by the electric field resulting from the applied potential. It is then threaded through the pore and extruded at the other end, called the trans-side.

Two forces should be considered. First and most importantly, the electrophoretic force induced by a positive electric potential applied at the trans-side attracts the negatively charged DNA and pulls it in. As negative particles leave the cis-side and positive particles simultaneously move in the opposite direction, a positively charged zone forms around the cis entrance of the pore, strengthening attraction of DNA strands. Secondly, strand translocation is influenced by the electro-osmotic flow (EOF), the force induced by the net water and ion flow through the pore. While a DNA strand is in the pore, the EOF normally opposes the direction of the electrophoretic force and thus of translocation; however, this effect is relatively minor.

Through iterative optimization of internal architecture, it was found that positive internal surface charges are important for efficient DNA capture [63; 64], while base recognition was found to improve with bulky or hydrophobic amino acid side chains placed at the constriction site, as these direct ion flow toward the DNA strand [65]. Although the structures of the modified pores used in MinION flow cells have not been publicly released by ONT, modifications to these properties have likely been made. Currently, ONT maintains two types of flow cells containing different modified CsgG pores [62], designated R9.4 and R9.5. Reportedly, alterations between R9.4 and its successor R9.5 were solely made to facilitate a novel sequencing mode (dubbed $1D^2$, see below) and should not influence sequencing accuracy in any other way. These alterations thus likely pertain to different properties of the pore.

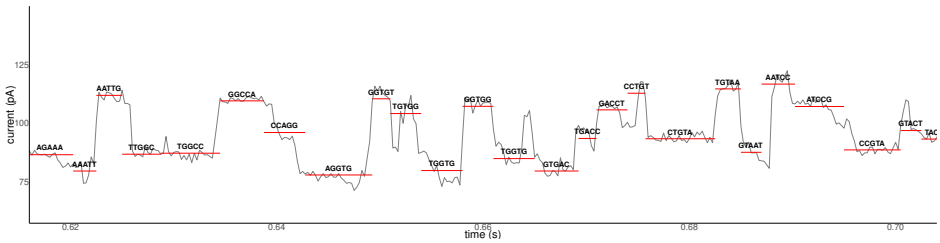


Figure 2.3: Example of a MinION DNA read as raw data (grey line) and the event data (red lines) extracted from it, corresponding to discrete sets of bases. For the sake of illustration it is assumed that five bases influence the current at a given time, although in reality this assumption may not always hold. Data used in this figure was obtained from the Nanopore WGS consortium (third release)[59].

2.2.3 Processive control

It should be noted that the processive speed of the strand without any further modifications is too high for the sensor to accurately detect changes in electrical current (between $2 \cdot 10^6$ and $10 \cdot 10^6$ bases/s in wild-type MspA) [63]. Currently, the most successful way to exert control over the speed has proven to be the addition of a motor protein, such as phi29 DNA polymerase [43] or a helicase [42]. In a preparatory step, poly-T or "leader" adapters are attached to the double-stranded DNA. Motor proteins attach to these adapters, but due to specialized bases in the adapter sequence (possibly acridine residues as used by [66], but left unspecified by ONT [42]), they cannot unzip it at this stage. Once one end of the complex is adjacent to the cis-side of the pore, the leader adapter previously blocking the motor protein is released, presumably due to the force exerted on the strand as demonstrated by [67] and described in [68]. The DNA is then fed base-by-base through the pore by the motor protein as it processes the strand, where it can now be read at a regular pace. A modified helicase is currently used as motor protein in the MinION [42]. The latest release of this motor protein at the time of writing (dubbed E8) maintains an average throughput speed of 450 bases/s (as noted in e.g. [59]).

2.2.4 Reading the DNA strand

During a MinION sequencing run, the potential over the membrane is kept stable, while the electrical current (in the pA-range) is sampled at a frequency in the kHz range (Figure 2.3). This signal is characteristic for the subsequent bases moving through the pore and will ultimately serve as the basis for basecalling. As the amount of electrolyte is increasingly depleted during the run, the applied potential (typically starting at -180mV) is further decreased by 5mV per two hours of runtime and increased by 5mV when the MinION switches to another set of wells filled with fresher buffer (see next section).

While the MinION can read the first strand of a dsDNA-stretch that is threaded through the pore - by definition, the template strand - and discard the complementary strand, it is possible to instead read the complementary strand immediately after the template, thus performing a second read of the same stretch in reverse complement (Figure 2.4). Combining reads of both strands has been shown to increase sequencing accuracy significantly [70]. The currently implemented method for doing so is referred to as $1D^2$ sequencing (versus 1D sequencing if only the template strand is read). The $1D^2$ chemistry provided by ONT includes different adapters that allow the complement strand to attach to the membrane while the template strand is read. Shortly after the template strand has completely left the pore, the complement strand is pulled in and sequenced. The mirrored reads are then decoded jointly so that any sequencing errors may be corrected. A previously offered method with the same aim, referred to as 2D-sequencing, involved covalently connecting the 3'-end of the template and the 5'-end of its complement using an abasic hairpin adapter, thus allowing the complement strand to be pulled in automatically after the template strand. However, due to several issues, including the hairpin's tendency to ligate different

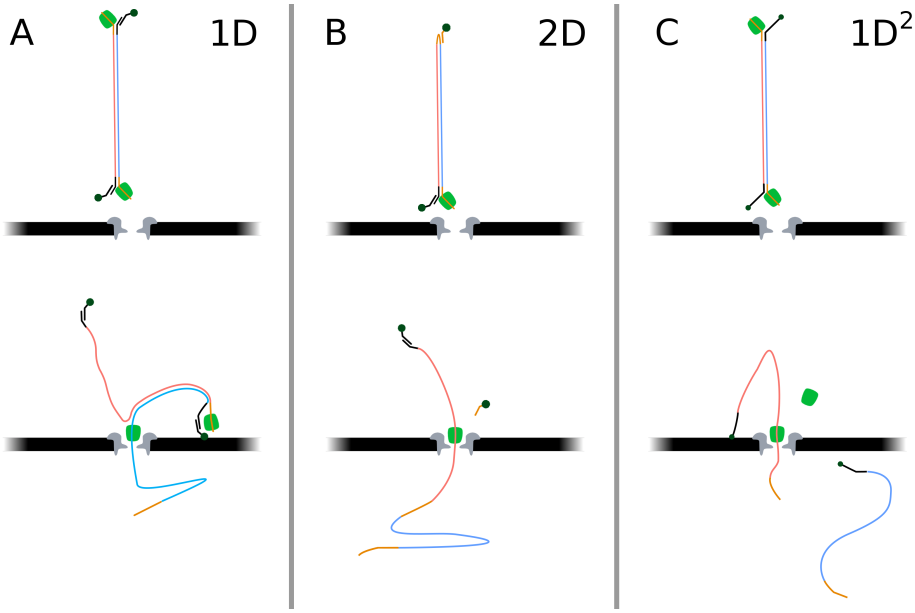


Figure 2.4: The three categories of DNA reading chemistries for the MinION. **(A)** When using 1D chemistry, only the template strand (blue) is threaded by its motor protein (green) and read. The complement strand (red) is discarded at the cis side of the pore. The tethers (dark-green) allow for selection of properly ligated complexes during sample preparation and attach to the membrane to increase the availability of strands near pores during sequencing. **(B)** The now-deprecated 2D chemistry connected template and complement strand using a hairpin, thus allowing sequencing of the complement strand immediately after the template strand. An additional tether that attached to the hairpin allowed for selection of correctly ligated strands during sample preparation. **(C)** $1D^2$ chemistry, the successor of 2D, also allows sequencing of both strands, but rather than attaching the two, the complement strand is tethered to the membrane while the template is sequenced. After the template strand is threaded through, the complement strand is drawn in and the tether is pulled loose. Based on [69] by permission from Macmillan Publishers Ltd: Nature Methods, copyright(2015), the ONT kit content description, and ONT's technical update of March 2017.

strands into chimeric reads [71] and a lower read quality and sequencing speed for the complement strand [70] reportedly caused by secondary structure changes in the strand while rezipping after sequencing, this approach was deprecated in favor of $1D^2$-sequencing in May of 2017.

2.2.5 Channel parallelization

Lastly, throughput can be greatly increased by reading the signal from multiple pores in parallel. The current generation of the MinION's disposable cartridges, called flow cells, can read the signal of up to 512 pores in parallel (Figure 2.5). The flow cell is equipped with 2048 wells, which are connected in groups of four to multiplexers (MUXs), the switches that control which of the four cells per group is controlled and read out by the circuits. During the initial platform quality check, DNA strands (of unreleased source and sequence), present in the buffer with which the flow cells are shipped, are sequenced to discern wells suitable for sequencing (i.e. containing an intact membrane and precisely one correctly inserted, properly functioning pore) from wells in which correct pore insertion has failed (see ONT platform quality check explanation). The latter scenario may occur, as the insertion of pores is a stochastic process. In a second quality check, the MUX scan, each MUX chooses up to three wells in order of signal quality and begins readout in the best-quality well. As well quality is expected to decline during the run, the standard protocol switches to the second-best quality pore after eight hours, and the third-best quality after another eight hours. This way, the best and most output is expected in the first part of the run. While a run using a group of wells is in progress, the circuits connected to the MUXs regulate the current in each selected well individually. This also allows expelling of eventual blockades from a pore, by temporarily reversing the current in the affected well while the rest of the wells continue to function normally.

2.3 Currently available software for MinION basecalling and *de novo* assembly

Following the process in section 2.2, a current signal is obtained that is subsequently translated into the underlying DNA sequence by a so-called basecaller. Next, the read sequences may be *de novo* assembled using assembly tools that can make use of the long read length while mitigating the error-prone nature of the reads. This is often followed by a last error correction or 'polishing' step, in which a better consensus between the assembly and the raw reads is sought. In this section, these steps are detailed and a selection of available software tools to fulfill each step is explored.

2.3.1 Basecallers

Before basecalling takes place, some preparatory steps may be required. First, if the (now deprecated) 2D chemistry was used, the signal derived from the

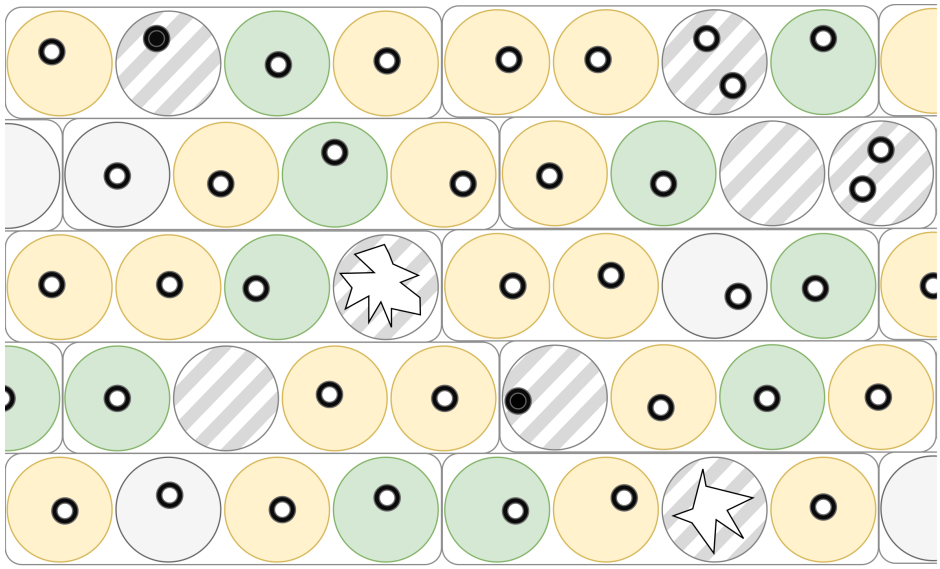


Figure 2.5: Layout of a MinION flowcell grid. Large circles denote wells in the grid, small black circles denote inserted nanopores. In reality, the pore diameter (12 nm) is much smaller with respect to the well diameter (about 10 μm). Each group of four wells is controlled by a multiplexer (MUX). During an initial quality check, wells that are unusable e.g. due to erroneous pore insertion, membrane defects or pore blockades are marked as unusable (hatch pattern). Right before sequencing, the wells are tested a second time and three wells per MUX are ranked on signal quality (if possible). Sequencing of the sample will then commence, starting read-out from the best-performing well (green) and switching to second and third best (yellow) after eight hours each. The white wells are usable for sequencing, but are left unused unless the user designates otherwise in the MinION protocol.

template strand should be separated from that of the hairpin and the complement strand. This process is commonly referred to as segmentation. Furthermore, older basecallers require the signal to be subdivided into discrete averaged stretches, or events, each corresponding to a particular set of k bases. Both segmentation and event detection can be performed by MinKNOW, the MinION control software provided by ONT. For event detection, MinKNOW was reported to calculate a simple t-statistic between sliding adjacent windows of set size. Peaks in the t-statistic above a certain threshold are then assumed to signify the borders between adjacent events.

Initially, basecallers were designed to find the most likely set of k bases for each event detected in this manner [23; 24]. As it became clear that the number of bases per event is too variable for this approach, newer tools generally infer the events and the underlying sequence simultaneously from the raw signal (e.g. Albacore v$\geq$2.0.1, Chiron, BasecRAWller) [72; 25].

To assess the quality of basecalling performance, a 3.6 kbase calibration strand derived from the Lambda genome may be added to the sample [73; 70]. MinKNOW automatically detects reads derived from the Lambda genome and separates those from the sample reads. Software tools may also use these strands for parameter optimization (e.g. as PoreSeq does to adjust its basecall correction algorithm [73]).

Several dedicated basecalling tools are available to MinION users. In this section, the underlying principles and implementation of these tools are explored, along with their reported strengths and weaknesses. Unfortunately, most basecallers do not support calling $1D^2$ reads, thus performance measures will focus on 1D calling. Wick *et al.* have provided a benchmark for basecallers on a 1D, R9.4 *Kleibsiella pneumoniae* dataset generated with a SQK-LSK108 chemistry kit. To the author's knowledge, this is currently the only comprehensive and up-to-date benchmarking effort. Comparisons made in this section are based on their analysis and reports made by the authors of the open-source basecallers in their publications. In the latter case, the used read type, pore and chemistry kit is listed between brackets each time, e.g. for the Wick *et al.* study: (1D, R9.4, SQK-LSK108).

Metrichor basecallers Metrichor, a spin-off company of ONT and its main developer of proprietary analysis software, maintains a range of basecallers that have remained the go-to option for most MinION users. Currently, four Metrichor basecallers are available to users: Albacore, the MinKNOW integrated basecaller, Nanonet and Scrappie. A cloud-based version was previously integrated in the EPI2ME platform, but this service has been discontinued. Both Nanonet and Scrappie are unsupported development basecallers, while Albacore and the MinKNOW version are stable tools intended for regular MinION users.

Initially, the Metrichor basecallers relied on hidden Markov models (HMMs) to assign k-mers of set size k to event-called data. As of early 2016, the HMM model was replaced by a more accurate recurrent neural network (RNN)-implementation. This approach was first introduced in Nanonet (source code publicly available), a basecaller written in Python and using the CURRENNT library [74] to implement its RNN. It is able to perform all steps from raw MinION signal to base sequence (i.e. segmentation, event-calling and basecalling). The next major advancement was the addition of a transducer after the RNN in April of 2017 which, rather than assigning a k-mer to each event, uses the newly input signal and the bases it previously emitted to determine whether to output none, one or multiple bases for the next event. Importantly, this allowed the detection of homopolymer sequences longer than a given k-mer size [59]. This was previously impossible, as the sliding window t-test used in event detection could not discern individual events in homopolymer stretches, effectively merging them into a single event which would then be assigned a single k-mer [73; 75; 76].

From June of 2017, event-based calling was abandoned all together in favor of a more accurate raw signal-based approach. Both the transducer and raw signal-based calling were first introduced as options in Scrappie, a newer developer basecaller written in C (source code publicly available), and were later implemented in Albacore (transducer as of v1.0.1, raw signal interpretation as of v2.0.1). To

date, Albacore also remains the only basecaller able to make use of $1D^2$ reads. The MinKNOW basecaller lags slightly behind Albacore but is otherwise identical. The source code of Albacore and the MinKNOW basecaller is currently only open to developer users.

Metrichor's up-to-date basecaller implementations (i.e. Albacore, MinKNOW and Scrappie) first center and scale the raw signal using the median signal over the entire read (as first described in [77]) and then consecutively feed it through a strided convolutional filter and unidirectional RNN layers of gated recurrent units (GRUs) which receive their memory from alternating directions. The stacked unidirectional layers and use of GRUs allows the RNN to interpret the convolved signal in a long-range context from both sides, while remaining computationally efficient in use. The output of the RNN is fed into a transducer, which assigns a number of bases to each raw data point as described above [2]. Lastly, the translocation speed of the strand is estimated using found non-homopolymeric events, which is then used to detect and correct probable collapsed homopolymer sequences.

The processing speed [25] and accuracy of Albacore, MinKNOW and Scrappie is currently considered to be the highest of all available basecallers. Wick *et al.* estimated median identity with the reference genome of a transducer-based raw signal-processing Albacore version (v2.0.2) at 87.6%. The introduction of raw data interpretation lead to some increase in accuracy; Albacore v2.0.1 scored 87.6% identity versus 86.5% for v1.2.6 (the last version without raw calling included by Wick *et al.*) and a similar difference was seen between Scrappie v1.1.1 processing event-called (85.8% identity) and raw data (88.1%). The effect of the introduction of the transducer at v1.0.1 can be seen in the read length, which is closer to the reference read length, and the higher corrected assembly identity, which indicates that fewer systematic errors are made. Both observations can be explained by the fact that the transducer allows for more accurate calls in homopolymer regions in particular, as was also shown by [59]. As expected, the outdated Nanonet (v2.0.0) does not perform as well as Scrappie and Albacore (85.6% identity). Albacore's median identity rate on $1D^2$ reads has been reported by ONT at around 97%, however this has yet to be confirmed by thorough independent studies.

Chiron Chiron [25] is a third-party basecaller that shows high similarity with current Metrichor basecallers. It was written in Python and its neural network is implemented using the TensorFlow library [81].

Chiron first centers the raw signal around the mean and scales it over the standard deviation, after which the signal is divided up in partly overlapping batches to allow parallel processing. Much like current Metrichor basecallers, it then feeds the signal through a convolutional filter, several RNN-layers and a transducer which outputs probabilities for each base (or the absence of a base) for each raw data point. Finally, the returned base sequences for the split signal are fused into a single sequence for the entire read by finding the largest overlap.

[2] A thorough discussion of neural network architectures and their respective properties is outside the scope of this article. Interested readers are referred to [78] and [79] for introductions to RNNs and convolutional networks respectively, and [80] for more information on transducers.

Although Chiron's overall structure is similar to that of Metrichor basecallers, its multiple convolutional layers, the usage of the more elaborate long short-term memory (LSTM) cells instead of GRUs and the more conventional bidirectional RNN architecture make Chiron more complex. Indeed, the benchmark published by Chiron's authors shows that it performs slightly slower than Albacore v1.1.1 but similarly in terms of accuracy; on reads of lambda phage DNA, *E. coli* and *Mycobacterium tubercolosis* (all 1D, R9.4, SQK-LSK108), the difference between sequence identities of Albacore and Chiron did not rise above 1.2%. Albacore did do slightly better than Chiron on a human dataset generated with the same chemistry; Chiron's authors hypothesize that this could be because Chiron was not trained on human data. These results are largely in line with the benchmark by Wick *et al.*; indeed Chiron (v0.2) performs similarly to Albacore v1.1.2, but the raw data-based Albacore v2.0.2 performs notably better. In terms of sequencing speed, Chiron's authors showed that Albacore (2975 bases per second on a CPU) easily outperformed Chiron (21 bases per second on a CPU, 1652 on a GPU).

BasecRAWller While other basecallers prioritize accuracy, BasecRAWller's [72] primary goal is to allow "streaming basecalling", i.e. basecalling during sequencing directly from the raw signal. As its authors note, streaming basecalling may prove highly advantageous in selected applications, such as rejection of strands from the pore during sequencing if, based on the retrieved base sequence, it is decided that the strand is not of interest to the user. BasecRAWller is written in Python and uses the TensorFlow library [81] for its neural network implementation.

Like Metrichor basecallers, BasecRAWller uses a median-based normalization method [77] to pre-process the raw signal. However, as the median of the signal of the entire strand (as used by Metrichor) is not available in streaming basecalling, it is approximated by using the median unoccupied pore signal, as these values were found to correlate sufficiently. The normalized signal is then consecutively fed into a unidirectional LSTM-RNN and a fully connected feed-forward network, which assigns a 4-mer to each measurement and a probability that the measurement should be recognized as the start of a new event. This information is ultimately passed on to another unidirectional LSTM-RNN which assigns zero, one or multiple bases to each event. Although bidirectional RNNs have the advantage of utilizing both past and future measurements to place a prediction in a proper context, the choice for a unidirectional network was consciously made to retain the ability to basecall in a streaming fashion.

As its authors state in their own assessment of BasecRAWller's performance, some accuracy was surrendered to allow for streaming basecalling; Metrichor basecallers reached significantly higher accuracy on both an *E. coli* dataset (1D, R9, SQK-NSK007) and a human dataset (2D, R9.4, SQK-LSK108) (89.4% and 76% respectively, versus 82.9% and 72.5% for BasecRAWller). It should be noted that Albacore was able to take advantage of the 2D chemistry used for the human dataset, while BasecRAWller could not. Similarly, Wick *et al.* found a median identity of 74.0% for BaseCrawller (v0.1) versus Albacore's (v2.0.2) 87.6%. An assessment by Teng *et al.* found slightly higher identity rates for BasecRAWller (v0.1) of around 82% on Lambda phage *E. coli*, *M. tuberculosis* and human

datasets (all 1D, R9.4, SQK-LSK108), which were still 2% lower than that of Albacore (v1.1.1) on human data and around 8% lower for the other datasets [25]. BasecRAWller's authors indicate a processing speed of up to 900 bases per second using the current MinION throughput speed and sampling frequency, while Teng *et al.* indicated a maximum sequencing speed of 81 bases per second [25]. The cause of this large difference is unclear, but important to investigate further, as a speed below 450 bases per second (the current average throughput speed of the MinION) would indicate that BasecRAWller is currently not able to function as a true streaming basecaller.

2.3.2 Assemblers

Once nanopore reads have been basecalled, they may serve several purposes. If SGS reads are available, one of several approaches to hybrid assembly (i.e. combining long error-prone and short accurate reads) may be chosen; short reads may be mapped to the nanopore reads to correct sequencing errors pre-assembly [82] or to create large low-error contigs. The latter goal may be achieved by using nanopore reads to close gaps and resolve repeat regions in SGS assemblies [83], by using them as scaffolds to properly align short reads [84–86], by correcting a long read-only assembly using short reads [87; 88] (referred to as "polishing", see also next section), or by creating short accurate seed regions from short reads, which are then bridged by nanopore reads [76]. All described approaches were shown to result in accurate and highly contiguous *de novo* assemblies and in identification of repeats that were collapsed in SGS-only assemblies [82; 76].

If no SGS reads are available, nanopore-only assembly pipelines can be used. It has been shown that using these pipelines, a cheap and highly contiguous MinION-only *de novo* draft genome can already be sequenced and assembled within one week (e.g. as was done for the 54 Mbase fungal genome of *Rhizoctonia solani* [89]). If speed, cost or only the general structure of the genome are of major importance, a MinION-only approach may thus already be adequate. However, it should be noted that MinION-only assemblies are still generally inferior to those of hybrid methods in terms of accuracy, due to the error-prone nature of the reads [90; 91]. If the goal is the construction of a highly accurate and contiguous assembly and SGS reads can be obtained, hybrid assemblies should be preferred. This accuracy gap is expected to diminish in the future due to the steadily increasing quality of MinION reads. With this and the cost- and time-effectiveness of the MinION in mind, the focus of this review lies on tools that can be used in *de novo* MinION-only sequencing.

As PacBio sequencers were available before nanopore sequencing had come to fruition, most assemblers able to work with MinION reads were either initially intended as PacBio tools or were written with both technologies in mind. Some tools offer specific parameter settings to account for differences in read properties between the two technologies, most importantly the differing error distributions. Giordano *et al.* showed that, on datasets of comparable size and read length distribution, assemblers consistently constructed more accurate assemblies with SMRT reads than with MinION reads (although the latter were generated with

older chemistries and basecallers, see also Table 2.1) [90]. While the difference in accuracy is in large part attributable to the higher number and less random distribution of sequencing errors, it does seem that those adapted for use with MinION reads are better able to mitigate its sequencing errors.

Assembly of MinION and SMRT reads requires a different approach than that of SGS reads; as the reads are longer, finding a correct overlap should be easier, yet they are more error-prone, which increases the uncertainty of overlaps. Because of these differences, a return of interest in overlap-layout-consensus (OLC) algorithms - which were at the peak of their popularity in the era of Sanger sequencing - is seen. Traditional De-Bruijn graph (DBG) assemblers, the more popular choice for SGS reads, were reported to return lower quality assemblies of MinION reads than OLC-based methods, but proved faster in some cases [92]. A selection of available long read OLC and DBG assemblers is discussed in this section.

Software using traditional greedy extension algorithms (e.g. SSAKE) is rarely used in MinION read assembly as it was found to perform decidedly less well in a *de novo* assembly setting, both in terms of assembly quality and required computational resources [92], and is therefore not further discussed here. Furthermore, only tools that provide a full solution to their respective step in the assembly pipeline are reported here. As current assemblers either include their own error correction module [93] or work with uncorrected reads [76; 94–96], stand-alone pre-assembly error correction tools are excluded as well. A short summary of each assembler's characteristics and the limited number of available benchmarks is given in Table 2.1, although it should be noted that a proper evaluation is difficult due to the different and outdated chemistries and basecallers used. Thus, while performances noted here may provide an initial orientation in the available choice in long read assemblers, results are likely to differ when using current technology.

PBcR & Canu Originally developed for the first human genome draft, the Celera assembly pipeline [99] and its extensions [100; 101; 93] have remained a popular choice in a growing landscape of OLC assemblers. Briefly, the Celera assembler uses read overlaps to find contigs of which the structure can unambiguously be derived from overlap information, referred to as unitigs. It then separates unitigs that were found to occur multiple times from unique ones and attempts to orient the unique unitigs with respect to eachother. Where possible, gaps between unique unitigs are filled with non-unique unitigs. As a high read error rate is detrimental to the quality of the assembly [102], two different modifications to the pipeline are available. The PacBio corrected Reads (PBcR) algorithm, originally developed for the correction of PacBio reads suffering from similar error rates, uses accurate short reads mapped with high confidence to the long reads to correct errors. The assembly then proceeds as usual by Celera [98]. Celera's successor, Canu [93], provides a more accurate solution that does not require short accurate reads. Like PBcR, Canu was shown to succesfully assemble both MinION and PacBio reads [90]. The pipeline includes three stages; correction, trimming and assembly. Overlaps are found using the efficient minhash alignment process (MHAP) [103], which hashes k-mers using different hash functions and for each hash function stores the smallest integer to which a k-mer of the sequence

A	Judge *et al.*[97]			Istace *et al.*[91]			Giordano *et al.*[90]		
	subs/kbase	indels/kbase	N50 (Mbase)	subs/kbase	indels/kbase	N50 (Mbase)	subs/kbase	indels/kbase	N50 (Mbase)
PBcR	1.0	12.2	1.20				0.2	17	0.616
Canu	**0.3**	**7.8**	**2.80**	**0.105**	**10.0**	0.610	**0.1**	17	0.698
SMARTdenovo				0.580	11.1	0.783	0.3	**14**	0.625
Minimap & miniasm	6.7	18.6	6.60	0.207[1]	13.5[1]	0.736[1]	34	67	0.739
ABruijn				0.130	10.1	**0.816**	0.1	15	**0.769**
Chemistry		MAP006			MAP005/MAP006			MAP006/007	
Read type		2D			2D			2D	
Pore		R7.3			R7.3			R7.3/R9	
Basecaller		EPI2ME			EPI2ME			EPI2ME	
Organism		*Enterobacter kobei*			*S. cerevisiae*			*S. cerevisiae*	

B	Description	Ref.
PBcR	Celera OLC assembler adapted for long error-prone reads.	[98]
Canu	The more accurate successor of PBcR.	[93]
SMARTdenovo	Fast and reasonably accurate assembler without prior error correction step.	Github
Minimap & miniasm	Fast assembly pipeline without error correction and consensus steps.	[96]
ABruijn	DBG assembler that fuses unique strings prior to assembly, produces highly contiguous assemblies.	[95]
TULIP	uses seed extension principle to efficiently assemble large genomes.	[76]
HINGE	Assesses coverage of low complexity regions prior to assembly and processes them more efficiently.	[94]

Table 2.1: Summary of comparisons between long read assemblers. (**A**) Selected metrics for three benchmarking efforts on MinION reads, including chemistries used in the respective studies. Bold values denote the best score per metric. (**B**) Short descriptions and reference papers for all assemblers discussed in this paper.
[1]: reads were corrected by Canu prior to assembly.

is hashed. Comparing the hashed k-mers per read results in initial overlap hits, which are then used to perform error correction by consensus seeking. By selecting overlaps for correction on quality, but limiting the number of overlaps a read can contribute to, Canu attempts to prevent masking of true repeat variants. Shorter reads are used at this stage to improve accuracy of longer reads. In the trimming step, overlaps are recalculated to locate and filter out regions of low coverage and high error. Reads are overlapped two more times to correct specific types of errors (i.e. missed hairpin sections for 2D reads, adapters, chimeric reads) and to adjust the error rate per overlap, before the actual assembly phase starts. With adjustments to account for erroneous alignments and residual errors, assembly essentially follows the same procedure as CABOG, another Celera-based pipeline [100].

Due to its thorough yet relatively efficient correction steps, Canu is significantly more accurate than both its predecessor Celera/PBcR and most other tested assemblers. In benchmarks on *Enterobacter kobei* and *S. cerevisiae* reads, it often produced an assembly with fewer indels and mismatches than others, often with higher contiguity [97; 91; 90]. These results are in line with the author's own assessment [93].

SMARTdenovo SMARTdenovo is a long read OLC-assembly pipeline that was originally intended to work with PacBio reads, but has been shown to produce assemblies of reasonably high continuity from MinION reads as well [90]. Surprisingly, it does so without an error correction step prior to assembly, making SMARTdenovo a faster alternative to Canu.

As detailed on its Github page, SMARTdenovo first attempts to find read overlaps for each read in three steps at increasing accuracy by first searching hits in sorted k-mer tables twice and then using a banded Smith-Waterman algorithm. To find overlaps that were missed in this process, it subsequently repeats the process for pairs of reads that should overlap, given the extent to which they are overlapped by other reads. Next, low quality or chimeric read ends are identified by their decreased coverage by other reads and removed. Finally, SMARTdenovo borrows PacBio's directed alignment graph consensus (DAGCon) algorithm [104] to produce the consensus assembly.

As expected, SMARTdenovo was shown to outperform Canu in terms of computing efficiency [90; 105]. However, benchmarks on *S. cerevisiae* reads demonstrated that assemblies by Canu generally show higher identity with the reference sequence [91; 90]. This is possibly due to the fact that the HGAP algorithm leveraged for error correction was originally intended to work with PacBio reads, which have a different error distribution. Notably, Schmidt *et al.* showed that SMARTdenovo produced an assembly of higher contiguity for the large tomato (*Solanum pennellii*) genome and, when preceded by Canu's pre-assembly error correction module, obtained an even more contiguous assembly with fewer predicted errors than either Canu or SMARTdenovo could, while still remaining faster than Canu alone [105].

Minimap & Miniasm In terms of speed and computational efficiency, the OLC-based pipeline consisting of Minimap and Miniasm [96] has a definite advantage over other existing tools [97; 90; 91]. This efficiency was reached through the omission of the consensus step and the use of minimizers. Much like the k-mer hash table used by Canu's MHAP [93], a minimizer is a memory-efficient hashed representation of a sequence. Minimap computes the set of minimizers of a sequence, the "sketch", by finding the k-mers represented by the smallest hash value within a certain window size of each position of the sequence. The complement of each k-mer is also considered. Decreasing the window size will increase the returned number of minimizers and allow for more accurate alignment, at the cost of increased computational requirements. Minimap then performs all-versus-all mapping by identifying hits between minimizers of different sequences. The found overlaps are passed on to Miniasm, which constructs an assembly graph. First, potential artefacts are removed from each read by identifying the longest stretch with a coverage of three or more other reads, and then clipping off the ends that fall outside this region. Then reads contained within other reads are removed and small bubbles, less than 50 kb in length, are popped (i.e. a consensus is taken in cases where paths split and later join up again). Finally, sequences can be extracted from stretches of the graph without multi-edges to form unitigs. The error rate at this point is practically the same as that in the raw reads, emphasizing that correct basecalling is essential for the eventual quality of the assembly. The graphical fragment assembly (GFA) output format of Miniasm conveniently allows both graphing of the uncorrected assembly and addition of consensus error correction tools, such as Nanopolish or Racon, to the pipeline.

In March of 2016, the authors of Minimap and Miniasm reported assembly of MinION reads of an *E. coli* genome in a single contig. In May of the same year, Judge *et al.* assembled an *Enterobacter kobei* genome in 16 contigs with an N50 of 662 kbase in two minutes, while the next fastest assembler (Canu) took two hours, however their benchmark showed that the omission of an error correction step caused the eventual assembly quality of *E. kobei* to be too low to properly assess by the QUAST analysis tool [97].

ABruijn While more traditional DBG assemblers performed worse than OLC assemblers on assembling long error-prone reads [92], the approach taken by the ABruijn assembler has shown more promise [95]. To account for the high error rate, ABruijn filters all k-mers occurring in the reads by their frequency; if a k-mer occurs few times for given dataset and genome sizes, it is assumed that it contains basecalling errors and it is removed. Then k-mers are fused into so-called "solid strings", sequences that contain no other occurring sequences as substring. The ABruijn graph is then drawn by representing solid strings as vertices and connecting them where connections exist in the reads. The edges are weighted by the number of positions between the first bases of the connected solid strings. The assembler consults the weights in this graph to quickly identify overlaps between reads, allowing to select on a minimum overlap length and maximum overhang length. The assembly graph is constructed by starting with the graph for an arbitrary read and iteratively extending it by overlapping it with other reads.

ABruijn also includes an error correction routine, during which a best consensus between reads is found by identifying low-error stretches and, in between those stretches, choosing the consensus sequence that maximizes the likelihood of the read sequences.

In two independent benchmarking efforts (2D, R7.3, MAP005/006 and 2D, R7.3/R9, MAP006/007), ABruijn assembled an *S. cerevisiae* genome with higher contiguity than other included assemblers (Canu, Minimap/Miniasm, SMARTdenovo and PBcR) [91; 90] (Table 2.1). However, ABruijn was also the only assembler to produce chimeric contigs. Furthermore, Canu's assemblies showed higher identity with the reference genome. Thus ABruijn's assembly routine tends to return longer contigs, while Canu is less error-prone.

TULIP As more reads are required to cover larger genomes, and as the time required for all-vs-all overlapping increases quadratically with an increasing number of reads, it follows that the overlap step of OLC assemblers may take unfeasibly long for very large genomes. To tackle this issue, The Uncorrected Long read Integration Process (TULIP) takes a different approach to read overlapping [76]. Instead of all-vs-all alignment, short seed sequences are selected, which the assembler then attempts to align with long reads. This drastically cuts down the overlapping complexity and makes efficient use of long reads to cover long stretches of the genome between the seed regions. The resulting graph represents seeds as vertices and the connecting reads as edges. In a graph cleaning step, vertices with multiple in- or outgoing edges are revisited. Spurious and superfluous edges are removed aggressively, thus producing a linear graph. Note that, as the name implies, TULIP does not perform basecalling error correction.

The success of assembly using TULIP highly depends on proper seed selection. To avoid spurious connections between reads, the seeds need to be sufficiently unique in the genome and contain few sequencing errors. If available, SGS reads may be used to construct seeds, although with the increasing accuracy of MinION reads, the ends of long reads may be used as well. Apart from cutting out the need for SGS methods, the latter approach has the added advantage that pairs of seeds are connected by at least one long read. Furthermore, as TULIP is not able to assemble regions in which the gap between seeds is larger than the read length, a proper seed density over the entire genome is required. If a marker map is available for the genome, this information can be used to control the distribution of seeds in the selection process.

As a first demonstration of TULIP's efficiency, Jansen *et al.* assembled the genome of the European eel *Anguilla anguilla* (approximately 850Mbp) with 18x coverage in three hours (excluding sequence polishing), requiring only 4.4GB of RAM and four threads [76]. The resulting assembly was more continuous than the SGS-based reference genome. As was the case with Minimap/Miniasm however, the current quality of MinION reads combined with the lack of an error correction step necessitates post-assembly correction. The authors further showed that missed seed alignments were the most commonly encountered issue during graph simplification, followed by tangled alignments due to repetitive seeds and spurious alignments. The seeds, constructed from short SGS reads, only

underwent selection by uniqueness, which did not lead to an equal distribution over the entire genome; however, density remained high enough for successful assembly. The authors noted that assembly using the tips of MinION reads as seeds proved successful for *Escherichia coli* genomes, but this has not been attempted for larger genomes yet (personal communication, May 1, 2017).

HINGE Although long reads provide a definite edge when attempting to resolve repeat regions, issues may still occur if not all individual repeats are spanned by at least one whole read. In such cases, HINGE may provide a solution. Rather than attempting to resolve frayed rope structures in the assembly graph afterwards, HINGE pre-processes the reads to separate repeat regions that are entirely spanned by a read (and are thus more easily resolvable) from those that are not, and collapses the latter beforehand [94].

First, HINGE attempts to identify reads that wholly or partly overlap a repeat region. It does so by performing all-vs-all alignment and then selecting those reads of which a stretch aligns to a proportionally larger number of other reads than the rest of the read. The intuition behind this is that reads from all copies of a repeat region existing in the genome align to each other, thus causing a characteristic abrupt increase in alignments for reads that overlap these repeat regions. Repeat regions covered entirely by at least one read can be easily resolved and are omitted from the following procedure. Of the reads lining the same repeat region, the reads that extend furthest into the repeat region (regardless of the location of the actual copy), are designated "hinges". In the subsequent greedy extension of the hinges, the contigs will split at the hinge regions. Like Miniasm, HINGE outputs its assembly in the form of a graph. As its authors show, this is particularly useful for circular genomes.

HINGE provides an elegant solution to long repeat resolution, by separating resolvable regions from unresolvable ones beforehand. Its authors compared HINGE to Miniasm on PacBio reads of 997 circular bacterial genomes and found that overall, HINGE produced a completed genome in more cases than Miniasm could [94]. Whether the precaution taken by HINGE is necessary is dependent on the genome under consideration and the used reads; if the genome is known to contain repeats longer than most of the reads, the described approach would be justified.

2.3.3 Post-assembly correction tools

A number of tools attempt to improve, or "polish", assemblies by remapping long reads to the assembly and adapting the assembly to increase local resemblance to the reads. These polishing tools may be essential to use after assembly pipelines that do not include a consensus step themselves, such as Minimap/Miniasm, but have also frequently been used to polish assemblies produced by assemblers that do include this step. In this section, a selection of polishing tools is described. Notably, ONT recently published the source code for their own neural network-based polisher, Medaka. Although this tool may become a valuable addition to assembly pipelines in the future, it is currently in an early stage of development.

Nanopolish Nanopolish attempts to find an optimal consensus between an assembly and the raw current signal output by the MinION, by iteratively proposing and evaluating small adaptations to the assembly based on the original reads [75]. The proposal mechanism for adaptations works in two steps. First, reads are aligned to the assembly and the resulting multiple alignment is divided in 50 bp subsequences of the assembly. For each read aligning partly or fully to a subsequence, sections in which events perfectly align to the assembly are detected. The consensus sequence between each pair of aligning sections is replaced by the aligned read subsequence, creating an initial set of alternative candidate sequences. In the second step, this set is further extended by proposing every possible one-base deletion, insertion and substitution in the previously generated candidate sequences. Of this set, the sequence maximizing the likelihood of observing the raw signal is picked. This process allows Nanopolish to explore a decent number of likely modifications, while remaining computationally tractable. As of v0.8.4, available information on methylation sites can be used to improve the quality of those sites even further. As epigenetic modifications were shown to influence the current signal [77], this may result in a significant improvement.

Nanopolish was found to improve assembly quality, regardless of the assembly tool used. One study on *E. coli* sequencing data reported that identity to the reference genome rose from 89% to 99% when Nanopolish (v0.4.0) was applied after Minimap/Miniasm, while improvement after Canu was more modest (98.2% to 99.6%) [106]. Notably, the previously mentioned Wick *et al.* benchmark showed that methylation-aware polishing brought the identity of reference-based assemblies up significantly to 99.9% versus 99.7% after polishing without methylation-awareness. An assessment on a *de novo* assembly has yet to be made.

Despite its efficient searching heuristic of block replacement and mutation, running Nanopolish remains a time-consuming step; in two separate benchmarking efforts, one on an *E. kobei* assembly produced by Minimap/Miniasm and one on a *S. cerevisiae* assembly by Canu, running Nanopolish (v0.4.0 and v0.5.0 respectively) required more than a month of extra CPU time [97; 90]. Later versions of Nanopolish (especially v0.7.0 and up) were reported by its authors to work much faster.

Racon Racon [107] corrects MinION assemblies by finding a consensus sequence between reads and the assembly through the construction of partial order alignment (POA) graphs. After alignment of the reads by a mapper of choice (e.g. Minimap or Graphmap), Racon segments the sequence and finds the best alignment between a POA graph of the reads and the assembly. By default, the alignment is performed using the Needleman-Wunsch algorithm, which can align sequence and POA graph with little adaptation. The alignment process is sped up by parallelization. Racon was reported by its authors to be two orders of magnitude faster than the popular (yet currently deprecated) Nanocorrect [75] after assembly of an *E. coli* genome by Miniasm, albeit not quite as good at diminishing the error rate (to 1.31% versus 0.62% for Nanocorrect). Compared to consensus steps in Falcon [108] and Canu [93] on that same assembly, Racon remains an order of magnitude faster while producing similar error rates. A closer look at the remaining errors reveals that

the majority consists of indels. As indel basecalling has drastically improved in newer basecallers (versus the pre-transducer basecallers used by Racon's authors), these would likely allow Racon to reach even lower error rates. Finally, the total genome size estimate following application of Racon was closer to the reference genome size than the estimates of Canu, Falcon and Nanocorrect.

2.4 Discussion

Nanopore sequencing is a promising new venue in biology research. Inexpensive, small, capable of producing long reads and freed from the need for nucleotide labeling or amplification, it is conceivable that the MinION will make cost-effective, fast and portable *de novo* whole genome sequencing of even complex genomes possible in the future. In this review, an attempt was made to give an updated overview of the progress in this field, focusing in particular on *de novo* whole genome sequencing.

Available basecaller tools have been improving rapidly in accuracy. Notable recent improvements include the move toward raw signal-based calling and the inclusion of a transducer. For the next step in a typical sequencing routine, assembly, OLC-assemblers are currently considered the best option for accurate *de novo* nanopore-based assembly. The choice of assembler should be adapted to the characteristics of the genome and the priorities of the user. Canu is a complete and accurate solution, although SMARTdenovo was shown to be much faster against slightly diminished accuracy. The best of both methods may be obtained by combining Canu's error correction module with SMARTdenovo. Minimap/Miniasm is by far the fastest option available, but as it lacks any form of error correction, cannot produce a usable genome draft without any post-assembly correction. For large, complex genomes, TULIP may be the more tractable alternative. Lastly, stand-alone post-assembly consensus error correction tools Nanopolish and Racon are a worthwhile addition in *de novo* sequencing pipelines and a necessity in combination with assemblers that do not contain a sequencing error correction step of their own.

Currently, the most prominent obstacle for *de novo* sequencing using the MinION is the high error rate of the reads. Improving basecalling accuracy would not only improve assembly quality in a direct manner, but may also allow more computationally efficient assembly.

The active research community surrounding the MinION has booked great progress in both the development of new applications and improvements on accuracy of existing ones. ONT also continuously works on improvements for both its hardware and software platforms, and regularly updates its users on this. Although these updates often entail welcome new features or some form of accuracy improvement, it should be noted that this policy has also lead to some difficulties. Developers may not be able to keep pace with ONT when evaluating, updating or calibrating their tools, and users may not always know which tool is suited best to their data and needs. As a result, most published studies, including tool benchmarking efforts, were conducted using older or multiple chemistries.

Although such growing pains are to be expected for a novel fast-developing field of research, the MinION's current state of development may allow for some increase in stability, thus giving the user community the time for proper evaluation.

Competing interests

No competing interests were disclosed.

Grant information

The authors declared that no grants were involved in supporting this work.

Acknowledgements

Giovanni Maglia (University of Groningen, Groningen, The Netherlands) provided helpful advice related to the physical basis of nanopore sequencing.

CHAPTER 3

FRETboard: semi-supervised classification of FRET traces

This chapter has been published as:

Carlos de Lannoy, Mike Filius, Sung Hyun Kim, Chirlmin Joo and Dick de Ridder. "FRETboard: semi-supervised classification of FRET traces" *Biophysical Journal* 120 (2021): 3253-3260

Supplementary material available at:

https://doi.org/10.5281/zenodo.6773393

Abstract

Förster resonance energy transfer (FRET) is a useful phenomenon in biomolecular investigations, as it can be leveraged for nano-scale measurements. The optical signals produced by such experiments can be analyzed by fitting a statistical model. Several software tools exist to fit such models in an unsupervised manner, but lack the flexibility to adapt to different experimental setups and require local installations. Here we propose to fit models to optical signals more intuitively by adopting a semi-supervised approach, in which the user interactively guides the model to fit a given dataset, and introduce FRETboard, a web tool that allows users to provide such guidance. We show that our approach is able to closely reproduce ground truth FRET statistics in a wide range of simulated single-molecule scenarios, and correctly estimate parameters for up to eleven states. On *in vitro* data we retrieve parameters identical to those obtained by laborious manual classification in a fraction of the required time. Moreover, we designed FRETboard to be easily extendable to other models, allowing it to adapt to future developments in FRET measurement and analysis.
Availability: Source code is available at `https://github.com/cvdelannoy/FRETboard` (DOI:10.5281/zenodo.4006487). The FRETboard classification tool is also available as a browser application at `https://www.bioinformatics.nl/FRETboard`.

3.1 Introduction

Over the past decades, single-molecule Förster resonance energy transfer (smFRET) experiments have provided fundamental insights in biomolecular structure and many molecular mechanisms [e.g. 7; 8; 109; 9; 110; 111; 31; 112]. Although all experiments essentially rely on the same principle – that of distance-dependent energy transfer efficiency between fluorescent donor and acceptor dyes – the use of different labeling schemes allows for versatile application. For instance, dyes fixed to two points on a single molecule may provide spatial information valuable in solving its structure [8; 9], or register structural dynamics as the molecule exerts its biological function [7]. Fixed to separate molecules, FRET may provide information on the occurrence and nature of molecular interactions [109–111]. If a single dye pair provides insufficient information, a multitude of dye pairs may even be read out simultaneously by making use of stochastically blinking dyes [113]. As each labeling scheme produces data of a different nature, it follows that widely applicable smFRET data analysis software should be flexible enough to adapt to these varying natures.

smFRET users currently may choose from a wide array of software packages, which mostly vary in scope and underlying trace analysis algorithms [32; 114–118; 34; 119; 35]. The core utility included in all packages is the estimation of FRET efficiency distributions and transition rates given a set of traces, for which most rely on some flavor of hidden Markov model (HMM). While packages differ in how HMMs are fitted, the overall consensus is that the influence of the user in the fitting process should be minimized to safeguard objectivity. However an

automated fitting procedure may find one of several good fits of which some may not make sense given the experimental context, a context which the user could provide.

Here we show that an HMM may be given that context for any particular FRET data set using a semi-supervised fitting approach, i.e. by allowing the user to manually curate classification of a limited number of traces to steer the model (Figure 3.1). Such direct intervention at the classification level makes model fitting a flexible, intuitive and computationally light-weight process. We further increase accuracy by introducing a more elaborate HMM structure that, to our knowledge, has not previously been applied to smFRET data. Using several additional features derived from the original signal further boosts accuracy and increases the flexibility to adapt to data of different labeling schemes. Our method is available for use through our web tool, FRETboard. FRETboard is a smFRET trace analysis solution which also supports data filtering and graphing utilities. As the smFRET field is rapidly developing and diversifying, we designed FRETboard to grow with the needs of the user community; by allowing anyone to easily extend FRETboard with existing or future classification algorithms, our tool may continue to serve as a unifying web front-end for high-level users with both niche and general classification needs.

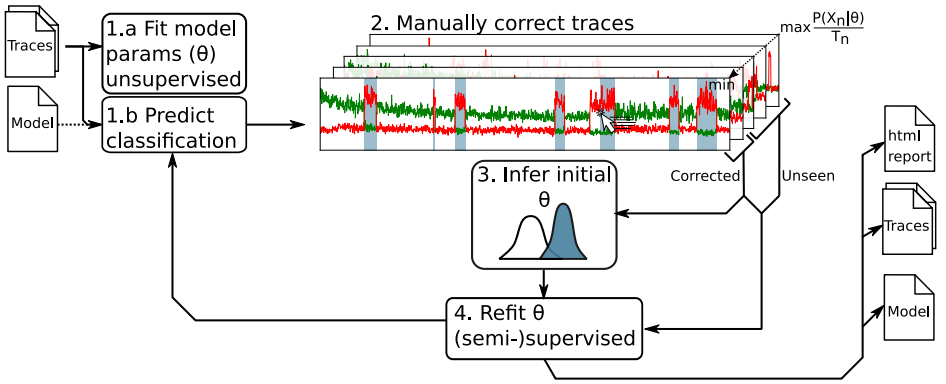


Figure 3.1: Our semi-supervised classification workflow for FRET trace classification, as implemented in FRETboard, is divided in 4 steps. **[1a]** FRET traces are uploaded to the web server, after which parameters (θ) of an initial model are fitted unsupervised and **[1b]** an initial trace classification is predicted. A suitable model generated during a previous FRETboard run can also be supplied to skip initial fitting. **[2]** The user is then shown the predicted classification of the trace at index n, X_n, with duration T_n, for which model fit was poorest based on duration-normalized trace probability given current model parameters ($P(X_n|\theta)/T_n$), and is asked for manual correction. **[3]** The curated trace is then used to reinitialize the model, which **[4]** is then trained in a (semi-)supervised fashion. Steps 2 to 4 may repeated until model fit is deemed satisfactory by the user.

3.2 Methods

The FRET trace analysis method presented here is different from previous methods in three respects: the semi-supervised training approach, the structure of the models that are fitted and the features on which they were trained. Here we explain and justify our choices in each of these respects.

3.2.1 Semi-supervised model fitting

We introduce semi-supervised fitting of models for the classification of FRET traces. The aim is to utilize the user's insight in the data structure to steer repeated light model fitting procedures, so that the resulting model will match the user's intuition. The fitting procedure is summarized in Figure 3.1 and a high-level description is given below. A more detailed description is given in Supplementary section 3.1.3.

The procedure is initialized by fitting an unsupervised HMM on all loaded traces in a traditional manner, using randomly generated initial parameters which are then fitted using an implementation of expectation maximization (EM). The only user-provided guidance at this point is the number of states that should be recognized. Traces are classified and the trace for which the state path probability normalized over sequence length $(P(X_n|\theta)/T_n$ was lowest is presented to the user for manual correction of the classification.

The probability of assigned states for a given trace may be poor due to the presence of noise. If this is the case a user may choose to assign noisy measurements to a state they deem appropriate. However if it is more appropriate to remove the noise, as is the case in bleaching and blinking events, a the user may filter these measurements out by assigning them to a separate state reserved for such events. Such a state may be discarded prior to FRET distribution and transition rate analysis. Alternatively, model fit may suffer if the trace contains more or less states than those included in the current HMM. For example, we show this to be the case for simulated traces containing three states if the model contains only two states (Supplementary figure 3.S3). In that case the user may simply adjust the number of states and adjust classification appropriately.

After applying manual corrections, the first semi-supervised training round on all loaded traces is started. State distributions and transition rates can now be deduced from the corrected trace and be used as initial parameters, after which the HMM is refitted on supervised and unsupervised traces simultaneously using semi-supervised EM. After refitting, traces are reclassified and the trace now marked by the lowest state path probability is presented to the user. The procedure is repeated until the user finds that presented traces are correctly classified.

3.2.2 Features

We trained our models on a combination of four features. The proximity ratio E_{PR} is included as an approximation of FRET efficiency and is defined as:

$$E_{PR} = \frac{F_{A_{em}}^{D_{ex}}}{F_{sum}}$$

Here $F_{D_{em}}^{D_{ex}}$ and $F_{A_{em}}^{D_{ex}}$ are the original donor and acceptor emission and F_{sum} denotes the summed donor and acceptor intensities $F_{D_{em}}^{D_{ex}} + F_{A_{em}}^{D_{ex}} = F_{sum}$. We also include the summed intensity as a separate feature, as it is expected to aid in the detection of bleaching.

Furthermore we used two time-aggregated features that capture the variability of features over a sliding-window of five measurements; the Pearson correlation coefficient between $F_{A_{em}}^{D_{ex}}$ and $F_{D_{em}}^{D_{ex}}$ (C) and standard deviation of F_{sum} $(\sigma_{F_{sum}})$. These features may aid models in capturing feature distributions characteristic for state transitions (Figure 3.2B). We specifically refrain from using $F_{D_{em}}^{D_{ex}}$ and $F_{A_{em}}^{D_{ex}}$ as features, as systematic variations frequently occur between experiments or even within the same experiment, which decreases the generalizability and re-usability of a trained model.

3.2.3 Model structures

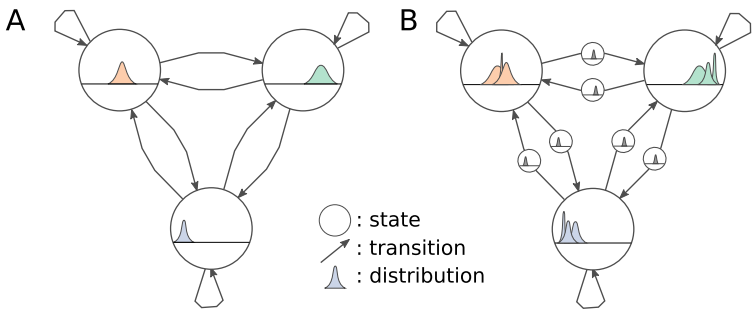


Figure 3.2: Hidden Markov model graph structures used in this work: (**A**) the plain "vanilla" structure, and (**B**) the "GMM-HMM" structure, in which each state contains a Gaussian mixture distribution and additional states for feature distributions at state transitions. Circles denote states with a characteristic feature distribution, arrows denote transitions. Three-state structures are shown here, however similar structures with an arbitrary number of states can be constructed.

We evaluate the performance of two HMM structures in a semi-supervised learning setting (Figure 3.2). Each structure implements transitions and emissions differently and can be extended to an arbitrary number of states. The "vanilla" structure produces a straight-forward fully-connected HMM sporting no further modifications. The "GMM-HMM" structure models emissions using a Gaussian mixture model (GMM), which add the flexibility to classify noisier distributions as a single state, using multiple Gaussians. The number of Gaussians per GMM is determined per state using a Bayesian information criterion (BIC)-selection procedure. Furthermore this structure adds additional "edge states" between states,

which are trained on measurements around a detected transition. Transitions between states may only occur through these edge states. If state transitions are marked by a signature distribution in a certain feature, this distribution is captured by the edge state, which allows for more accurate detection of state transitions. For example, under some labeling schemes transitions between FRET events may be marked by negative correlation between donor and acceptor signal, thus training a GMM-HMM on such data while using pearson correlation coefficient as a feature would make use of this fact.

3.2.4 Implementation

To facilitate application of our method we developed FRETboard, a browser-based graphical user interface (GUI) for semi-supervised training of segmentation and classification algorithms (Supplementary figure 3.S1). In addition to intuitive example supervision, FRETboard offers users the flexibility to choose between model structures and opt which features to include. As we foresee that more suitable supervise-able classifiers may be proposed for the growing number of labeling schemes in the future, we also offer users the option to write custom algorithms and train them through the same FRETboard front-end. However due to the security risk of code injection that is inherent to running such custom code, users are advised to only allow this option on private machines that are not exposed to the public network.

Traces may be loaded in plain text, binary 64-bits or photon-HDF5 [120]-format, and may be corrected for background emission using our DBSCAN-based filter (Supplementary section 3.1.4).

After the training procedure, the user may generate a report detailing feature distributions per state and transition rates. Transition rates are derived by deriving a transition matrix (A) from classified data, converting from discrete to continuous rates and multiplying by the frame rate f_s, thus arriving at corrected transition rates F (Equation 3.1) [114].

$$F = I + f_s \cdot \log A \tag{3.1}$$

Here I is the identity matrix and log denotes the natural matrix logarithm operation. 95% confidence intervals (CIs) of transition rates are estimated by repeatedly extracting transition rates from bootstrapped data. The CI is then reported using the bootstrap standard deviation on each parameter. Note that bootstrapping CIs is applicable to almost any (semi-)supervised model, thus any user-defined algorithm can make use of the same method. FRETboard is available as a web tool (`https://www.bioinformatics.nl/FRETboard`), thus freeing users from the burden of installation and maintenance, but it can also be used and hosted on a private server. FRETboard was written in python 3.7 (`https://www.python.org`). The GUI was implemented using the Bokeh interactive visualization library (v1.4.0)[121] (Supplementary figure 3.S1). Included HMM model structures were implemented using pomegranate (v0.13.4) [122] and scikit-learn (v0.21.2) [123].

3.3 Results

Below we validate our analysis method on four *in silico* and four *in vitro* data sets. To demonstrate the flexibility of our method, the different sets were simulated or recorded assuming a variety of realistic labeling schemes. All data sets used here are freely available (`https://git.wageningenur.nl/lanno001/fretboard_data`). All FRETboard runs were performed on a laptop running Ubuntu 18.04, on four CPU cores (Core i7 1.80GHz, Intel Corp.) with 4GB of memory. In total we supervised ten traces for each data set (3% and 10% of the total number of reads for *in silico* and *in vitro* data sets respectively), making use of the four described features (E_{PR}, F_{sum}, C, and $\sigma_{F_{sum}}$).

We assessed how well semi-supervised HMMs were able to reproduce ground truth parameters for simulated state sequences or, in the case of *in vitro* data, parameters acquired by manual labeling. To test whether predicted E_{PR} distributions attain a mean comparable to either reference value, we apply the two one-sided t-tests (TOST) procedure [124]. That is, for a given state s the predicted mean of E_{PR} ($\hat{\mu}_s$) and the reference mean (μ_s) are calculated and two one-sided t-tests are employed to test $H_{01} : \hat{\mu}_s - \mu_s \leq -\frac{\delta}{2}$ and $H_{02} : \hat{\mu}_s - \mu_s \geq \frac{\delta}{2}$ versus $H_1 : -\frac{\delta}{2} < \hat{\mu}_s - \mu_s < \frac{\delta}{2}$. A rejection of both null hypotheses implies that the difference between emission means is significantly smaller than δ E_{PR} percentage points. Here we test for a maximum deviation of five or ten percentage points ($\delta = 0.05$ and $\delta = 0.1$ respectively), which we consider sufficiently accurate for many current applications. We report the TOST p-value for a given $\delta = \delta^*$ as $p_{\delta=\delta^*}$. Reported 95% CIs around estimated transition rates were calculated using FRETboard's built-in bootstrapping method, using a bootstrap size of 100.

3.3.1 Performance on *in silico* data

To demonstrate the flexibility of our approach, we simulated realistic FRET traces based on three different labeling schemes and classified them using a semi-supervised vanilla HMM. Briefly, the first two data sets contain two and three FRET states respectively, which are separable based on E_{PR} only (Supplementary figure 3.S2A, B). The third data set contains three states, of which the third is identical to the second in its proximity ratio but has a different transition rate, making it a "degenerate state" (Supplementary figure 3.S2C). For a full description of the simulation methodology see Supplementary section 3.1.1.

In all cases, estimated mean E_{PR} significantly differed less than 5 percentage points from the ground truth mean($p_{\delta=0.05} << 0.001$, Figure 3.3A-C). Most ground truth transition rates fell well within bootstrapped 95%-CIs around the predicted rates (Figure 3.3E-G). If a degenerate third state was present it was identified as such, however occasional misclassification between the two high-FRET states led to transition rates slightly differing from ground truth values (Figure 3.3G). In general we find that two rounds of semi-supervised training suffices to obtain parameter estimates close to ground truth values, while further rounds account for minor adjustments (Supplementary figure 3.S4). Apart from the manual correction, no further parameter tuning or other user input was required,

demonstrating that semi-supervised training provides the expected flexibility while maintaining accuracy.

To stress-test our method on a more difficult case, we generated a data set in which eleven FRET states of differing E_{PR}-levels were present (Supplementary figure 3.S2D). The mean E_{PR} values were distributed such that their corresponding donor-acceptor distances were evenly distributed, thus causing lower and higher ends of the E_{PR} spectrum to be more densely crowded with states. As both visual examination during the training procedure and the produced results indicated poor performance by the vanilla HMM (Supplementary figure 3.S5), we repeated our analysis using the GMM-HMM structure. Six of eleven E_{PR} distribution means significantly differed less than five percentage points from the ground truth means ($p_{\delta=0.05} << 0.001$, Figure 3.3D), with the remaining five differing less than ten percentage points ($p_{\delta=0.1} << 0.001$) and transition rate estimates were closer to ground truth values than vanilla HMM estimates (Figure 3.3H). This demonstrates another strength of our method; if the user discovers upon visual inspection that simpler models cannot capture a user's classification, the training procedure is light enough that a more elaborate model can be selected on the fly, after which the analysis can continue without extra effort.

Finally we assessed how well the semi-supervised approach of FRETboard mitigates the effects of decreased signal-to-noise ratios (SNRs). We find that the GMM-HMM structure correctly estimated transition rates down to an SNR of 4.0, while E_{PR} levels could still be deduced down to an SNR of 2.75 (Supplementary figure 3.S6). Interestingly the accuracy of manual labeling of traces decreased with SNR as well, thus increasingly erroneous supervised examples further contribute to the innate difficulty of classifying low-SNR data.

3.3.2 Performance on *in vitro* data

We further validated our method on experimental data generated under immobilization schemes often used in single-molecule FRET, each marked by different classification challenges. Similar to our simulations, our *in vitro* data contains up to two types of FRET events and one ground state, which may be discernible by proximity ratio or transition rate. Lacking knowledge of the state sequence, we manually classified our data sets and used this classification to estimate proximity ratios and transition rates. A more extensive description of experimental methods can be found in Supplementary section 3.1.2. All experimental data was analyzed using the GMM-HMM model structure, as the vanilla structure did not show a satisfactory increase in classification quality as training progressed (Supplementary figure 3.S7).

First we designed an experiment in which a donor (Cy3)-labeled single-stranded (ss) DNA, containing a target site A, is immobilized through biotin-streptavidin conjugation on a quartz slide (Figure 3.4A). During measurement, we add (Cy5)-labeled eight-nucleotide 'imager' strands, which upon binding to site A produce FRET events marked by anti-correlated donor and acceptor signals (Figure 3.4E). Similar labeling schemes have previously seen application in point accumulation for imaging in nano-scale topography (PAINT) methods and the study of on- and

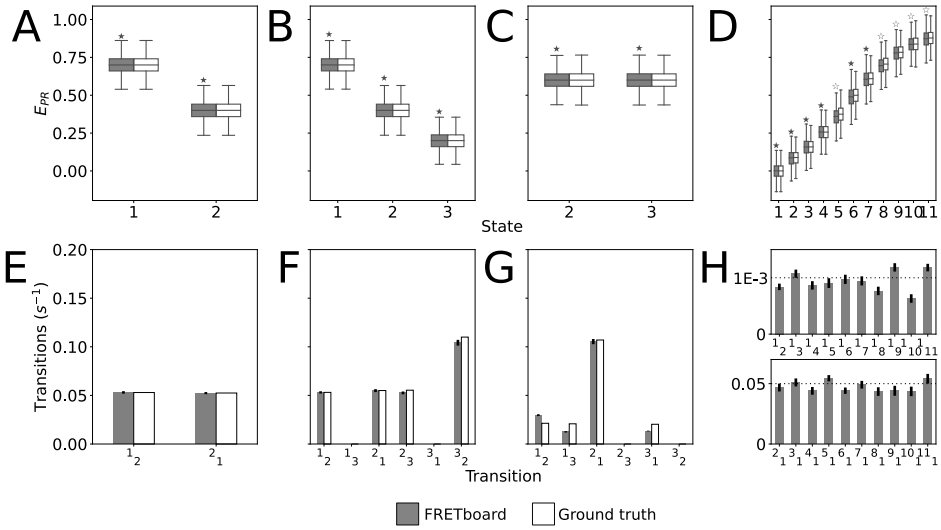


Figure 3.3: (E_{PR}) distributions per state in boxplots (**A-D**) and transition rates (**E-H**) as estimated by a semi-supervised hidden Markov model on four simulated data sets based on different labeling schemes; (**A**,**E**) producing two types of spatially separable FRET events, i.e. with different E_{PR} distributions, (**B**,**F**) three spatially separable states, (**C**,**G**) three states of which one containing no donor or acceptor signal and the others exhibiting only a kinetic difference, i.e. separable by transition rate and (**D**, **H**) eleven spatially separable states. Symbols above E_{PR} boxplots denote how much estimated means significantly differ from the ground truth at most ($\star$: 0.05, ☆: 0.1). Solid black lines in **E-H** indicate 95% bootstrapped CIs. Dotted lines in **H** denote ground truth values.

off-rates (k_{on} and k_{off}) in biological systems [125; 126]. In this labeling scheme, bleaching of the donor dye due to continuous excitation occurs frequently, which may negatively impact kinetics analysis. Instead of requiring the user to remove bleaching events prior to analysis, we capture them in a separate state while training our GMM-HMM. This bleached state may then be discarded. Following this approach, we find that manually obtained transition rates for ground to high-FRET state and vice versa indeed fall within their respective estimated CIs ($0.119s^{-1}$ versus $CI : (0.095 - 0.122)$ and $0.546s^{-1}$ versus $CI : (0.468 - 0.578)$ respectively, Figure 3.4M). Estimated E_{PR} values (0.151 and 0.816 for ground and high-FRET states respectively) significantly differ by less than five percentage points from ground truth values ($p_{\delta<0.05} << 0.01$, Figure 3.4I).

We also performed the reverse experiment, in which the acceptor is immobilized and the donor is attached to the imager strand (Figure 3.4B). Although this labeling scheme does not suffer from dye bleaching as much, the lack of anti-correlation in the signal is expected to increase the difficulty of classification. Nonetheless, here too the predicted E_{PR} distribution mean of the high-FRET state (0.844) differed from the manually obtained value by less than 5 percentage

points ($p_{\delta=0.05} << 0.01$, Figure 3.4J). As no dye is observable in the ground state under this labeling scheme, its E_{PR} value is meaningless and not analyzed here. The manually obtained transition rate from high-FRET to ground state fell within its predicted 95% CI ($0.367s^{-1}$ versus $CI : (0.333 - 0.429)$), while the rate for ground to high-FRET state was slightly underestimated ($0.012s^{-1}$ versus $CI : (0.007 - 0.011)$, Figure 3.4N).

Next, we evaluated performance in two scenarios where two FRET states are present. For these experiments we followed the same experimental procedure, but simultaneously flushed in two types of donor-bound free-floating imager strands, at a 1:1 ratio (Figure 3.4C,D).

In the first experiment, the second imager strand was complementary to a second target site B at 15nt from the acceptor – 10nt further than target site A – where imager strand binding should produce an intermediate E_{PR} (Figure 3.4G). Upon analysis, the GMM-HMM model found E_{PR} means of 0.85 and 0.72 for states 2 and 3 respectively, matching the manually obtained state means ($p_{\delta<0.05} < 0.01$, Figure 3.4K). Most transition rate estimates fell within the predicted 95% CIs, except for that from mid-FRET to ground state ($0.439s^{-1}$ versus $CI : (0.483 - 0.780)$, Figure 3.4O). Upon inspection of traces, we find that some short mid-FRET events had been erroneously detected in noisy ground state stretches – a common occurrence in smFRET analysis and therefore not explicitly accounted for by e.g. removing traces from analysis manually.

In the second experiment, site A was targeted with a second imager strand of 7nt – 1nt shorter than its counterpart – (Figure 3.4D) which should increase the off-rate and produce a degenerate state (Figure 3.4H). Here too our GMM-HMM produced parameter estimates close to manually obtained values on traces containing degenerate states, which is surprising given our results on *in silico* data. Predicted transition rates from state 2 to ground state were higher at $2.22s^{-1}(CI : 1.87 - 2.57)$, than that of state 3 – $0.43s^{-1}(CI : 0.382 - 0.479)$ –, which resembled transition rates seen in other *in vitro* experiments (Figure 3.4O). Presumably, state 2 corresponds to the annealing of the shorter 7nt imager strand. Both are in close agreement with manually obtained rates ($2.24s^{-1}$ and $0.411s^{-1}$ respectively).

Finally, we compared FRETboard estimates for our experiments against those of three other tools, each of which employs a different solution to the classification problem: ebFRET [116], which fits an HMM using a Bayesian approach, MASH-FRET [115; 119], a flexible software suite that includes several other tools and infers transition rates through exponential fitting, and DeepFRET [35], which filters traces from noise using a neural network before classifying them with an HMM. As we focus on transition rate and emission distribution analysis only, we fed the same background-subtracted traces to each tool and forced them to use the correct number of states if it allowed us to do so. We find that for each experiment, FRETboard returns transition rate and E_{PR} distribution estimates that are equally close or closer to the manually derived values as estimates of other tools (Supplementary figure 3.S9). Analysis procedures followed for other tools are detailed in Supplementary section 3.1.5.

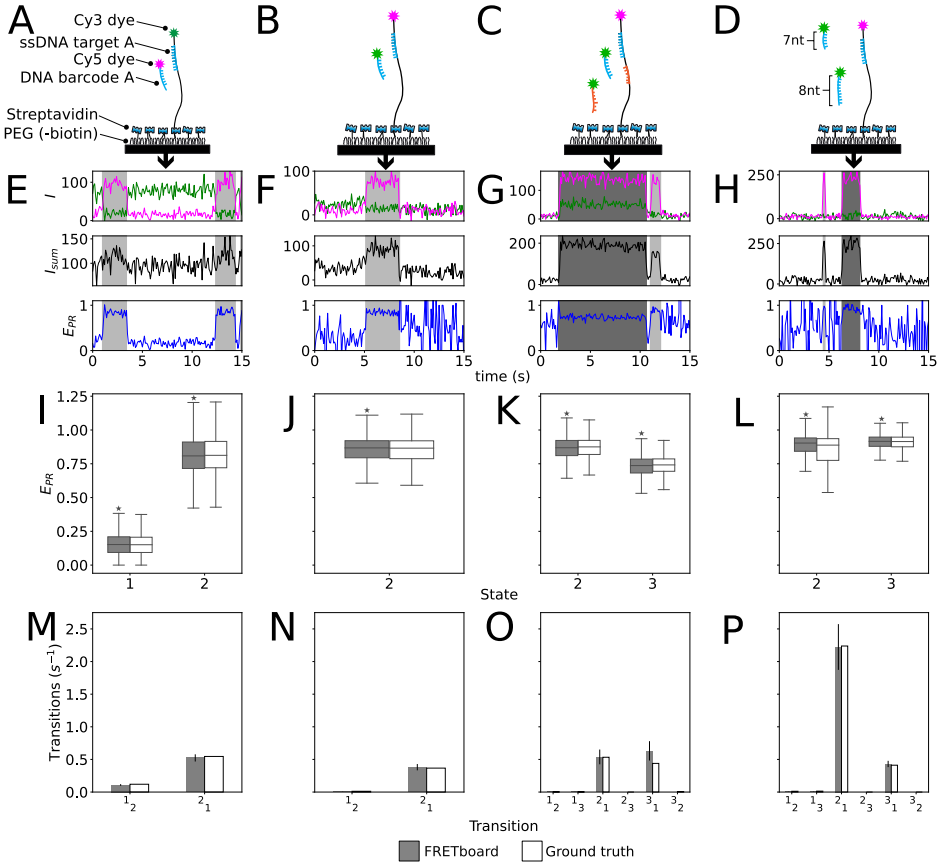


Figure 3.4: Used labeling schemes (**A-D**), examples of events produced as correctly found by semi-supervised HMM classification (**E-H**), E_{PR} distribution boxplots per state (**I-L**) and estimated transition rates (**M-P**) for four different labeling schemes on which our semi-supervised HMM fitting method was evaluated. From left to right, these labeling schemes were producing single-type FRET events using an immobilized donor (**A**) or an immobilized acceptor (**B**), two types of events producing high- and mid-FRET events (**C**) and two types of kinetically different events (**D**). In **J-L**, the non-FRET state [1] is omitted as E_{PR} values are meaningless for labeling schemes in which the donor is not immobilized. In **I-J**, $\star$ marks E_{PR} distributions for which estimated means significantly differ less than 0.05 from values acquired through manual analysis. In **M-P** solid lines denote bootstrapped 95% CIs.

3.4 Discussion

We show that semi-supervised classification models, in particular hidden Markov models (HMMs), are capable of capturing properties of FRET events in a wide array of realistic experimental scenarios, using a combination of input features derived from the original donor and acceptor dye emission intensities. We also provide an HMM structure that is better suited for semi-supervised learning than the straight-forward fully connected model, and provide a particular advantage in noisy real-world and complex data sets containing more than two states. To accommodate for the intensive user interaction required for this method we developed FRETboard, an intuitive browser-based tool that allows data filtering, model training, classification and report generation.

We placed our method in the landscape of existing analysis tools by comparing its analysis results to three other often-used tools. We find that FRETboard model parameter estimates were as accurate or more accurate than those of other tools, using manual classification as a reference. Furthermore, FRETboard sports several other features not previously seen in smFRET analysis, including implementation as an analysis server that can be used remotely in the browser and the possibility to run custom algorithms. However it should be noted that FRETboard is focused on trace analysis and does not currently support any functionality for trace extraction from microscope images. Several other tools such as MASH-FRET [119; 115], iSMS[117], and SPARTAN [118] do include this functionality, and therefore provide a more complete software solution.

Another important caveat particular to our method is related to the quality of the user's supervision. Contrary to other approaches we give responsibility for proper classification to the user, embracing the pros and cons of user input; on the one hand, it allows for efficient training and yields results that match the user's intuition, on the other hand it matches mistakes that the user may make. Users may derive the knowledge necessary for supervision (e.g. number of states) from their experimental setup as was done in the validation shown here, or start analysis agnostically and estimate such information from the traces itself. To test our method in the latter scenario we have entered FRETboard into the KinSoft challenge (`https://sites.google.com/view/kinsoftchallenge/home`), the first blind assessment of smFRET kinetic analysis tool performance. A publication on the challenge results is pending.

Lastly, we encourage users to design their own classifiers and test them through the FRETboard interface; many more supervise-able HMM flavors and entirely different classifiers exist and may be a better fit than the models currently included for certain experimental data. In consultation with the authors of such custom classifiers, these may also be included in future releases of FRETboard. This would allow it to become a unifying front end for FRET trace analysis, with back end support for the expanding variety of smFRET experimental methods.

Author contributions

C.L., M.F., C.J. and D.R. conceived and designed the project. C.L. developed algorithms and software, and performed analyses. M.F. designed and performed experiments. S.H.K. designed, wrote and performed the simulation procedure. D.R. supervised algorithm development. C.L., M.F. and S.H.K. wrote the manuscript. All authors discussed results and improved the manuscript.

Acknowledgements

This project was supported by the Foundation for Fundamental Research on Matter, vrije programma (Single Molecule Protein Sequencing). The authors declare no conflicts of interest.

CHAPTER 4

PoreTally: run and publish *de novo* Nanopore assembler benchmarks

This chapter has been published as:

Carlos de Lannoy, Judith Risse and Dick de Ridder. "poreTally: run and publish *de novo* Nanopore assembler benchmarks" *Bioinformatics* 35 (2019): 2663-2664

Supplementary material available at:

`https://doi.org/10.5281/zenodo.6773393`

Abstract

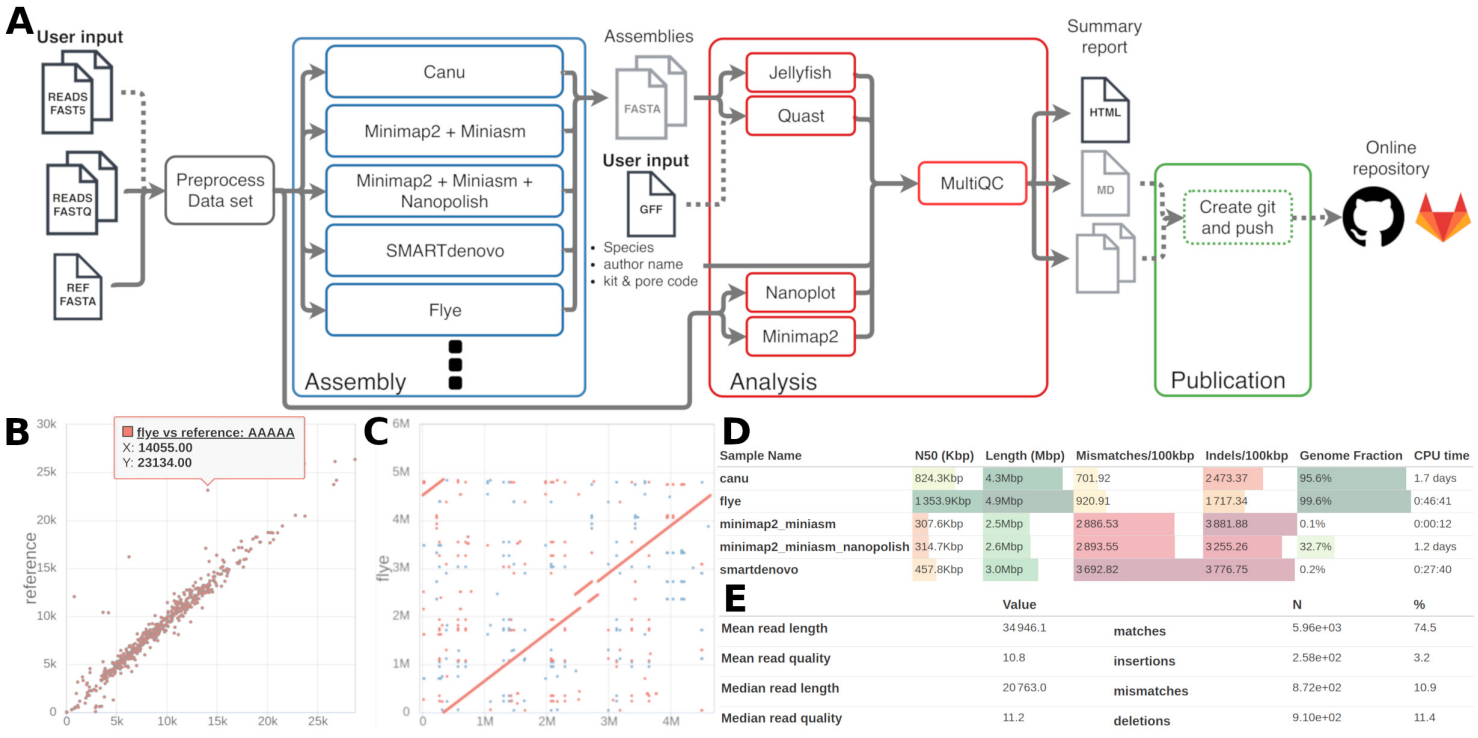
Nanopore sequencing is a novel development in nucleic acid analysis. As such, nanopore sequencing hardware and software are updated frequently and extensively, which quickly renders peer-reviewed publications on analysis pipeline benchmarking efforts outdated. To provide the user community with a faster, more flexible alternative to peer-reviewed benchmark papers for *de novo* assembly tool performance we constructed poreTally, a comprehensive benchmarking tool. poreTally automatically assembles a given read set using several often-used assembly pipelines, analyzes the resulting assemblies for correctness and continuity, and finally generates a quality report, which can immediately be published on Github/Gitlab.
Availability: poreTally is available at https://github.com/cvdelannoy/poreTally, under an MIT license.

4.1 Introduction

Nanopore sequencing is a third-generation nucleic acid sequencing method that produces error-prone long reads of consistent quality. From 2014 onwards, Oxford Nanopore Technologies (ONT) introduced the three first commercial nanopore sequencers: the MinION, GridION and PromethION. As is to be expected of the first attempts at a radically different approach to sequencing, the hardware, software and sample preparation practices for these devices are updated frequently.

Although ONT's update schedule provides the user community with tools of steadily increasing value, it requires downstream analysis tool developers to repeatedly re-parameterize their tools. Providing updated benchmarks for said tools has proven difficult as well. In the case of *de novo* assembly pipelines, several benchmarks of the most popular tools at a given point in time have been run and published in peer-reviewed journals [e.g. 91; 127]. However, in each case ONT significantly improved some aspect of read quality before or shortly after the publication appeared, rendering it partially outdated. Furthermore, each benchmarking effort focused on one species, while it has been shown that read quality and best assembly practices may differ from one taxon to the next [128; 129].

To address this issue, we propose a community-driven frequent benchmarking practice and present a tool to facilitate this. We encourage research groups that make use of nanopore sequencing to benchmark assembly pipelines for their organisms of interest following a standardized routine and publish their results directly on-line. This will provide other users working on the same or similar taxa with an indication of the best assembly pipeline for their case with the most up-to-date hard- and software. Our tool, poreTally, supports this practice by offering often-used assembly pipelines, a performance analysis routine, report generation and publication on free repository hosting services Github or Gitlab in one package. Optionally, users may submit their results to a collective benchmark effort. Submitted benchmark results will periodically be summarized and reported back to the community.



Sample Name	N50 (Kbp)	Length (Mbp)	Mismatches/100kbp	Indels/100kbp	Genome Fraction	CPU time
canu	824.3Kbp	4.3Mbp	701.92	2 473.37	95.6%	1.7 days
flye	1 353.9Kbp	4.9Mbp	920.91	1 717.34	99.6%	0:46:41
minimap2_miniasm	307.6Kbp	2.5Mbp	2 886.53	3 881.88	0.1%	0:00:12
minimap2_miniasm_nanopolish	314.7Kbp	2.6Mbp	2 893.55	3 255.26	32.7%	1.2 days
smartdenovo	457.8Kbp	3.0Mbp	3 692.82	3 776.75	0.2%	0:27:40

	Value		N	%
Mean read length	34 946.1	matches	5.96e+03	74.5
Mean read quality	10.8	insertions	2.58e+02	3.2
Median read length	20 763.0	mismatches	8.72e+02	10.9
Median read quality	11.2	deletions	9.10e+02	11.4

Figure 4.1: Schematic representation of the three-step poreTally workflow **(A)**, which consists of assembly pipeline running (blue), assembly quality analysis (red) and online publication (green), and examples of four elements in the HTML summary report generated by poreTally: a Jellyfish-based *k*-mer abundance graph **(B)**, Nucmer/Quast-based contig alignment plot **(C)**, aggregated assembly pipeline performance metrics **(D)** and read set quality measures based on Nanoplot and Minimap2 output **(E)**. Nanopore reads used in this figure were obtained from Loman Labs at University of Birmingham, UK, reads and constructed assemblies were aligned to NCBI reference sequence NC_000913.3 to obtain displayed metrics.

4.2 Methods

poreTally is written in Python 3 (Python Software Foundation, `https://www.python.org/`). Its workflow is divided in three steps. First it executes a number of assembly pipelines on a provided dataset. Then several analysis tools are run on the produced assemblies. In the final step, summarized results are automatically published on Github or Gitlab in a readable format (Figure 4.1A).

Installation and running

poreTally can be run using the provided docker container or by installing it through pip. In the latter case, some minimal requirements have to be met first; Python 3.6, a miniconda/anaconda installation and Git.

poreTally relies on the Snakemake workflow management system [130] and its excellent integration with conda environments. First a Snakemake workflow is generated, containing a set of commands for every assembly pipeline, and executed. If required, dedicated conda environments containing the necessary tools for a particular pipeline are generated on the fly.

After the pipelines have finished running, a second workflow is executed to assess the quality of the original read set and that of the produced assemblies. Nanoplot [131] and Minimap2 [132] are used to evaluate raw read quality and a combination of Quast [133] and Jellyfish [134] is used to assess assembly quality. Finally, results are summarized in an interactive HTML-report using MultiQC [135] (Figure 4.1B–E). If a git repository address was provided, the results are now uploaded to Github or Gitlab. At this point the user is also asked permission to upload the results to a collective benchmarking effort (see Supplementary section 4.1). A detailed step-by-step walkthrough of the full benchmarking process is given on the poreTally Github.

4.3 Discussion

To adapt to the high pace of development in nanopore sequencing, its users have sought out faster and more transparent methods to disseminate knowledge. With poreTally we aim to support this movement. poreTally automates benchmarking of *de novo* nanopore read assembly pipelines and immediate publication of benchmark results in public repositories, thus allowing users to conduct frequent benchmarking and independently map the landscape of nanopore assembler pipelines.

Funding

This work was funded by the Foundation for Fundamental Research on Matter (Single Molecule Protein Sequencing).

CHAPTER 5

BaseLess: lightweight detection of sequences in raw MinION data

This chapter has been submitted for publication as:

Ben Noordijk, Reindert Nijland, Victor J. Carrion, Jos M. Raaijmakers, Dick de Ridder and Carlos de Lannoy. "baseLess: lightweight detection of sequences in raw MinION data"

Supplementary material available at:

https://doi.org/10.5281/zenodo.6773393

Abstract

With its candybar form factor and low initial investment cost, the MinION brought affordable portable nucleic acid analysis within reach. However, translating the electrical signal it outputs into a sequence of bases still requires high-end computer hardware, which remains a caveat when aiming for deployment of many devices at once or usage in remote areas. For applications focusing on detection of a target sequence, such as infectious disease or GMO monitoring, the computational cost of analysis may be reduced by directly detecting the target sequence in the electrical signal instead. Here we present baseLess, a computational tool that enables such target-detection-only analysis. BaseLess makes use of an array of small neural networks, each of which efficiently detects a fixed-size subsequence of the target sequence directly from the electrical signal. We show that baseLess can accurately determine the identity of reads between three closely related fish species and can classify sequences in mixtures of twenty bacterial species, on an inexpensive single-board computer.
Availability: baseLess and all code used in data preparation and validation is available at `https://github.com/cvdelannoy/baseLess`, under an MIT license.

5.1 Introduction

Nucleic acid (NA) sequencing is no longer the costly endeavor it once was; while two decades ago analysis of a single genome could occupy multiple labs over several years [136], technological innovations have now driven the per-base cost down sufficiently to allow routine sequencing for other purposes than scientific discovery, including forensics [137] and clinical diagnoses [138–140]. The case for such usage was strengthened further with the introduction of Oxford Nanopore Technology (ONT)'s MinION, a low-cost, small-size sequencing device. No longer inhibited by high initial investment costs or poor portability, small laboratories and individual users may now opt for in-house sequencing and on-site analysis in remote locations [141–143].

This development was possible due to the introduction of a new sequencing mechanism; rather than the fluorescence-based sequencing-by-synthesis approach employed by previous devices, the MinION sequences DNA strands of arbitrary length by ratcheting them through a nanopore while reading out the electric current [144]. This readout is colloquially referred to as a "squiggle". As the nucleotide combination residing in the nanopore at a given moment influences the electrical resistance, the squiggle carries information on the sequence. In a process termed "basecalling", the NA sequence is deduced from the squiggle.

Although the MinION itself is an inexpensive NA sequencer, real-time data analysis currently still requires at least a high-end laptop. For some applications, e.g. the distribution of thousands of devices for infectious disease screening, this may bring along prohibitively high additional costs. It would therefore be beneficial if inexpensive computing hardware could be used instead. Depending on the intended purpose, a computationally lighter analysis pipeline may be a solution. As fast computing hardware is mainly required for basecalling, some basecallers

have been developed that trade off lower resource requirements against a decreased basecalling accuracy. DeepNano-blitz [145] is the most recent open-source example of such an implementation, while ONT's proprietary basecaller guppy has a "fast" running mode for this purpose.

Not all applications require information on the full read sequence however. If only detection of a set of known sequences is required, these sequences could be detected directly in the squiggle instead, potentially reducing the computational load even further. Several direct-from-squiggle sequence detection methods have been proposed. Kovaka *et al.* developed UNCALLED [27], which assigns a probability for each 5-mer potentially matching to each squiggle segment and then compares probable series of 5-mers to a pre-indexed genome to quickly map the read to its likely location. Its original purpose is to facilitate "adaptive sampling", that is, to rapidly detect the likely origin of a read while the strand is still being sequenced, so that sequencing of strands from non-target sources may be terminated early [146]. UNCALLED can easily be repurposed to perform general sequence detection; however, the index-based approach carries several disadvantages. Efficiency decreases for larger and more repetitive genomes and re-indexing is required to attune the tool to a new target sequence. Moreover, accuracy was found to be low for short sequences [26]. Similarly to UNCALLED, SquiggleNet was designed for adaptive sampling [26]. Following a more straightforward approach, it uses a neural network trained for the recognition of a given genome to decide whether squiggles belong to a species or not. Previously SquiggleNet was found to outperform UNCALLED in terms of both accuracy and processing speed, but the required re-training of SquiggleNet for a given species is a highly resource- and time-consuming process.

Here we introduce baseLess, a computationally efficient and flexible approach to direct sequence detection (Figure 5.1). Using an array of small neural networks, each pre-trained to recognize a single k-mer, baseLess can determine whether a read can be mapped to a given sequence or not. Configuring our tool to detect a sequence requires only the selection of target k-mers and their associated pre-trained neural networks. We show that baseLess can perform species detection on eukaryotic whole genome sequencing data against a background of similar species, as well as 16S-based species detection of prokaryotes agnostic of background sequences. BaseLess is more accurate than direct sequence detection pipelines and less computationally demanding than basecalling-and-mapping, allowing it to run on more affordable (∼$100) analysis hardware. As such, it removes an important economical bottleneck for highly distributed and remote field analysis using the MinION.

5.2 Results

5.2.1 Tool structure

BaseLess deduces the presence of a target sequence by detecting squiggle segments corresponding to salient short sequences, k-mers, using an array of convolutional neural networks (CNNs) (Figure 5.1A). Each CNN detects a single k-mer, a

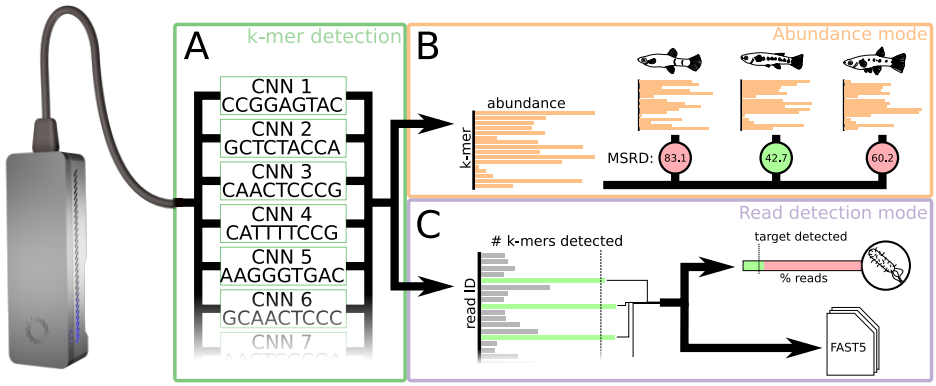


Figure 5.1: Schematic overview of the baseLess sequence detection tool. **(A)** baseLess detects sequences using an array of pre-trained interchangeable neural networks, each of which detects a specific k-mer. These k-mers have been specifically selected to allow discrimination of a target sequence. **(B)** In abundance mode, network outputs are summed over all reads and presented as an estimate of k-mer abundance. This estimate is compared against genome-based estimates for several closely related species by calculating the mean squared rank difference (MSRD). The species for which the MSRD is lowest is the most likely source of the reads. **(C)** In read-based detection mode, a target sequence is sought in each individual read. A minimum fraction of k-mers needs to be detected in a read before it is classified as a target read. A target species is detected if a minimum fraction of analysed reads can be assigned to it. Reads assigned to the target species are also stored in a FAST5 file for further analysis.

relatively simple task, thus the network complexity can be kept low. This divide-and-conquer strategy has several advantages. All CNNs can process a read in parallel, which makes baseLess computationally efficient. Furthermore, given a library of pre-trained CNNs, baseLess can easily be reconfigured to detect a different target sequence by combining a different set of CNNs. Finally, sufficient data to train the CNNs is usually available; shorter sequences generally occur more often than longer sequences, thus a read set of any source, once corrected for basecalling errors (see Methods), provides sufficient data to train for a wide range of k-mers.

To complete the baseLess network, the outputs of the CNN array are combined using one of two aggregation rules. If configured in "abundance mode", baseLess returns the number of occurrences found for each k-mer, which may then be compared to abundance estimates derived from a target genome (Figure 5.1B). In "read detection mode", the network is configured to decide whether a sufficiently large fraction of its k-mers have been found in a given read to conclude that it contained the target sequence (Figure 5.1C). These modes are explained and evaluated in more detail below.

5.2.2 Abundance-based species detection

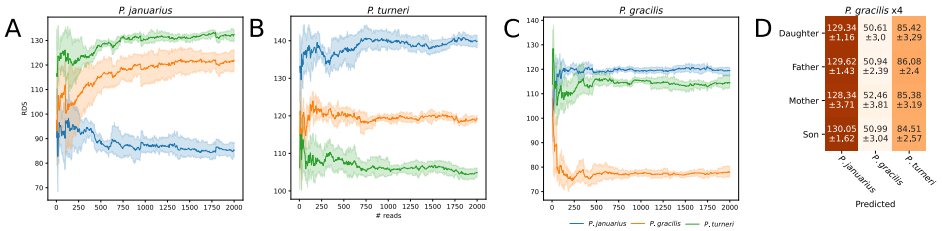


Figure 5.2: Mean squared rank differences (MSRD) based on a comparison of k-mer abundances estimated from reads by baseLess and the abundances in genomes of three closely related fish species. A low MSRD indicates that k-kmer abundances in sample and genome are alike and that reads are thus more likely derived from that genome. Results are presented for **(A)** *Phalloptychus januarius*, **(B)** *Poeciliopsis turneri*, **(C)** *Poeciliopsis gracilis* and **(D)** a family of four *P. gracilis* individuals of which no assembled genome was used in the configuration of baseLess. In A, B and C colored areas denote the 95% confidence interval. In D, numbers are formatted as mean ± standard deviation over 2000 reads.

As baseLess provides fast and accurate inference on low-cost hardware, it is highly suited to determine the species or strain to which a given individual belongs at remote sampling locations, or at many locations simultaneously. Practical applications of such usage may be found in ecological monitoring of visually similar species or forensic investigation of patented crops. For such tasks, baseLess should be configured in abundance mode, which requires the target species' genome and a set of background genomes – genomes of species from which the target species must be discerned. The k-mer set used for discrimination is then found by combining k-mers that are highly abundant in the target genome yet found less than average in the background genomes, or vice versa. To determine the origin of a sample, baseLess ranks k-mers by abundance as measured in the reads and compares it to their abundance ranking in the target and background genomes, using the mean squared rank difference (MSRD):

$$MSRD = \frac{1}{N}\sum_{N}^{n=1}(m_{b,n} - m_{r,n})^2$$

Here $m_{b,n}$ and $m_{r,n}$ are the rank for k-mer m based on abundances in analyzed reads and in a reference respectively. N is the total number of k-mers analyzed in the reads.

To test baseLess' performance in this scenario, we analysed unamplified whole-genome MinION reads from three related guppy species: *Phalloptychus januarius*, *Poeciliopsis gracilis* and *Poeciliopsis turneri* [147]. In three separate analyses, we configured our tool for detection of one of the species against the other two, using Illumina short-read assemblies of the same individuals as target and background genomes to avoid the risk of detecting species based on MinION-specific sequencing

errors. We then analyzed a set of 2,000 MinION reads originating from the target species. We found that baseLess consistently calls the correct species for each analyzed readset (Figure 5.2A-C). Moreover, baseLess did not need the full 2,000 reads for any classification; stable MSRD values were attained after 52, 352 and 84 reads for *P. gracilis*, *P. januarius* and *P. turneri* respectively. To test whether baseLess indeed detects differences between species and not between individuals, we also ran classification on samples of a family of four *P. gracilis* individuals, using a k-mer set selected using the genome of an unrelated *P. gracilis* individual. baseLess consistently called the correct species while requiring less than a hundred reads.

Interestingly, the k-mer rankings also followed the phylogenetic relation between the species; in all detection experiments, MSRD values for *P. gracilis* and *P. turneri* were consistently closer to each other than to *P. januarius*, which is indeed of a different genus. This implies that, even if the genome of the correct species is not included, the relative identity of a sample may be inferred by comparing measured abundances to several related species.

5.2.3 Read-based species detection

In specific applications, a sample may contain a mixture of DNA of many species, from which a species of interest must be detected. Possible scenarios include the screening for infectious disease agents at events or at national borders, or detection of indicator species for environmental health. In 16S cDNA samples, baseLess may be configured to detect such a species of interest by selecting a combination of k-mers unique to the target's 16S sequence, and running it in read detection mode. In this configuration, baseLess detects each k-mer on a per-read basis, rather than summing occurrences over all reads as is done in abundance mode. If a minimum fraction of target k-mers is found in a read, it is attributed to the target species. The raw squiggle of found target sequences is stored to allow more in-depth analysis at a later stage, while non-target reads are discarded to decrease data storage footprint. To allow reliable detection of a wide range of species against an arbitrary genomic background, we composed a list of k-mers which both varied in sequence composition and produced easily differentiable squiggle segments. This list was further filtered to only contain k-mers that are present in NCBI 16S sequences, yet sufficiently rare to allow for species discrimination (see Methods).

To test this approach we amplified and sequenced the 16S rRNA regions of an artificial microbial community of twenty-one known species on the MinION (Supplementary table 5.S1). 400,000 reads were fully basecalled and mapped to the twenty-one genomes of the species to determine their likely origin. No reads were mapped to the *Porphyromonas gingivalis* genome, thus this species was left out of subsequent analysis. Read numbers for other species varied between 11 and 51,040.

We reconfigured baseLess and ran inference for each of the species in a five-fold cross validation scheme, to determine how well it could identify the origin of reads. Running speeds were benchmarked on two different classes of hardware; the Nvidia

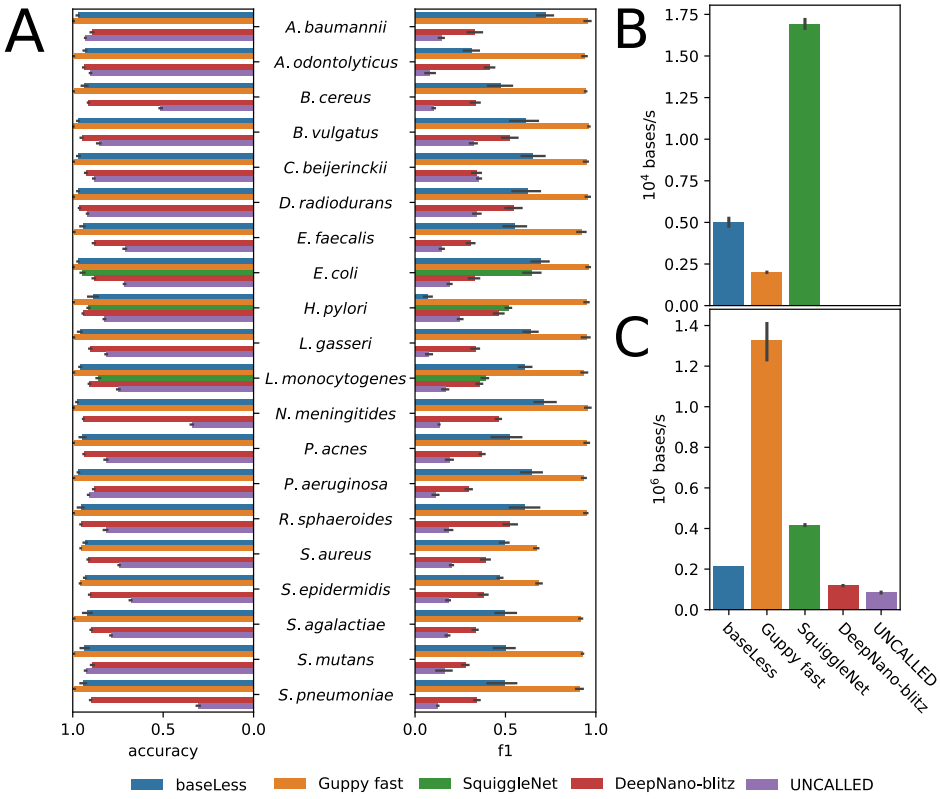


Figure 5.3: Performance in 16S-based species detection for a community of 20 species, of baseLess and four competing analysis tools; DeepNano-blitz; Guppy (fast mode); UNCALLED and SquiggleNet. As DeepNano-blitz and Guppy are basecallers, not read mapping tools, minimap2 is used to obtain the mapping. **(A)** Accuracy and F_1 score per species for all four tools. Species are sorted by read abundance in the dataset, from high (top) to bottom (low). Black bars denote standard deviation over five cross-validation folds. **(B)** Analysis speed for each tool on the Nvidia Jetson Nano (2GB) single-board computer and **(C)** on a high-end desktop computer. Black bars denote standard deviation over ten analysis runs on 1000 reads.

Jetson Nano (2GB), a ∼$100 single-board computer with dedicated GPU (Nvidia Maxwell, 128 cores@921MHz), and a mid-tier desktop computer with dedicated GPU (Nvidia GeForce RTX 3070, 5888 cores@173GHz). To allow straightforward comparison, all tools were given access to 3 CPU cores and the GPU if required. We compared the performance of baseLess on 16S read classification to that of four other pipelines: full basecalling by either DeepNano-blitz [145] or guppy in "fast" mode, followed by mapping using minimap2 [132] ("DeepNano+minimap2" and "guppy+minimap2" respectively); UNCALLED [27]; and SquiggleNet [26].

BaseLess consistently identified its target reads with more than 95% accuracy

and an F_1 score of 0.54 on average (Figure 5.3A), with the exception of *Helicobacter pylori*. Guppy+minimap2 outperformed other pipelines consistently, however compared to DeepNano+minimap2 and UNCALLED, baseLess yielded a higher accuracy and a higher F_1-score for all species. Under default settings SquiggleNet only had sufficient data to classify reads of the three species for which the most reads were available: *E. coli*, *H. pylori* and *L. monocytogenes*. On these species, baseLess preformed similarly to, or better than SquiggleNet.

In the speed benchmark baseLess processed 5.0 kilobases per second (kbps) on the Jetson Nano. We found that baseLess was more than twice as fast as Guppy+minimap2 (2.0 kbps), which could not make use of the Jetson GPU due to software compatibility issues. SquiggleNet ran the fastest at 17 kbps. None of the tools were able to match the theoretical maximum throughput of the MinION (230 kbps). We were unable to install DeepNano-blitz and UNCALLED on this hardware, possibly due to incompatibility with the energy-efficient AARCH64 CPU architecture used in the Jetson Nano and most other single-board computers. As expected, processing speeds were much higher on high-end desktop hardware with the three GPU-accelerated tools – baseLess, guppy+minimap2 and SquiggleNet – performing best. At 1.3 megabase per second (Mbps), guppy+minimap2 was faster than all other tools. SquiggleNet (420 kbps) was again faster than baseLess (210 kbps), which in turn out-competed DeepNano+minimap2 (120 kbps) and UNCALLED (84 kbps).

5.3 Discussion

In this work we proposed a method to identify whole genomes or amplified sequences in nanopore reads by detecting salient k-mers using an array of individual, interchangeable neural networks. We show that baseLess, our implementation of this method, is capable of correctly classifying single-species whole-genome sequencing samples, given the target species' genome and a set of off-target genomes. This is useful for species determination of larger organisms, though not for environmental samples of microbes, which contain many species of which most may be unknown. We therefore also implemented an alternative running mode, which allows microbial species detection against an unknown background, suitable for smaller genomes or PCR-amplified samples.

The world-wide demand for microbe screening, most prominently for infectious disease agents, is currently filled mostly by lateral-flow antibody and qPCR tests. Antibody tests have a turnaround time of minutes, require little training to use and can be mass-produced at low expense, but require a redesign and subsequent re-distribution for the detection of different targets. Moreover, detection is not as reliable as that of nucleic acid analysis [148]. qPCR is generally more reliable, but also requires newly designed primers for different targets. We propose that MinION-based sequence detection using baseLess could fulfill a role similar to that of qPCR; while quick and inexpensive mass produced antibody tests are difficult to improve on for sustained monitoring of a single biological agent, the fast turnaround times and adaptable nature of MinION-based sequence detection

would provide an improvement over the logistics and organization required for qPCR testing sites. Reconfiguring baseLess for the detection of a new agent or variant only requires loading the networks for a different set of k-mers. Moreover, as found target reads are stored, these may be analyzed in depth afterwards, giving researchers an unprecedented wealth of information on mutations from each detected occurrence of the agent. Especially relevant in this context, though left unexplored here, would be microbe detection using direct DNA or RNA sequencing, as omission of PCR steps would bring down turnaround time even further.

To our knowledge, baseLess is the first tool built specifically to perform MinION-based sequence detection, but other tools can be repurposed to perform the same task. We thus compared baseLess to two fast full basecalling-and-mapping pipelines and two adaptive sequencing tools. Of these competitors, only the guppy+minimap2 pipeline could consistently classify reads with a higher accuracy than baseLess. However, guppy was not optimized for use with low-powered hardware such as the Nvidia Jetson Nano featured in this work, thus it ran slower. This may change if current software compatibility issues are resolved in future releases, although then still the 2GB memory limit of this hardware may prove problematic. SquiggleNet performed similarly to baseLess in terms of accuracy and was more than three times faster on the Jetson Nano. However, due to its high training data requirements it could only be evaluated on three of the twenty species tested here. BaseLess did not suffer from this disadvantage as it only needs examples of k-mers to train, which may be obtained from any source. Furthermore, SquiggleNet requires retraining to detect new species, while baseLess only needs reconfiguration for a different set of k-mers. We thus argue that baseLess has more potential to be developed into an accurate yet flexible sequence detection tool than its competitors.

Several venues may be explored to further optimize our workflow. Importantly, baseLess' computational efficiency can be further increased; we ran our tool using Tensorflow, a fully equipped deep learning library, however to run inference on low-powered hardware more efficiently, light-weight frameworks such as Tensorflow-lite and TensorRT may be employed. Further optimization would allow baseLess to analyse reads at a speed more similar to the MinION's throughput, or preserve computational resources for other tasks, such as driving the MinION itself. Furthermore, we note that the amplification of 16S sequences used in our 16S performance evaluation remains a bottleneck in sequence detection. Instead, the MinION may also be used to directly sequence RNA. As ribosomal RNA makes up a large part of the total RNA content of prokaryotes [149], it would be interesting to evaluate classification based on unamplified RNA content instead.

In summary, the results obtained inspire confidence that using baseLess, the MinION can be turned into a mobile species detector for under $1,000, thus paving the way for large-scale nucleic acid-based detection of biological agents in any environment.

5.4 Methods

5.4.1 Network design procedure

Individual k-mers are recognized using 1D convolutional neural networks implemented in Tensorflow 2.3 [81]. We optimized hyperparameters through 100 rounds of training and evaluation on $33,549$ and $3,241$ held-out training and test reads respectively to obtain the final network architecture (Supplementary figure 5.S1). After each round of training and evaluation, the next hyperparameter set was selected using a tree-structured Parzen estimator implemented in hyperopt [150]. The objective function was designed to increase the F_1 score while decreasing network size:

$$L = (1 - F_1) + \lambda \cdot \frac{p_c}{p_{max}}$$

Here L denotes the loss to be minimized, p_c denotes the number of parameters in the current iteration of the network and p_{max} denotes the maximum number of parameters attainable given the boundaries of the parameter search space. The parameter λ controls the trade-off between accuracy and network size and was set to 0.01.

Networks output the posterior probability of their target k-mers being present in a squiggle segment. The threshold above which this posterior probability is considered sufficiently high to detect the presence of a k-mer was chosen to maximize the F_1-score, using a grid search on training data for probabilities between 0.75 and 0.999 with a step size of 0.001. For read detection mode, the fraction of k-mers to be detected before the target sequence is considered present must be set as well. This parameter was optimized simultaneously with the posterior probability threshold.

5.4.2 False positive rate simulation

An optimal choice for the value of k should balance the abundance of a k-mer, such that it is rare enough to discriminate sequences, yet not so rare that it never occurs at all. We approximate this optimal value by considering the probability of detecting target sequences in random sequences by chance.

Assuming all canonical k-mers are equally represented, the expected number of k-mer occurrences in a read of length L_{read} is $L_{read} \cdot 2 \cdot 4^{-k}$ and the probability of a k-mer occurring in the sequence at least once can be estimated using a Poisson distribution. Assuming we draw networks detecting k-mers from a library of pre-generated networks A, the expected number of k-mers found in a target sequence at least once can be calculated:

$$E = P(X_{target} \geq 1) \cdot |A|$$

Here X_{target} is the number of occurrences of a k-mer in the target sequence, $|A|$ is the size of the k-mer network library and E is the expected number of k-mers in A found in the target sequence. A false positive occurs when a read

that does not contain the target sequence contains all the selected k-kmers of the target read by chance. The rate at which this occurs can be estimated as follows:

$$\text{FPR} = P(X_{non-target} \geq 1)^E$$

Here $X_{non-target}$ denotes the number of occurrences of a k-mer in the non-target sequence and FPR denotes the false positive rate. We performed FPR simulations for different values of k, representative values for non-target sequence lengths – 30 and 50 kb, representing full nanopore read lengths – and target sequence lengths – 0.6, 1.5 and 30 kb, representing BOLD barcodes [151], 16S sequences and whole coronavirus genomes respectively – and selected the value for k that minimized FPR. Both $k = 8$ and $k = 9$ returned good FPR values, thus we included k-mers of both sizes in subsequent steps.

5.4.3 *k*-mer library design

For 16S sequence detection we composed a library of 1,500 suitable k-mers, which should allow detection of a wide range of species. Similar to Doroschak *et al.* [152], we used an evolutionary algorithm to select k-mers that are dissimilar in sequence and produce easily distinguishable squiggles. To enforce sequence dissimilarity, only k-mers with a maximum Smith-Waterman score of 6 (assuming gap penalty, match score and mismatch score of -4, 1 and -1 respectively) to other selected k-mers are allowed, while squiggle dissimilarity is enforced by comparing simulated squiggles as produced by guppy (v. 5.0.11+2b6dbff). That is, we only accept modifications made to k-mers by the evolutionary algorithm if both the minimum and the average dynamic timewarping score between its squiggle and the other squiggles in the set increase. Furthermore, for the bacterial case study, we remove the outer 10 percentiles of most abundant k-mers based on 20,959 16S rRNA sequences obtained from NCBI (Bioproject:PRJNA33175) because these k-mers are excessively rare or ubiquitous. Additionally, k-mers containing four or more of G/C or 5 or more of A/T in a row are rejected as the length of homopolymer stretches can be difficult to detect in squiggles. Starting with a set of random sequences, we ran the evolutionary algorithm for ten rounds of decreasing numbers of proposed mutations per sequence; the initial two rounds applied 5 mutations in each sequence, after which the number of mutations decreased by one for each two rounds.

5.4.4 Nanopore sequencing

Poeciliidae reads were obtained from a previous study and have been obtained as described in [147]. For 16S reads, we sequenced pre-made DNA isolate of microbial mock community A (v3.1, HM-278D, BEI resources) on a MinION (Mk.1B, Oxford Nanopore plc.) using accompanying flowcell (FLO-MIN106) and 16S sequencing kit (SQK-RAB204).

5.4.5 Data preparation

All reads were basecalled using guppy (v5.0.11+2b6dbff) in high-accuracy mode. To obtain a ground truth species assignment for 16S reads, we mapped them using BLASTN (v2.9.0+) to the expected twenty-one bacterial GenBank genomes (Supplementary table 5.S1). The species to which the sequence identity was highest was selected as the ground truth species for that read. To correct sequencing errors and assign individual bases to each squiggle segment, we aligned reads to reference genomes using tombo (v1.5.1). *P. gracilis*, *P. januarius* and *P. turneri* reads were aligned to genomes constructed from the nanopore reads, while 16S reads were aligned to their respective GenBank genomes. These genomes were also used for salient k-mer detection in the evaluation of read detection mode. For abundance mode validation, k-mers were selected from GenBank short read genomes, built from Illumina reads of the same three individuals (GCA_903067085.1, GCA_902982915.1, and GCA_903068135.1 for *P. gracilis*, *P. januarius* and *P. turneri* respectively).

5.4.6 Benchmarking

We compared BaseLess performance on 16S reads to four other tools; UNCALLED (v2.0-127-g0fc1cab), SquiggleNet (v1.0), DeepNano-blitz (v1.0) and Guppy (v5.0.11+2b 6dbff, "fast" mode). As the latter two tools are basecallers and not mapping tools, the basecalled reads returned by these were mapped to target genomes using minimap2 (2.17-r941) to produce the final prediction.

We performed accuracy and F1 score benchmarks in stratified 5-fold cross validation on 335,000 reads. Tools were run on a PowerEdge R740 server (Dell), on three Xeon Gold 6242 CPUs @2.80GHz (Intel). As Guppy and SquiggleNet were optimized for GPU usage, they were run on a Tesla T4 GPU (NVIDIA). We ran all tools in a Snakemake [130] workflow.

Speed benchmarks were performed on two systems, an Nvidia Jetson Nano System-on-Module (2GB RAM, ARM CPU, 4 cores@1.43GHz, Nvidia Maxwell GPU, 128 cores@921MHz) and a mid-tier desktop computer (32GB RAM, AMD Ryzen 3700x CPU, 16 cores@3.6GHz, Nvidia GeForce RTX 3070, 5888 cores@173GHz). Inference was performed 10 times per tool over 1,000 reads. Tools were given access to 3 CPU cores and the GPU if they were configured to use it.

5.5 Acknowledgements

We thank Elio Schijlen and Bas te Lintel Hekkert for help with nanopore sequencing. We also thank Henri van Kruistum for the provision of raw nanopore reads and nanopore assemblies for *P. gracilis*, *P. januarius* and *P. turneri*.

CHAPTER 6

Evaluation of FRET X for single-molecule protein fingerprinting

This chapter has been published as:

Carlos de Lannoy*, Mike Filius*, Raman van Wee, Chirlmin Joo and Dick de Ridder. "Evaluation of FRET X for Single-Molecule Protein Fingerprinting" *iScience* 24 (2021): 103239

*: authors contributed equally

Supplementary material available at:

https://doi.org/10.5281/zenodo.6773393

Abstract

Single-molecule protein identification is an 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 efficiencies 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 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.
Availability: lattice modeling code and data are available on Github at `https://github.com/cvdelannoy/FRET_X_fingerprinting_simulation`, and on Zenodo at `https://zenodo.org/record/5330741`. Trace analysis code is available on Github, at `https://github.com/kahutia/transient_FRET_analyzer2`, experimental data is available on request.

6.1 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. [153]. 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 [154]. Not all of these proteoforms can be 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 [155; 156].

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 [157]. 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 [52; 49; 13]. In particular, simulations indicated that the majority of human proteins were uniquely identifiable if cysteine and lysine residues were orthogonally labeled and read out sequentially [49].

Several novel protein fingerprinting methods based on the read-out 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 [158; 159]. This linearization can be achieved by translocating the polypeptide chain through a nanopore [156] or by using a fluorescently labeled motor protein [49] to recognize 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 [12]. 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 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 task 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 [31]. 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 312 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 X fingerprinting allows for the reliable identification of proteins in complex mixtures.

6.2 Approach

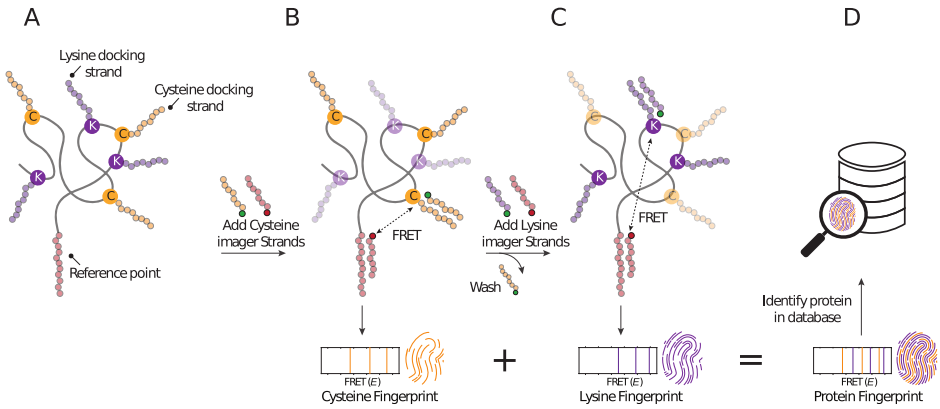


Figure 6.1: **(A)** A subset of amino acids (here cysteines and lysines) are labeled with orthogonal DNA sequences that function as docking sites for complementary, fluorescently labeled imager strands. Another orthogonal DNA sequence is conjugated to one of the protein termini, which serves as an acceptor 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 lysines. 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. FRET

6.2.1 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 [160; 161]. 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 [162; 163]. 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 6.1 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 6.1A). One of the protein 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 X 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 6.1B) 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 simultaneously 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 X fingerprint can be constructed, reporting on the distance of each target amino acid to 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 6.1C). The FRET X cycle can be repeated for any number of different amino acids, as long as they are labeled with orthogonal DNA docking 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 from a reference database (Figure 6.1D).

6.2.2 Fingerprinting simulations

The usefulness of our method hinges on its ability to discern FRET X fingerprints derived from many different proteins, and we run simulations to assess this. 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 [164; 165] 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 modeling approach [166], in which each residue is represented as a single pseudo-atom, restricted in space to only occupy the vertices of a lattice (Supplementary figure 6.S7). Structures are assigned an energy which is lower for structures more likely to occur in vitro. Pseudo-atoms may interact with the solvent or with pseudo-atoms on adjacent vertices, incurring either energy bonuses or penalties depending on the residue types involved. A structure can then be efficiently energy-minimized

using a Markov chain Monte Carlo process. That is, random modifications to the structure are proposed (Supplementary figure 6.S9), and for each modification the incurred change in energy determines the probability of accepting it. Despite their simplicity, past investigations have shown that lattice models can reproduce native protein folding behavior [167–169; 41; 170].

The attachment of DNA tags to selected residues, as required to accurately model our approach, has not previously been included in lattice models. Coarse-grained models have been used to study the effect of dyes linked directly to residues using short linkers, which were found to be minor [171; 172]; however, the additional effect of the longer, bulkier DNA tags on structure may be more significant. Although data on DNA-tag-protein interaction is lacking, we find that implementation at the coarse granularity required by lattice models may be built on two basic assumptions: that tags require sufficient unoccupied space to avoid steric hindrance and that they repel each other if situated closely together. Indeed, similar assumptions may be found in other models of ssDNA interaction [173]. A residue marked as tagged loses its ability to interact with other residues and is outfitted with a long, bulky side chain (Supplementary figure 6.S11), which incurs heavy energy penalties for clashes with the main structure and attempts to orient itself away from nearby tags. 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 X fingerprints associated with different residue types. Analogously to experimentally obtained fingerprints, simulated FRET X fingerprints for several residue types are then combined to serve as features for automated classification algorithms.

The simulation and classification procedures are described in more detail in the methods section.

6.3 Results

6.3.1 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 6.2A). 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 6.2A). The donor and acceptor imager strands were designed to exhibit a dwell time of ∼2s (Supplementary figure 6.S2), so that dyes could be frequently replenished.

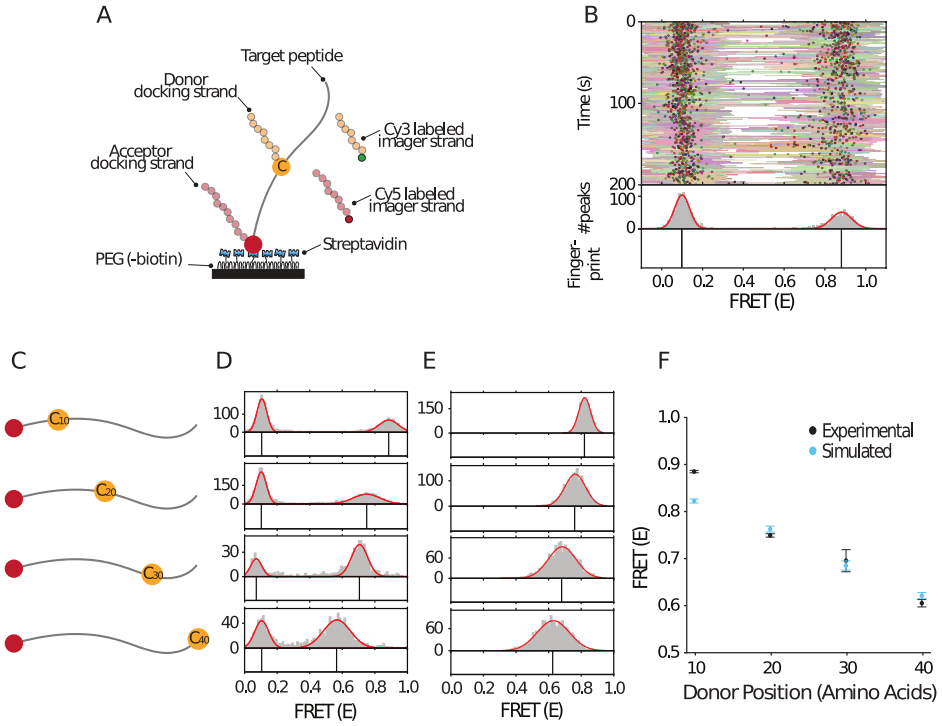


Figure 6.2: **(A)** Depiction of the experimental system for peptide fingerprinting. The target peptide is immobilized through conjugation of its N-terminal biotin with the streptavidin on the PEGylated surface. The donor (Cy3)-labeled imager strand (yellow) can bind to the DNA-docking site on the cysteine, 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 themean FRET efficiency from all data points in a binding event (dots) are indicated as a function of time. A Gaussian distribution (0.88 ± 0.14) 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. **(C)** Our four model peptides have a lysine at the N terminus and a cysteine at position 10, 20, 30 or 40. See Table S1 for the full amino acid sequences of the model peptides. **(D)** Experimental distributions and fingerprints for each peptide show a downward trend in mean FRET **(E)** for increasing FRET pair separation (mean ± FWHM of the Gaussian fit: 0.89 ± 0.14, 0.75 ± 0.20, 0.72 ± 0.11, 0.57 ± 0.20). See also Figures S2 and S3 for imager strand dwell times and kymographs for single peptides, respectively. **(E)** The simulated distributions and fingerprints for the four peptides show a similar downward trend in distribution means (0.82 ± 0.08, 0.76 ± 0.15, 0.68 ± 0.20, 0.62 ± 0.23). **(F)** Experimental and simulated data correlate well. Whiskers denote ± one standard deviation. Standard deviation of experimental data points is over four kymographs (each consisting of hundreds of events). Experiments were performed on separate days.

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 6.2B). The FRET efficiency for each data point (Figure 6.2B, lines) and the mean efficiency per binding event are calculated (Figure 6.2B, 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 [161], which together constitute the fingerprint of the peptide (Figure 6.2B, 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 6.2C). First, we performed single-molecule experiments to obtain experimental FRET X fingerprints and found a clearly discernible peak for each peptide (Figure 6.2D and Supplementary Figure 6.S3). Then we simulated FRET X 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 6.2E). While each histogram showed a wide distribution (FWHM of $\sim$0.1-0.2, Figure 6.2D and E), the Gaussian fit can be used to resolve the peak with high precision of $<$0.01 (standard error of mean), where the achievable precision depends on the number of binding events [162]. Furthermore, 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 6.2F). 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 X fingerprint.

6.3.2 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 [176] 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 [176; 177]. 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 6.3A 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 6.3C). Fingerprints do vary across individual molecules of the same spliceoform; however, the fingerprints remain sufficiently characteristic

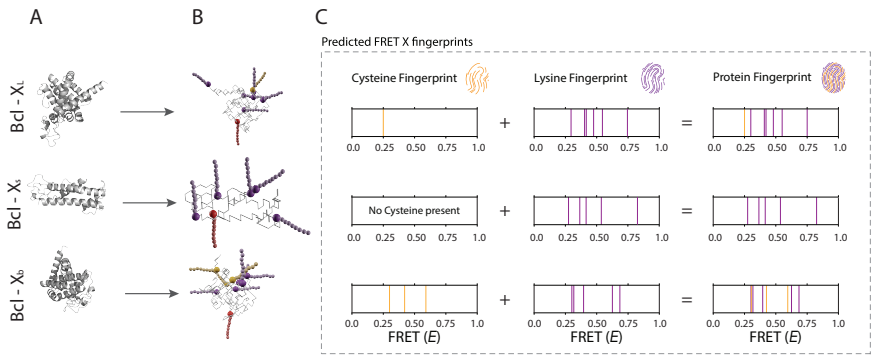


Figure 6.3: **(A)** Fully atomic structure for BCL XL, Xs, and Xb (from top to bottom) as predicted by the RaptorX structure prediction tool [174; 175]. **(B)** Energy-optimized lattice model structures with DNA-docking strands attached to cysteines (orange) and lysines (purple). The reference acceptor 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 (distributions shown in Figure S4). Fingerprints for a second set of spliceoforms (PTGS1) are shown in Figure S5.

to identify each spliceoform by eye (Supplementary figure 6.S5A). 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) [178]. 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 (Supplementary figure 6.S5B).

6.3.3 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 = 312). 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 measure overall classifier accuracy. To identify the subset of proteins for which our method works well, we also analyze the number of well-identifiable

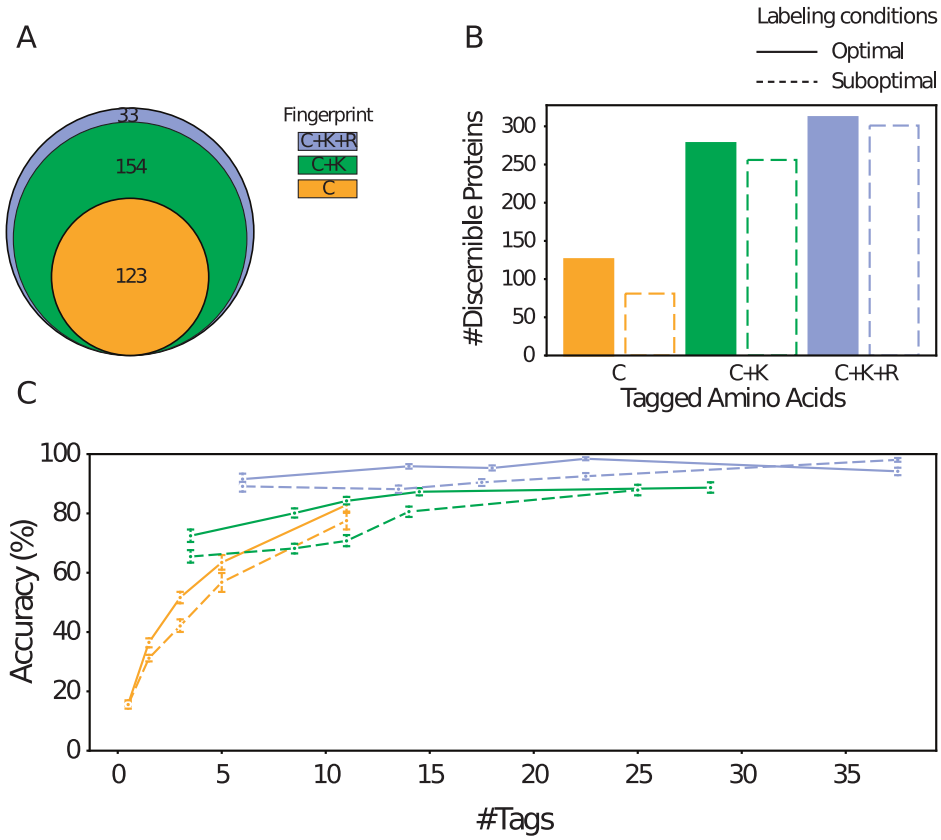


Figure 6.4: FRET X fingerprint classifier cross-validation performance measures are shown for three combinations of tagged residue types, C, C + K, and C + K + R, and two labeling qualities, "optimal", where all targeted residues and no off-target residues were labeled, and "suboptimal", where erroneous labeling occurred following the rules in Table S3. **(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 312. **(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 Gone standard deviation. Accuracies for different resolutions and suboptimal labeling scenarios are shown in Figure S6.

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 312 proteins, 123 were well-identifiable, indicating that labeling only C-residues is sufficient to consistently recognize this subset of proteins (Figure 4A, orange circle). 57 proteins did not contain C residues and are thus impossible to identify using only C-labeling. The remaining 132 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 4B). 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 277 and 310 out of 312 respectively. Regardless of which residue types are labeled, we find that proteins containing more tagged residues can be identified with higher accuracy (Figure 6.4C).

6.3.4 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 6.S3). For C and K these probabilities were based on experimentally determined efficiencies and specificities found in literature [179–181].

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 X fingerprints - particularly those gained from C+K+R labeling - contain the redundant information required to mitigate the effect of imperfect labeling (Figure 4C). 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 (Supplementary figure 6.S6).

6.4 Discussion

Here we present a 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 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 [182], while the FRET dyes (Cy3-Cy5) used here are expected to be accurate at distances of up to ∼7 nm [31]. Therefore, our FRET X 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 X 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 [156; 179–181]. Interestingly, even if only cysteine is labeled we find that a considerable 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 [183; 184] may be employed to further increase accuracy or tailor our method to the detection of a given target protein.

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 [154]. 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 [185]. 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.

6.5 Limitations of study

Although care has been taken to account for the effects of our experimental method on target protein structures, and thus the produced fingerprints, we note that the nature of several potentially influential factors have yet to be elucidated. Importantly, 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. Furthermore, although we model the effects of lower labeling efficiency and specificity, we have insufficient information to model how adjacency of residues targeted for labeling will affect efficiency of labeling chemistries. Once proteins can be fingerprinted more routinely, more data will be available to support modeling choices accounting for these factors. We stress that it is primarily the uniqueness and reproducibility of a fingerprint that is important for protein identification, not necessarily its predictability from

a known structure. While our current simulations were performed on a set of 312 known protein structures, we envision that the number of proteins that can be fingerprinted using our FRET X approach will increase significantly due to recent developments in protein structure prediction tools [165; 186; 187]. 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 be important for more targeted approaches.

6.6 Methods

6.6.1 Peptide Labeling

Custom designed polypeptides were obtained from Biomatik (Canada) and had a constant backbone sequence (see Table S1), 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 Tables S1 and S2 for the full list of substrates.

6.6.2 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).

6.6.3 Single-Molecule Data Acquisition

Single-molecule flow cells were prepared as previously described[188; 126]. 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 $\pm$ 2 °C).

6.6.4 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[183]. 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.

6.6.5 Simulations

Fingerprinting simulations were generated using a lattice folding model written in Python 3.7. Simulation and analysis code are freely available at `https://github.com/cvdelannoy/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[164–168]. 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 (E_{tot}), as calculated by an energy function. Modifications that decrease E_{tot} are accepted, whereas those that increase E_{tot} are more likely to be discarded the more they increase E_{tot}. 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.

6.6.6 Lattice structure

The lattice modeling procedure employed here largely resembles those in previously published applications[167]. In particular, the model developed by Abeln *et al.*[167] was used as a starting point, however the cubic lattice was replaced by a novel body-centered cubic (BCC) lattice (Figure S6). 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[185]. 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.

6.6.7 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 S7). 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 20Å of each other.

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.*[13]; 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 S3).

6.6.8 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 C_α positions and placing each C_α-atom on the nearest lattice vertex, while remaining connected to its neighboring C_α-atom, starting from the residue with the lowest index. Alpha helices are forced to remain intact on the lattice, by first translating involved C_α-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 [189].

6.6.9 Folding simulation

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

$$E_{tot} = E_{AA} + E_{sol} + E_{SS} + E_{tag} + E_{req}$$

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[190]. The secondary structure formation energy term E_{SS} is adapted from Abeln et al.[167] 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.

6.6.10 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 S7). 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 54Å for the Cy3-Cy5 FRET pair [31]). 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[162]. If multiple residue types are tagged, each residue type generates its own fingerprint which is binned separately.

6.6.11 Classification

To classify simulated fingerprints a support vector machine (SVM) was implemented using the scikit-learn package (v0.23.2)[189]. 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.

Acknowledgements

We thank Sung Hyun Kim for fruitful discussions and feedback. C.J. and D.R. acknowledge funding from NWO-I, the Netherlands Foundation of Scientific Research Institutes (formerly FOM), part of the Dutch Research Council, under grant SMPS.

Author contributions

C.L and M.F. contributed equally to this work. C.L, M.F., C.J., and D.R. initiated and designed the project. C.L. and D.R. developed the simulation approach, C.L. ran the simulations and analyzed of simulation results. D.R. supervised simulations. M.F., R.W., and C.J. designed the wet lab experiments. M.F. and R.W. performed the wet lab experiments and analysed the data. C.J. supervised the wet lab experiments. C.L. and M.F. wrote the first draft of the manuscript, all authors read and improved the manuscript.

CHAPTER 7

Chop-n-Drop: *In silico* assessment of a novel single-molecule protein fingerprinting method employing fragmentation and nanopore detection

This chapter has been published as:

Carlos de Lannoy, Florian Leonardus Rudolfus Lucas, Giovanni Maglia and Dick de Ridder. "*In silico* assessment of a novel single-molecule protein fingerprinting method employing fragmentation and nanopore detection" *iScience* 24 (2021): 103202

Supplementary material available at:

`https://doi.org/10.5281/zenodo.6773393`

Abstract

The identification of proteins at the single-molecule level would open exciting new venues in biological research and disease diagnostics. Previously we proposed a nanopore-based method for protein identification called chop-n-drop fingerprinting, in which the fragmentation pattern induced and measured by a proteasome-nanopore construct is used to identify single proteins. In the simulation study presented here, we show that 97.1% of human proteome constituents are uniquely identified under close to ideal measuring circumstances, using a simple alignment-based classification method. We show that our method is robust against experimental error, as 69.4% can still be identified if the resolution is twice as low as currently attainable and 10% of proteasome restriction sites and protein fragments are randomly ignored. Based on these results and our experimental proof-of-concept, we argue that chop-n-drop fingerprinting has the potential to make cost-effective single-molecule protein identification feasible in the near future.

Availability: All generated data, ion current measurement data and analysis code was deposited at `https://github.com/cvdelannoy/chop_n_drop_simulation`, and on Zenodo at `https://zenodo.org/record/5116022`.

7.1 Introduction

Over the past decades, mass spectrometry (MS) has allowed for ground-breaking discoveries in proteomics, enabling such impressive feats as the definition of a human protein atlas [191] and large-scale screening for protein disease biomarkers [192]. However, not all protein-related research questions may be addressed by MS. Examples are found in the nascent field of single-cell proteomics which, following the example of single-cell transcriptomics, is expected to give unprecedented insight into cell functioning and pathology [193]. While MS has already made strides in this field by enabling the detection of proteins present at thousands of copies per cell [45], some important and clinically relevant proteins such as signaling molecules and transcription factors are expected to be present in the range of dozens of copies [194]. The development of novel single-molecule protein identification methods is therefore necessary to unlock the true potential of single-cell proteomics.

In the search for single-molecule alternatives to MS, two main venues are currently being explored. On the one hand, conceptual methods utilizing the read-out of fluorescent dyes attached to a subset of residue types have shown promising results [49; 12; 13]. However, methods using fluorescence-based readout strategies require efficient and specific labeling of residues. Optimizing labeling strategies is non-trivial [e.g. 180; 195] and less-than-perfect labeling may decrease accuracy, thus a label-free method would be preferred.

On the other hand, unlabeled proteins may be analysed using a nanopore, over which an electrical potential is applied; as a protein is passing through the pore, changes in electrical resistance may give information on the protein's properties [196]. Proteins may be analysed in their folded states [197–201], which is relatively

straightforward but does not provide sufficient information to discriminate between similarly shaped and charged proteins. Furthermore, no single pore aperture size is suitable for proteome-wide analysis due to the wide variety of protein sizes found in nature [202]. Alternatively, proteins may be unfolded and threaded single-file through the pore using a molecular motor [158; 51]. This approach allows for finer interrogation of the residue sequence and may analyse proteins of any size using a single pore aperture size.

In prior work, we showed that engineered complexes of heptameric nanopores and proteasomes can be readily assembled without loss of proteasome activity or electrical conductance of the pore [203]. Furthermore, we have shown that residual current through FraC pores correlates well with the molecular weight of passing protein fragments in the 500 to 1600Da range [204]. Presumably, this is because weight is correlated to properties that directly influence residual current, such as fragment size, shape and charge [205; 206; 200; 207–209]. We thus proposed that proteasome-nanopore constructs can be used to identify proteins, in a conceptual method dubbed chop-n-drop fingerprinting [203]. An unknown protein can be processed terminal-to-terminal by the construct, cleaving it at proteasome target sites, after which the molecular weight of sequentially released fragments can be estimated based on the residual electrical current as they pass through the nanopore. The sequence of measured fragment weights can then serve as a characteristic signature – a fingerprint – of the protein. Once proven, this fingerprinting method can easily be implemented in a highly parallel fashion by adapting existing hardware that was developed for nucleic acid sequencing. Compared to both MS and existing fluorescence-based measurement equipment, this hardware is inexpensive and has a small benchtop footprint, thus opening up opportunities for field diagnosis and in-house analysis for even small laboratories. It is as of yet however unclear whether chop-n-drop fingerprints are sufficiently characteristic to identify a single protein in highly complex mixtures.

Here we present a computational analysis of the chop-n-drop method, in which we show that simulated fingerprints of all proteins in the UniProt human proteome can be accurately classified using a simple alignment-based method. Considering these and previously published experimental results, we argue that chop-n-drop fingerprinting is a promising concept for cost-effective single-molecule protein identification.

7.2 Results

7.2.1 Simulation and classification method

To estimate the performance of the chop-n-drop fingerprinting method on a highly complex protein identification task, we developed a simulation pipeline mimicking the experimental procedure, including several sources of biological and technical noise that we expect to encounter (Figure 7.1).

In essence, the chop-n-drop fingerprint of a protein only consists of a sequence of weights, which are deduced from pore current blockades caused by sequentially cleaved-off fragments passing through the nanopore. The simulation of this process

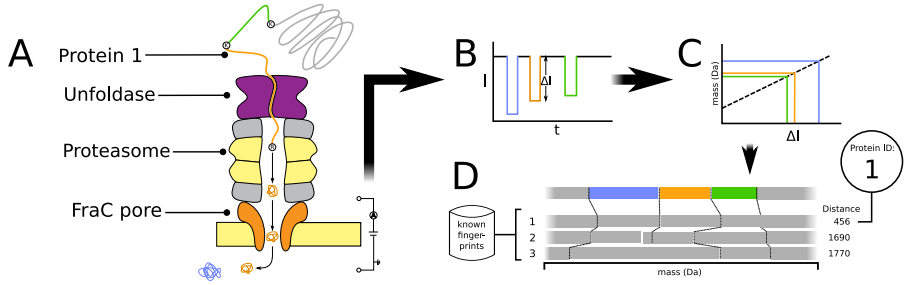


Figure 7.1: **Schematic overview of the chop-n-drop fingerprinting method.** A protein is unfolded by an unfoldase and fragmented by a proteasome directly introduced above a nanopore. The protease is engineered to lyse proteins at particular residues. **(B)** As the fragments pass the pore, a change in electrical current through the pore is measured. **(C)** The molecular weights of the fragments are estimated from the magnitudes of the current changes. **(D)** Finally the produced sequence of fragment weights is aligned to database fingerprints of known proteins, to identify the protein.

follows a straight-forward two-step process. First, akin to the proteasome cleaving a protein into fragments, we divide a given protein sequence into sub-sequences by splitting it at the proteasome's target sites. We assume here that we can force it to exhibit only trypsin-like behavior, by mutating proteasome subunits exhibiting chymotrypsin- and caspase-like activities, while leaving its subunits exhibiting trypsin-like activity intact [210; 211]. To account for the fact that the proteasome will likely fail to cleave at a fraction of target sites, we only cleave each target site with a certain probability, which we refer to as the proteasome efficiency (e_p).

Subsequently we mimic the passing of fragments through a heptameric FraC pore, the readout of the current blockade and the estimation of the fragment weight, by simply translating the sub-sequences into corresponding fragment weights. Although weights can be calculated from sequences with high accuracy, experimental measurements may be less accurate and marked by a given resolution (r), the smallest detectable weight difference. In experimental setups, this parameter is dependent on pore and measuring equipment properties. The smallest weight difference we have detected with FraC so far is 4Da (Supplementary figure 7.S1), thus in simulations we consider r-values above 4Da attainable. To account for resolution in simulation, Gaussian noise is added to fragment weights, where the standard deviation of the noise is related to r (see Methods). Fragments weighing less than 500Da are removed, as they typically escape detection of heptameric FraC nanopores [204]. Furthermore, as it has not been shown that the relation between weights above 1.6kDa and current blockades remains monotonic [204], all fragment weights larger than this value are reduced to 1.6kDa. Lastly, although we expect the seal between proteasome and pore to be extremely tight based on molecular dynamics simulations [203], fragments may fail to enter the pore after cleavage. We account for this by only retaining each fragment with a certain probability, which we refer to as the capture rate (C). Although C is

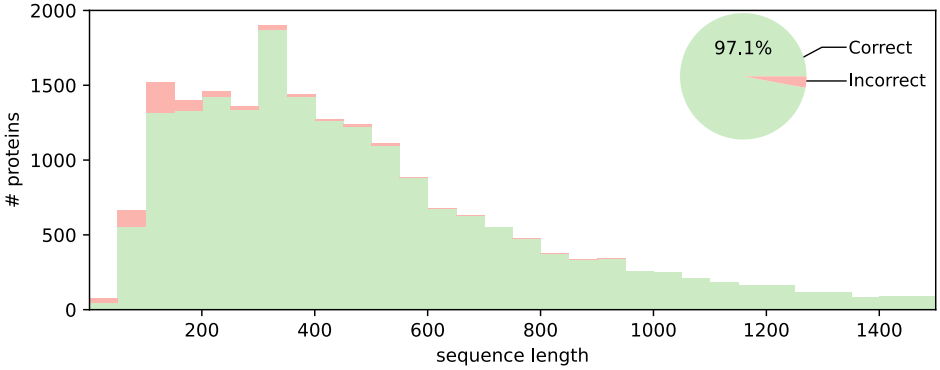


Figure 7.2: **Simulated fingerprint identification accuracy assuming low-noise conditions.** Cumulative histogram of correct and incorrect classifications of simulated chop-n-drop protein fingerprints for all human proteome constituents, assuming low noise parameters; resolution $r = 5$Da, capture rate $C = 0.99$ and proteasome cleaving efficiency $e_p = 0.99$. Numbers are shown distributed over sequence length (bars), and relative to the total number of proteins (pie chart). Alignment examples and sequence identity distribution for erroneous alignments are shown in Supplementary figures 7.S3 and 7.S4 respectively. Results assuming charge-dependent fragment capture are shown in Supplementary figure 7.S5A.

likely dependent on the size and charge of individual fragments, the relationship between these factors is unclear, thus we assume C to be constant. The resulting sequence of fragment weights returned by this process constitutes the fingerprint for a protein.

We used fingerprints generated using our pipeline to develop a classification method, which assigns a protein identity to a given fingerprint. We follow an alignment-based approach, where a query fingerprint is aligned to a database of previously generated fingerprints, using a custom dynamic programming implementation (Supplementary figure 7.S2). The database fingerprint that is most similar to the query fingerprint is assumed to have come from the same protein.

7.2.2 Simulations under low-noise conditions

We ran our simulation pipeline and classification method on all sequences in the UniProt human proteome ($n = 20,395$). Under close to ideal simulated noise parameters ($e_p = 0.99$, $r = 5.0$Da, $C = 0.99$) we find that our alignment based approach retrieves the correct identity for 97.1% of fingerprints (Figure 7.2). Inspection of made alignments shows that our algorithm correctly handles missing and fused fragments (Supplementary figure 7.S3A). 77% of misclassifications occurs for shorter proteins, under 250 residues in length. Of misclassified fingerprints, 42% shows more than 80% amino acid sequence identity to the protein as which it was wrongly identified, indicating that the resolution of 5Da assumed here is insufficient to consistently separate such similar entities (Supplementary figure

7.S4). Upon inspection of these cases, we find that many misclassifications were in fact mix-ups between paralogous sequences. The remaining misclassifications are caused by chance alignments with different fingerprints (Supplementary figure 7.S3B). This is expected to occur more often if a protein is shorter, as it will generally produce a fingerprint of fewer elements, which is less likely to yield a unique pattern.

7.2.3 Simulations under high-noise conditions

We subsequently probed how resistant chop-n-drop fingerprinting is to higher levels of experimental noise, by varying one noise parameter at a time while keeping all others near their low-noise values ($e_p = 0.99$, $r = 5.0$Da, $C = 0.99$). To keep computations tractable these simulations were run on a random subset of 200 query sequences, while the reference database still contained all UniProt sequences so that the classification task was no less challenging. In each case we find that accuracy deteriorates gracefully with parameter value (Figure 7.3A). Interestingly, we still attain an accuracy of 90.9% at a resolution of 50Da, which is worse than the 44Da resolution we reported previously [204] and more than tenfold worse than the current-best resolution of 4Da, as reported in this work (Supplementary figure 7.S1). Similarly, we find that a lower proteasome efficiency or capture rate of 90% still results in 90.7% and 87.1% accuracy on average respectively. We then repeated the simulation on the entire dataset with all noise parameters at high-noise values ($e_p = 0.90$, $r = 10.0$Da, $C = 0.90$). Even under these circumstances, we find that 69.4% of proteins are correctly classified (Figure 7.3B). Here too, it should be noted that most incorrectly classified proteins were of lower sequence length.

Finally, we investigated the effect of fragment charge on accuracy. In FraC pores, Electro-osmotic flow (EOF) is sufficiently strong to overcome an opposing electrophoretic force (EF) to some degree, so that even negatively charged fragments are pulled through the pore towards an anode at the trans side. However fragments carrying larger negative charges increase the EF so that the EOF can no longer cancel it out, and thus such fragments cannot enter the pore. We investigated the effect of the omission of fragments that are too negatively charged ($< -1e$) at FraC's operating pH-levels (pH=4.0) [204] and found that the effect on accuracy is negligible in both low-noise and high-noise scenarios (97.1% and 69.3% respectively, Supplementary figure 7.S5).

7.3 Discussion

Single-molecule (SM) protein fingerprinting holds great promise to revolutionize biological research and diagnostics [155]. We have previously proposed that this may be accomplished using a proteasome-nanopore construct, which cleaves a target protein into fragments and subsequently reads out the fragment weights [203]. Here we present simulation results indicating that the produced sequence of fragment weights contains sufficient information to identify a protein.

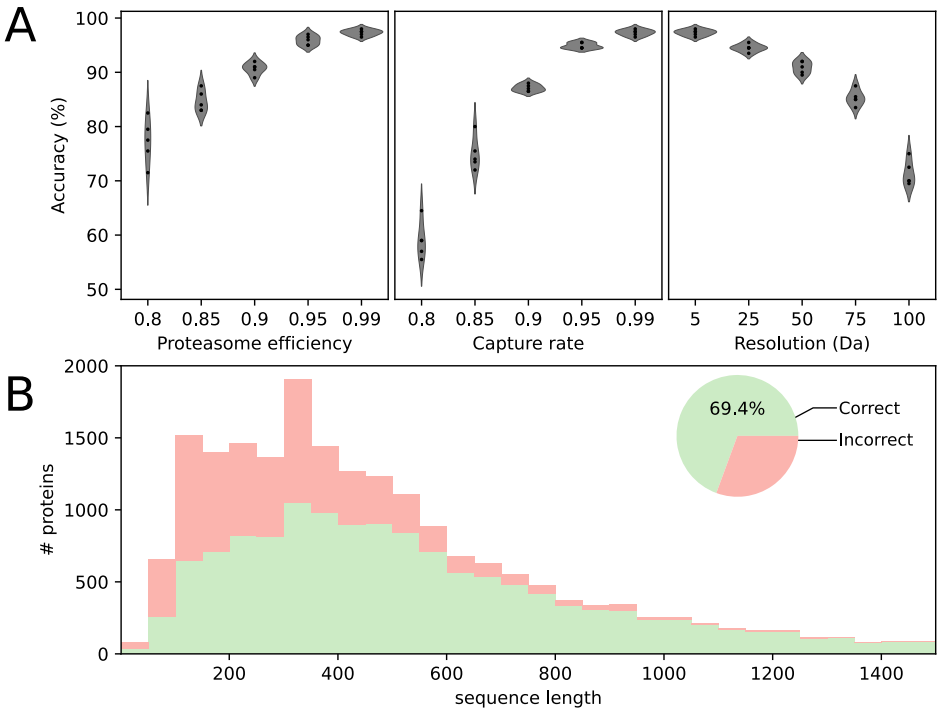


Figure 7.3: **Simulated fingerprint identification accuracy under noisy conditions. (A)** Fingerprint classification accuracy over a range of noise parameter values; resolution (left), capture rate (mid) and proteasome efficiency (right). For each case the unvaried noise parameters are set to low-noise values (capture rate $C = 0.99$, resolution $r = 5.0$Da and proteasome efficiency $e_p = 0.99$). Five replicates were generated for each parameter combination. **(B)** Cumulative histogram of correct and incorrect classifications of simulated chop-n-drop protein fingerprints for all human proteome constituents, assuming more realistic noise parameters; $r = 10$Da, $C = 0.90$ and $e_p = 0.90$. Numbers are shown distributed over sequence length (bars), and relative to the total number of proteins (pie chart). Results assuming charge-dependent fragment capture are shown in Supplementary figure 7.S5B.

The hypothetical construct investigated in our simulations consists of a heptameric dihelical FraC-pore, which we show to be well-suited to detect differences between fragment weights at high resolution, and a proteasome exhibiting trypsin-like cleaving activity. In prior work, we have shown that constructs of artificial heptameric beta-barrel PA-pores and proteasomes can be built without loss of function of either component. Given the structural similarity, we are confident that our hypothetical construct can be built in a similar way, by replacing PA-pore monomers with FraC monomers and by making use of engineered proteasomes [210; 211].

In the presented simulations, we included sources of noise that may hamper fingerprint measurements in practice. We assumed that the proteasome may not cleave each target site, that weight measurements may be inaccurate up to a given weight resolution and that not all cleaved-off fragments may be caught in the nanopore. Assuming higher noise parameter settings – a fragment capture rate and proteome efficiency of 90%, with a measurement resolution of 10Da – for each of these noise sources, we find that overall accuracy remains sufficiently high at 69.4%. As accuracy increases with protein length, we find that chop-n-drop fingerprinting should be particularly suitable to identify larger proteins.

Over the past years, the obstacles on the road toward SM protein fingerprinting have been attacked vigorously from multiple angles, with several groups showing promising initial results and proofs-of-concept. While each proposed method has shown particular strengths, we argue that chop-n-drop combines several properties not found together in other methods. First, unlike fluorescence-based methods [12; 49; 13] it does not require the implementation of any labeling chemistries as properties of the target protein are read out directly, thus evading issues with erroneous labeling and simplifying sample preparation. As a trade-off, fluorescence-based methods are more sensitive to differences between proteoforms as long as the difference involves the position or presence of a targeted residue type. As we show here that even at high resolution our method misclassifies proteins with high sequence similarity to other entries, it is likely that differences between highly similar proteoforms may also remain unnoticed.

Different methods based on the readout of folded proteins by electrical current blockage of a nanopore have been proposed as well [198; 212; 199]. These were unable to analyse a wide range of protein sizes however; as the pore lumen needs to be of an appropriate volume for the analysis of a given protein size, a single nanopore is not able to detect minute differences in both small and large proteins. Here this problem is mitigated by the fragmentation step.

Most importantly however, the hardware required to implement chop-n-drop fingerprinting in a highly parallelized setting can be readily borrowed from commercial platforms for DNA sequencing using nanopores, which are inexpensive and have already been miniaturized to a handheld format. As such we envision that our method could soon fill a niche that no other method currently can; that of small-scale, in-house single-molecule protein identification.

In conclusion, we provide evidence that chop-n-drop fingerprints can provide sufficient information to identify proteins in complex samples, and present a suitable alignment-based classification method. Upon optimization of the finger-

printing procedure, we envision that our method may see practical implementation in the near future.

7.4 Limitations of study

Our simulation builds on the assumption that fragment weight is correlated to the residual current measured while the fragment passes the nanopore. Indeed, we have previously shown that this is the case for fragments weighing between 500 and 1600Da [204]. However it should be noted that rather than the fragment's weight, its volume, shape, charge, hydrophobicity and interactions with the pore interior directly influence residual current [205; 206; 200; 207–209]. As these properties are more difficult to model, and considering how they apparently correlate to weight sufficiently well to in turn correlate weight and residual current [204], we consider using weight in simulated fingerprints justifiable. Once the experimental methodology has been further developed and protein fingerprints can be measured more routinely, we can define the relation between these properties and the residual current in more detail to predict fingerprints in a more robust manner.

The existence of different proteoforms, which was not accounted for in this simulation, presents both an opportunity and a challenge to chop-n-drop fingerprinting. Through alternative splicing and post-translational modification (PTM), multiple proteoforms with different functions may be generated from the same gene [154]. Depending on the spliceoform or the PTM types present, different proteoforms may generate distinct fingerprints. This allows their individual identification at SM resolution, which is an important potential application of SM analysis, but also adds tens of thousands of potential fingerprint patterns, which further complicates the task of fingerprint classification. A solution may be to fractionate samples prior to chop-n-drop analysis, after which each fraction may be analysed using a dedicated classifier which only considers the proteoforms that could be present in a given fraction.

Acknowledgements

This work was supported by the Vrije Programma of the Foundation for Fundamental Research on Matter under grant number 16SMPS05.

Author contributions

C.L. developed and wrote the code for simulations, D.R. supervised simulation development, F.L.R.L. designed and executed practical experiments and performed experimental data analysis, G.M. conceived the analysis method and supervised practical experiments, C.L. wrote the manuscript, all authors provided feedback on the manuscript.

Declaration of interests

G.M. holds a patent for biopolymer sensing and sequencing based on FRAC Actinoporin (patent number US20190292235A1) and has filed a patent for single-molecule recognition by chop-n-drop.

Data and code availability

- Ion current measurement data have been deposited at Github and are publicly available as of the date of publication. DOIs are listed in the key resource table.
- All simulation and analysis code has been deposited at Github and is available as of the date of publication. DOIs are listed in the key resources table.
- For fingerprinting simulations we used all available sequences in the UniProt human proteome (UP000005640).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

7.4.1 Methods

Fragaceatoxin C purification

Plasmid DNA containing the sequence for His6-tagged Fragaceatoxin C (FraC) with a phenylalanine modification on position G13 was transformed into electro-competent E. cloni ®EXPRESS BL21 (DE3) strain cells (Lucigen) for protein expression. Transformed cells were plated on Lysogeny broth (LB) plates containing 100 µg/mL ampicillin and incubated overnight at 37 °C. A single colony was cultured in LB medium containing 100 µg/mL ampicillin and finally protein expression was induced using 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Bacteria were harvested, treated with 0.2 units/mL DNase I and 20 µg/µL lysozyme, and sonicated, after which the mixture was incubated with Ni-NTA beads. FraC monomers were washed from the beads using elution buffer containing 0.300 M imidazole. Sphingomyelin (porcine, brain):1,2-Diphytanoyl-sn-glycero-3-phosphocholine (DPhPC) liposomes for FraC oligomerization were created by solubilizing equal weight parts sphingomyelin and DPhPC in a pentane-ethanol mixture and evaoporating the solvent to create a lipid film. The film was then solubilized in 0.015 M Tris-HCL dilution buffer (pH 7.5, 0.150 M NaCl), briefly sonicated, frozen at −20 °C and thawed. Liposomes and FraC monomers were mixed at a mass ratio of 10:1 solubilized in 0.6% (w/v) N,N-Dimethyldodecylamine-N-oxide (LDAO) and diluted in 0.02% (w/v) Dodecyl-β-D-maltoside (DDM). Finally oligomerized FraC was bound to Ni-NTA beads and eluted in oligo elution buffer (0.200 M Na_2EDTA, pH 8.0, 0.02% (w/v) DDM). This protocol was previously described in [213]

His6-tagged Fragaceatoxin C sequence

MASADVAGAVIDFAGLGFDVLKTVLEALGNVKRKIAVGIDNESGKTWTA
MNTYFRSGTSDIVLPHKVAHGKALLYNGQKNRGPVATGVVGVIAYSMS
DGNTLAVLFSVPYDYNWYSNWWNVRVYKGQKRADQRMYEELYYHRS
PFRGDNGWHSRGLGYGLKSRGFMNSSGHAILEIHVTKAGSAHHHHHH

Electrophysiological recordings

For planar lipid bilayer electrophysiology measurements, a constant voltage of -70 mV was applied over a single nanopore, and a buffer solution containing 1M KCl buffered to pH 3.8 using 50mM Citric acid titrated with bis-tris-propane was used. [Met5]-Enkephalin and [d-Ala2][d-Leu5]-Enkephalin (Sigma Aldrich) were added in 10 µM concentration. Ionic currents were recorded using an Axopatch 200B amplifier coupled with a Digidata 1440a or Digidata 1550B A/D converter (Molecular Devices). All data was recorded using Clampex 10 (Molecular Devices) with a sampling frequency of 50 kHz and an analogue Bessel filter of 10 kHz. Measurement procedures were first described in [50].

Electrophysiological recording analysis

Data was analysed using Python 3.7 and is contained within a Jupyter notebook, available in the Github repository noted in the key resources table. Events from translocating peptides were characterised using a threshold search algorithm combined with a generalised flat-top normal distribution fit [50]. The resolution between the two peptides was determined using the difference between the peak centres and their standard deviation [50].

In silico fingerprint generation

Code for *in silico* fingerprint generation and classification was wrtten in Python 3.8 (Python Software Foundation, `www.python.org`). We generate *in silico* chop-n-drop fingerprints by splitting protein sequences at protease target sites and calculating the weights of the resulting fragments from their sequences. We assume that fragments of a weight lower than 500Da are undetectable, thus these fragments are removed from fingerprints. Fragments of a weight larger than 1.6kDa are set to 1.6kDa, as prior investigations showed that the relationship between weight and current blockage is not monotonic above this weight [204]. In simulations where fragments were selected based on charge, the fragment charge was calculated using biopython (v1.78), which employs a concentration ratio-based calculation [214].

Three parameters are set to represent different noise sources; capture rate C, proteasome efficiency e_p and resolution r. The capture rate denotes the fraction of fragments that enters the pore after lysis and is measured. In our simulations each fragment is retained with a probability of C. The proteasome efficiency denotes the fraction of target sites at which the proteasome cleaves. In simulations, each target site has a probability of e_p of being cleaved. Note that a failure to

cleave will result in two fragments being fused together, after which they remain represented in the fingerprint as the sum of their weights. Finally, the resolution denotes the minimum difference in fragment weight that can still be detected by current blockage, expressed in Da. We adhere to an experimentally found minimum resolution of 4Da (Supplementary figure 7.S1). In our simulations, the resolution is represented by the magnitude of Gaussian noise added to fingerprint weights. Specifically, we define the standard deviation of the distribution from which a noise value n is drawn such, that the probability of a fragment size measurement deviating r or less from its actual size is fifty percent:

$$P(n \leq r) = 0.5 \tag{7.1}$$

The standard deviation enforcing this resolution r, σ_r, can be found using the Z-score formula, in which the Z-score is calculated using the inverse cumulative distribution function of the standard normal distribution, Φ^{-1}:

$$\sigma_r = \frac{-r}{\Phi^{-1}(0.5/2)} \tag{7.2}$$

As resolution is expressed as a positive number and equation 7.2 considers the lower tail of a distribution centered at 0, the resolution is multiplied by -1.

Simulation and classification

We ran *in silico* digestions on all sequences in the UniProt human proteome (UP000005640). To compile a database of fingerprints with known identity, we first performed an *in silico* digestion under noiseless circumstances (i.e. $C = 1.0$, $e_p = 1.0$ and $r = 0.0$Da). Then we ran several subsequent digestions for a range of values for C, e_p and r. Fingerprints from these runs were classified by aligning them to database fingerprints obtained from noiseless digestions.

We gauge the similarity of query and database fingerprints by aligning them using a dynamic programming algorithm (Supplementary figure 7.S2). The dynamic programming table is filled as follows:

$$S(i,j) = \min \begin{cases} |X_i - Y_j| + S(i-1, j-1) \\ |X_i + X_{i-1} - Y_j| + S(i-2, j-1) \qquad i \leq |X|, j \leq |Y| \\ |X_i - Y_j| + S(i-2, j-1) + G \\ |X_i - Y_j| + S(i-1, j-2) + G \end{cases} \tag{7.3}$$

with the following conditions for edge cases to ensure that the alignment is global:

$$\begin{aligned} S(0,0) &= 0 \\ S(i,0) &= \infty \quad \forall \quad 1 \leq i \leq N_X \\ S(0,j) &= \infty \quad \forall \quad 1 \leq j \leq N_Y \end{aligned} \tag{7.4}$$

Here $S(i,j)$ is the distance between query and database fingerprints X and Y respectively, up to fragments X_i and Y_j and G is a gap penalty. N_X and N_Y are

the numbers of fragments in X and Y respectively. At each step in the alignment one of three actions may be taken. First, a single fragment of each fingerprint may be aligned, in which case the absolute difference of their weights is added to the total score. Second, two fragments of X may be aligned to one fragment of Y, corresponding to a missed proteasome target site. This action increases the score by the difference between the summed weight of the former and the single weight of the latter. Third, a gap may be introduced in either X or Y at the cost of a penalty. The gap penalty G is dependent on the resolution used during digestion:

$$G = (1.96 \cdot \sigma_r)^2 + L \tag{7.5}$$

Here σ_r is the resolution-dependent standard deviation of Gaussian noise added to fragment sizes during *in silico* digestion (equation 7.2) and L is the lower detection limit ($L = 500$ Da). This means that introducing a gap is preferred over matching fragments if the difference between fragment weights exceeds the difference expected in 95 percent of correct matches. The addition of L is required to ensure that a match is still preferred if a normally undetected fragment (i.e. of which the weight is under L) is fused to another fragment due to a missed proteasome target site.

A query fingerprint is classified by aligning it to all fingerprints in the database and assigning it the identity of the database fingerprint to which the distance is smallest.

CHAPTER 8

Discussion

Methods for single-molecule fingerprinting and sequencing of biopolymers (BPs) have been in the making for more than thirty years [58; 215]. With the first implementations achieving commercial success, the future of the field and its impact on scientific research is starting to take shape. As noted in the introduction, analysis of nucleic acid (NA) BPs typically bears fruit more readily than that of proteins, and SM sequencing/fingerprinting has proven to be no different in that respect. Both electrical (nanopore-based) and fluorescence-based sensing methods have already entered and revolutionized mainstream NA sequencing. In contrast, SM protein fingerprinting is still in its infancy, with practical applications only beginning to appear on the horizon. Although the main causes of this chasm are fundamental differences between the two BP types – the more complex nature of proteins and the lack of readily available mechanisms for threading and amplification in nature – specific lessons from SM NA analysis can still be used to streamline the development of SM protein analysis.

Computational biology and bioinformatics are in an excellent position to distill these lessons and provide guidelines for the further development of SM analysis; through simulations it may steer experimental design, and by developing analysis tools it makes SM data analysis tractable and accessible. This thesis concerned both tasks. I started with an exploration of the current state of SM NA analysis and introduced several new data analysis methods in that field. This was followed by two simulation studies, exploring the potential of concepts of new methods for SM protein analysis. In the remainder of this discussion, I gauge the gap between SM analysis of NAs and proteins. I will first discuss to what extent current methods for NA and protein analysis are comparable, and how differences between the two BPs influence their analysis at the SM level, with a particular focus on *in silico* method simulation and data analysis. Building on these insights, I lay out a path for the future of SM protein analysis and describe how computational biology and bioinformatics efforts could support development.

8.1 Lessons from nucleic acids for protein analysis

8.1.1 Complementarity of readout methods

The current-day landscape of SM NA sequencing is divided between electrical and optical readouts, represented by nanopore sequencing devices, mainly produced by Oxford Nanopore Technologies (ONT), and the single-molecule real-time (SMRT) sequencing devices employing fluorescence readout, produced by Pacific Biosciences (PacBio). Although ONT and PacBio appear to be staunch competitors, each company currently has a particular niche in which it outcompetes the other: nanopore sequencers are less expensive and much easier to miniaturize, while SMRT devices produce reads with fewer systematic errors. This division between cost and accuracy is inherent to differences in the readout methods. The one-dimensional electrical readout of a nanopore device is easier to acquire but only provides an indication of size, shape and charge of its target, as well as its interactions with the nanopore interior [205; 206; 200; 207; 209], with no guarantee that biologically relevant differences are represented. Optical readouts can be made

highly purpose-specific by targeting a key property with a fluorescently labeled probe, however acquiring the signal requires more costly measuring equipment, as well as the development and production of highly specific probes.

It makes sense that similar niches will start to take shape in SM protein analysis as well. I presented two methods in this thesis that illustrate this. The FRET X fingerprinting method presented in chapter 6 employs probes that orthogonally target cysteine, lysine and arginine residues, which was theoretically sufficient to correctly classify most tested proteins in complex mixtures, even assuming suboptimal labeling efficiencies and reduced FRET resolution. The Chop-n-Drop nanopore-based method presented in chapter 7 also performed well under low-noise circumstances, but classification accuracy rapidly decreased under suboptimal conditions. In terms of implementation cost however, Chop-n-Drop would easily outcompete FRET X fingerprinting; while the latter requires a sensitive system of specialized optics and chemical labeling procedures, Chop-n-Drop's implementation may be as simple as introducing a different nanopore construct in existing NA nanopore sequencing devices.

The readout method is an important factor in deciding how to tackle the challenges encountered when moving from SM NA to SM protein analysis. In the following sections, I discuss each of these challenges and some possible solutions for both optical and electrical readout methods.

8.1.2 Primary structure differences

A first obstacle in translating NA analysis methods to proteins is caused by the more complicated primary structure of proteins – the sequence of monomers. As may be expected, the increase in the number of monomer types – from four nucleotides to twenty amino acids – has major consequences for analysis complexity.

Chapter 2 helps us to put these consequences into context for sequential electrical readouts. There, I described the development of basecallers, the computational tools that convert a squiggle into an NA sequence. This task is made more difficult by the fact that more than one nucleotide at a time influences the pore current, thus 4^k k-mers should be distinguished, where k is the number of monomers occupying the pore constriction. For the CsgG pore used in ONT devices, it was found that $k = 5$ explained the obtained current signal best, thus 1024 current levels should be modeled. This problem could be adequately solved using hidden Markov models (HMMs), which assigned Gaussian distributions to each k-mer and made use of the fact that subsequent k-mers should overlap to produce a final consensus sequence [23]. Even better performance was reached later using neural networks which, among other advantages, could account for the fact that k is actually variable [24; 25].

For proteins read out in a similar fashion, Ouldali *et al.* have shown that individually, most residue types generate different current blockades when in sequence with seven arginine residues [201]. However, in a naturally occurring sequence 20^k residue combinations may occur, which quickly increases with k. The MspA pore used in [51] analyzes 8-mers, which would yield $2.56 \cdot 10^{10}$ combinations.

This complexity may be impossible to capture in the limited bandwidth of a 1-dimensional squiggle, which would mean that no algorithm would be able to naively translate squiggles into protein sequences. Fingerprinting may still be possible however. Brinkerhoff *et al.* have shown that a single-residue difference in peptides was detectable using their DNA-tethered readout [51], thus it is certainly possible that a squiggle can contain sufficient information to detect some salient short sequences. If individual NNs could be trained for several subsequences, an array of NNs may be used to detect longer sequences as demonstrated for NAs in chapter 5.

If true *de novo* sequencing of proteins is to be achieved however, approaches employing orthogonal readouts of multiple properties have a higher chance of success. Indeed, several such methods are in development. For instance, Reed *et al.* describe a system in which color, pulse duration and cleaving time all inform on the identity of N-terminal residues, so that reading all three out gives a unique combination for each residue [14].

8.1.3 Higher-order structure differences

Although both NAs and proteins assume three-dimensional conformations, those of proteins are generally much more stable and diverse. For SM analysis methods, this can be either a challenge or an opportunity.

Methods that prescribe sequential processing of a protein have access to each individual residue, but a protein's stable structure may be difficult to remove, in which case it physically impedes sequential access. This is particularly problematic if a molecular motor is involved; processing may be stalled, or the protein may be threaded as hairpins [216]. To mitigate this issue, harsh chemical treatments may be used to sufficiently denature the protein for readout. Others have proposed pre-digestion of proteins to smaller fragments, which are less likely to assume rigid structures [16; 14]. However, like other BP analysis methods requiring digestion [217], this comes at the cost of information loss, as it introduces uncertainty on which fragments originated from the same protein.

Instead of removing structure, methods may also use the information present in the structure to identify a protein and accept that some internal residues will remain inaccessible for analysis. To my knowledge no successful higher-order structural fingerprinting methods for NAs exist to serve as inspiration, but then this strategy is more fitting for proteins as their structure is more dictated by sequence than is the case for NAs. FRET X fingerprinting is one method that leverages structure, while similar approaches with optical readouts can be imagined using recognizers for specific structural elements, such as aptamers or antibodies. Folded protein identification using an electrical readout is also being explored, although given the limited information gained from the signal and the wide variety of structures found in proteins, it is unclear how successful this strategy can be in a whole-proteome approach. The broad range of sizes of known proteins implies that no single pore will have the correct dimensions to effectively interrogate all proteins. To remedy this, an array of different pore sizes or a pore that allows fine control of the aperture size may be employed

[218]. Furthermore, extensive observation times would be required to extract as much information as possible from the protein. To this end, the indefinite protein trapping capability demonstrated by the NEOtrap – a nanopore partially obstructed by a DNA origami ball – may be leveraged [20].

From a computational analysis perspective, the fingerprinting of protein structural features has become more feasible in recent years. Importantly, rapid advances in protein structure prediction [219; 186] and molecular dynamics simulation [220] are making feasibility studies and signal prediction increasingly tractable. While these advances are partially made possible due to more elaborate models and continuously improving computational resources, some are built on the realization that predictions do not always require much complexity. For instance, the highly abstracted representations of whole proteins and nanopores in Wilson *et al.*'s steric exclusion model produced similar *in silico* electrical readouts as fully atomistic models, at a fraction of the computational cost [208]. Similarly, the coarse structural models used to evaluate the FRET X fingerprinting method in chapter 6 allowed for computationally inexpensive but realistic simulation of optical readouts, as verified by *in vitro* experiments.

8.1.4 Dynamic range

Thus far several methods have been proposed that, in theory, can discriminate between a large number of proteins. The issue of dynamic range is often not as thoroughly addressed, however; considering that the concentration between the most and least ubiquitous proteins differs by a factor of 10^9, how many proteins would we need to analyze to encounter all protein species, including the rarest? How long would it take to analyze all proteins in one cell? It is usually implied that, once discriminative power has been achieved, upscaling and streamlining the method until it can work at the required scale is an engineering issue, rather than a scientific one. Not all methods will prove to be equally easy to upscale however, so it would be beneficial at this stage to consider how these questions may eventually be answered.

Future nanopore-based protein analysis devices may follow the same path of development as current-day NA nanopore sequencers. Over the past five years ONT has rolled out several devices that engage different numbers of individually addressable nanopores at once – from 512 up to 128,400 – catering to different experiment sizes. To analyze the protein content of a single yeast cell, or about $\sim$42 million proteins [221], each pore would need to be engaged over 300 times. A method that utilizes a proteasome to pace the analysis speed of each protein, such as Chop-n-Drop, would need about 30 seconds per protein (assuming a median protein size of 350 residues). Thus, this hypothetical protein analysis machine, when constantly and fully engaged, would need over 2.5 hours to analyze the protein content of a single yeast cell. In a bulk sample and assuming fully random sampling, it would take slightly over eight hours (for 126 million proteins) to encounter a protein present at one copy per cell at least once with 95% probability. Of course some caveats apply: additional time should be added as pores will not be continuously engaged, detection accuracy will not be perfect and, above all,

it is unlikely that all proteins from a single cell could be captured for analysis. Nonetheless, this back-of-the-envelope calculation stems hopeful that proposed parallelization efforts will indeed yield a complete single-cell proteome within a reasonable amount of time in the near future. To increase analysis speed even further – or enable similar throughput on lighter hardware – inspiration may be drawn from the "adaptive sequencing" method that was recently introduced for NA nanopore sequencing [27; 26]; while a protein is processed through a nanopore, the squiggle obtained thus far may be compared in real-time to a database of frequently occurring large proteins. As soon as a match is found, the pore may eject the protein, thus freeing it to analyze a different protein.

Compared to electrical readouts, parallelization of optical readouts is often less involved; in TIRF microcscopy a single field of view may contain thousands of simultanously observed molecules [222]. To further increase throughput, an automated scanning stage may be used to move the FOV once sufficient information has been captured at the current coordinates. Similar to the automated ejection system used with nanopores, a feedback loop between analysis software and the scanning stage control may be introduced to move the FOV once enough events have been collected for identification of a set percentage of the proteins. Implementation of zero-mode waveguides (ZMWs) would allow the observation of vastly more proteins at once – a single PacBio cell currently offers eight million simultaneously operating ZMWs for NA sequencing – albeit at increased hardware cost. It is difficult to indicate how much time would be required to image a single FOV, as estimates differ between implementations from about 20 seconds for sequential processing methods [159] to several hours for stepwise degradation [52]. Although the latter approach is more likely to succeed at *de novo* sequencing, a significant speed-up would be required to reach a suitable throughput to screen for rare protein species.

8.2 Towards routine SM protein fingerprinting

8.2.1 The future of simulation studies

While the feasibility of a novel protein fingerprinting method will initially be judged based on scientific intuition, computational biology can help to more thoroughly evaluate feasibility in several scenarios and to identify targets for further development. An example is a study by Yao *et al.* which only showed that, theoretically, a noisy readout of a subset of residues is sufficient to discriminate between the majority of sequences in the UniProt human proteome [49]. Notably, the computational analysis did not include details on how the readout should be implemented. Then, several studies appeared that provided detailed experimental implementations, often supported by proof-of-concept experimental results [159; 12; 13]. Without exception, they were accompanied by their own computational simulations, detailing under which simulated conditions their method could discriminate the constituents of the human proteome.

As an initial indication of feasibility, such simulations are indeed useful and convincing. Their role does not have to end there, however; upon further development

of the practical methodology and acquisition of more preliminary data, it should be common practice to adjust the used model to account for new insights and re-evaluate which advances are required at minimum to reach a practically usable method. Advances in structural bioinformatics may also warrant re-evaluation of simulation studies. The most prominent example for this is the publication of AlphaFold2, a protein structure prediction algorithm which outperformed competitors by a large margin in blind performance evaluations [186]. AlphaFold2 was subsequently used to construct the AlphaFold Protein Structure Database, which contains predicted structures of 98.5 percent of known human protein sequences [223]. Although not all predicted structures will faithfully represent the actual structures, we can now effectively use a previously unimaginable amount of structural data, to *in silico* evaluate SM protein analysis methods that rely on structure. The FRET X simulation study described in chapter 6 for instance, was limited to work only with high-quality structures of proteins crystallized as single chains, of which several hundreds exist. Using the AlphaFold2 database and given sufficient computational power, this study could be expanded to include several thousands of proteins.

Furthermore, future simulation studies may explore the effects of alternative splicing (AS) on a method's ability to analyze a proteome. It is estimated that 93 percent of all human genes produce more than one spliceoform, which may be highly divergent from the canonical form [224], thus it is important to take AS into account when making claims on protein discernability. Again, for structure-based methods advances in *in silico* structure prediction may enable a deeper dive into this topic; although structures of multiple AS variants are not typically determined using conventional structure determination methods due to cost or experimental difficulty, it is comparatively easy to predict them.

Lastly, simulations also provide the opportunity to develop analysis software before *in vitro* methods have been developed far enough to generate sufficient data for this purpose [13; 16; 15; 14]. Doing so ensures that the development of analysis software is an integral part of method development. Moreover, it provides a valuable opportunity to test our understanding of the biophysical mechanism underlying the method; if analysis software developed using model parameters or output performs well on preliminary *in vitro* data, it indicates that we have an adequate working knowledge of the underlying biophysical mechanism.

8.2.2 The future of data analysis software

While future developments in data analysis depend heavily on the type of readout produced by *in vitro* methods, it is still possible to give generally applicable directions. Regardless of the implementation, it is likely that different phases in method development will require different data analysis solutions. In the concept development phase, the amount of acquired data will be low, and the quality will vary as *in vitro* methodology must still go through many iterations of improvement. Accordingly, data analysis software should rely on simple data models; while large neural networks containing many parameters may be able to fit patterns in the data well, the requirement of many labeled data points and potentially long

training times do not support the user in rapid development iterations and are thus not suitable for such a task. In TIRF microscopy, where prototyping of optical readout-based SM protein analysis takes place, most data analysis tools are therefore still based on hidden Markov models (HMMs); these algorithms may be fitted with little or no manual annotation and work adequately using few parameters. User friendliness is another important, yet frequently overlooked, aspect of analysis software for this phase; to keep turnaround times short, wet lab researchers may prefer to analyze their own data without intervention of a data analyst. To facilitate this, analysis tools may be outfitted with an intuitive user interface and be served remotely through a web browser to move the responsibility of software installation and maintenance from users to a single maintainer. It is with this precise purpose that the FRETboard tool described in chapter 3 was developed.

Once *in vitro* method design has stabilized, more data of a more stable quality will be generated. Rapid retraining of data models is no longer key, thus more opportunities arise to optimize data analysis methodology. Here an SM analysis method may find its first practical applications, such as the detection of a few biomarkers, thus analysis software would need to solve a classification problem for a few classes. Conventional machine learning or deep learning implementations are likely able to tackle this issue with relative ease. Labeled training data may be obtained by analyzing pure samples of each target protein, which is still tractable at this stage.

In the final stage of development full proteomes, including spliceoforms and post-translational modifications, may be analyzed. From the data analysis perspective, this would be a vastly more complicated undertaking; the number of output classes may now reach tens of thousands to millions [154]. While potentially a tractable problem for deep neural networks [225], such classifiers require very large amounts of annotated training data. Of course, analyzing pure samples for all possible proteins to this end is intractable.

For some methods, training on *in silico* generated data may be possible. More realistically, the first steps in SM full-proteome data analysis will lie in formulating suitable embeddings, i.e. translations of readouts to numerical representations that allow meaningful clustering of proteins. Such embeddings may take the shape of a machine learning implementation or may be grounded in physical theory. Indeed, protein fingerprint embeddings have been developed previously for MS spectra [47], which in turn drew inspiration from word embeddings in natural language processing [226]. If a fingerprint contains some structural data, such as relative positions of a subset of residues, 3D-invariant moments may provide a basis for a suitable embedding [227]. Combined with a database of known proteins and a suitable clustering method, this approach would allow protein identification for a growing number of classes, with the flexibility to focus on specific cell types or protein families.

8.2.3 The co-existence of SM and mass spectrometry

Although large strides have been made in the development of SM protein analysis methods, it is certain that mass spectrometry (MS) will remain the golden standard for protein analysis for some years to come. With tandem MS reaching mainstream status and further improvements focussing on single-cell applications [45] and natively folded protein assemblies [228], the technology has matured so well that it is difficult to outperform. The immediate future of new SM protein analysis methods will therefore not lie in competing with MS in its own field, but in filling niches that MS cannot, and in complementing MS analyses with secondary, orthogonal information.

Paradoxically, initial practical uses of SM protein analysis devices may not focus on the detection of proteins occurring in very low copy numbers. Rather, meaningful contributions can be made more easily by providing less expensive analysis tools that are operable by less specialized users. A single MS device may cost anywhere from several tens of thousands to a million euros and incur significant running costs, which puts in-house analysis far out of reach for smaller laboratories. Initiatives to streamline the spread and shared use of MS equipment promote broader adoption [229], but an inexpensive alternative – even if less accurate – may relieve pressure on existing MS equipment and suffice for some purposes. Potential applications could include targeted disease biomarker analysis in fractionated and filtered samples, or coarse sample profiling. In remote areas or developing countries, where reaching the nearest MS provider can be difficult, such a hypothetical protein analysis device may be the only option. This would mirror a similar development already seen in NA sequencing; nanopore sequencers were successfully used to track ebola in Liberia and Guinea [230], and SARS-CoV-2 spread worldwide [231–233] – not because they produce single-molecule data, but because they are light, inexpensive and easy to operate compared to second generation sequencing devices.

Furthermore, SM protein analysis methods may augment MS analysis, to enable the analysis of more complex samples or provide more in depth information. Despite improvements in MS sensitivity and data analysis methods, the m/z quantity read out remains a fingerprint, containing incomplete information on highly complex molecules. For instance, the bottom-up MS approach requires digestion of proteins prior to analysis, after which the original protein needs to be reconstituted, however this is not always possible based on MS fingerprints alone [234]. The orthogonal information provided by SM protein analysis methods employing electrical or optical readout may partially remedy this. Long-range information provided by an electrical readout of proteins passing through a nanopore may make this reconstitution easier and more reliable, even if only few residues are read out. In a way, this is analogous to the scaffolding of a genome using the minimal, error-prone, long-range information provided by optical mapping [235].

8.2.4 Closing remarks

There is much reason for optimism in the nascent field of SM protein analysis; although it has set itself the difficult task of studying the most diverse group of biomolecules at the ultimate resolution, several methods have already been conceived and brought to the proof-of-concept stage within the last decade. As it stands, the first commercial implementations of SM protein fingerprinting methods may arrive and mature in the coming decade, with the first SM sequencing methods following closely behind. As I speculate here, this rapid development may be aided by developments in SM sensing methods for NAs, but also by continuing the tight integration of computational and experimental method development. The work described in this thesis supports each of these elements, while simultaneously introducing several immediately useful tools for data analysis.

It is difficult to overestimate the impact that a mature SM protein sequencing method would have on molecular biology and medicine, though one can imagine it would surpass that of DNA sequencing; a transformative force empowering all fields of biology and enabling leaps in our understanding of disease mechanisms. With this much societal impact on the line, it will be our responsibility going forward to ensure an equitable distribution of benefits and opportunities in research, by weighing cost against resolution, and making access ubiquitous.

REFERENCES

[1] International Human Genome Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[2] Evan E Eichler, Royden A Clark, and Xinwei She. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Reviews Genetics*, 5(5):345–354, 2004.

[3] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

[4] Shivani Gupta and Raffaella Santoro. Regulation and roles of the nucleolus in embryonic stem cells: from ribosome biogenesis to genome organization. *Stem Cell Reports*, 15(6):1206–1219, 2020.

[5] Olga V Iarovaia, Elizaveta P Minina, Eugene V Sheval, Daria Onichtchouk, Svetlana Dokudovskaya, Sergey V Razin, and Yegor S Vassetzky. Nucleolus: a central hub for nuclear functions. *Trends in Cell Biology*, 29(8):647–659, 2019.

[6] Mikael S Lindström, Deana Jurada, Sladana Bursac, Ines Orsolic, Jiri Bartek, and Sinisa Volarevic. Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene*, 37(18):2351–2366, 2018.

[7] Achillefs N Kapanidis, Emmanuel Margeat, Sam On Ho, Ekaterine Kortkhonjia, Shimon Weiss, and Richard H Ebright. Initial transcription by rna polymerase proceeds through a dna-scrunching mechanism. *Science*, 314(5802):1144–1147, 2006.

[8] Allan Chris M Ferreon, Yann Gambin, Edward A Lemke, and Ashok A Deniz. Interplay of α-synuclein binding and conformational switching probed by single-molecule fluorescence. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5645–5650, 2009.

[9] Eitan Lerner, Antonino Ingargiola, and Shimon Weiss. Characterizing highly dynamic conformational states: The transcription bubble in RNAP-promoter open complex as an example. *Journal of Chemical Physics*, 148(12):123315, 2018.

[10] Philippe Haas, Patrick Then, Andreas Wild, Wilfried Grange, Sylvain Zorman, Martin Hegner, Michel Calame, Ueli Aebi, Josef Flammer, and Bert Hecht. Fast quantitative single-molecule detection at ultralow concentrations. *Analytical Chemistry*, 82(14):6299–6302, 2010.

[11] Tomas Hirschfeld. Optical microscopic observation of single small molecules. *Applied Optics*, 15(12):2965–2966, 1976.

[12] Jagannath Swaminathan, Alexander A Boulgakov, Erik T Hernandez, Angela M Bardo, James L Bachman, Joseph Marotta, Amber M Johnson, Eric V Anslyn, and Edward M Marcotte. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nature Biotechnology*, 36(11):1076–1082, 2018.

[13] Shilo Ohayon, Arik Girsault, Maisa Nasser, Shai Shen-Orr, and Amit Meller. Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. *PLOS Computational Biology*, 15(5):e1007067, 2019.

[14] Brian D Reed, Michael J Meyer, Valentin Abramzon, Omer Ad, Pat Adcock, Faisal R Ahmad, Gun Alppay, James A Ball, James Beach, Dominique Belhachemi, et al. Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. *bioRxiv*, 2022.

[15] Jarrett D Egertson, Dan DiPasquo, Alana Killeen, Vadim Lobanov, Sujal Patel, and Parag Mallick. A theoretical framework for proteome-scale single-molecule protein identification using multi-affinity protein binding reagents. *BioRxiv*, 2021.

[16] Jessica M Hong, Michael Gibbons, Ali Bashir, Diana Wu, Shirley Shao, Zachary Cutts, Mariya Chavarha, Ye Chen, Lauren Schiff, Mikelle Foster, et al. Protseq: Toward high-throughput, single-molecule protein sequencing via amino acid conversion into dna barcodes. *iScience*, 25(1):103586, 2022.

[17] Shuai Chang, Jin He, Ashley Kibel, Myeong Lee, Otto Sankey, Peiming Zhang, and Stuart Lindsay. Tunnelling readout of hydrogen-bonding-based recognition. *Nature Nanotechnology*, 4(5):297–301, 2009.

[18] Amit Meller. Dynamics of polynucleotide transport through nanometre-scale pores. *Journal of physics: condensed matter*, 15(17):R581, 2003.

[19] Alina Asandei, Mauro Chinappi, Jong-kook Lee, Chang Ho Seo, Loredana Mereuta, Yoonkyung Park, and Tudor Luchian. Placement of oppositely charged aminoacids at a polypeptide termini determines the voltage-controlled braking of polymer transport through nanometer-scale pores. *Scientific Reports*, 5(1):1–13, 2015.

[20] Sonja Schmid, Pierre Stömmer, Hendrik Dietz, and Cees Dekker. Nanopore electro-osmotic trap for the label-free study of single proteins and their conformations. *Nature Nanotechnology*, 16(11):1244–1250, 2021.

[21] Winston Timp, Jeffrey Comer, and Aleksei Aksimentiev. Dna base-calling from a nanopore using a viterbi algorithm. *Biophysical Journal*, 102(10):L37–L39, 2012.

[22] Jacob Schreiber and Kevin Karplus. Analysis of nanopore data using hidden markov models. *Bioinformatics*, 31(12):1897–1903, 2015.

[23] Matei David, L. J. Dursi, Delia Yao, Paul C. Boutros, and Jared T. Simpson. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics*, 33(1):49, 2017.

[24] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. Deepnano: deep recurrent neural networks for base calling in minion nanopore reads. *PLOS ONE*, 12(6):e0178751, 2017.

[25] Haotian Teng, Minh Duc Cao, Michael B Hall, Tania Duarte, Sheng Wang, and Lachlan JM Coin. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*, 7(5):giy037, 2018.

[26] Yuwei Bao, Jack Wadden, John R Erb-Downward, Piyush Ranjan, Weichen Zhou, Torrin L McDonald, Ryan E Mills, Alan P Boyle, Robert P Dickson, David Blaauw, et al. Squigglenet: real-time, direct classification of nanopore signals. *Genome Biology*, 22(1):1–16, 2021.

[27] Sam Kovaka, Yunfan Fan, Bohan Ni, Winston Timp, and Michael C Schatz. Targeted nanopore sequencing by real-time mapping of raw electrical signal with uncalled. *Nature Biotechnology*, 39(4):431–441, 2021.

[28] Renmin Han, Yu Li, Xin Gao, and Sheng Wang. An accurate and rapid continuous wavelet dynamic time warping algorithm for end-to-end mapping in ultra-long nanopore sequencing. *Bioinformatics*, 34(17):i722–i731, 2018.

[29] Michael J Levene, Jonas Korlach, Stephen W Turner, Mathieu Foquet, Harold G Craighead, and Watt W Webb. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299(5607):682–686, 2003.

[30] Denis Garoli, Hirohito Yamazaki, Nicolò Maccaferri, and Meni Wanunu. Plasmonic nanopores for single-molecule detection and manipulation: toward sequencing applications. *Nano Letters*, 19(11):7553–7562, 2019.

[31] Eitan Lerner, Anders Barth, Jelle Hendrix, Benjamin Ambrose, Victoria Birkedal, Scott C Blanchard, Richard Börner, Hoi Sung Chung, Thorben Cordes, Timothy D Craggs, et al. Fret-based dynamic structural biology: Challenges, perspectives and an appeal for open-science practices. *Elife*, 10:e60416, 2021.

[32] Sean A. McKinney, Chirlmin Joo, and Taekjip Ha. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophysical Journal*, 91(5):1941–1951, 2006.

[33] Jonathan E Bronson, Jingyi Fei, Jake M Hofman, Ruben L Gonzalez Jr, and Chris H Wiggins. Learning rates and states from biophysical time series: a bayesian approach to model selection and single-molecule fret data. *Biophysical Journal*, 97(12):3196–3205, 2009.

[34] Sonja Schmid, Markus Götz, and Thorsten Hugel. Single-molecule analysis beyond dwell times: demonstration and assessment in and out of equilibrium. *Biophysical Journal*, 111(7):1375–1384, 2016.

[35] Johannes Thomsen, Magnus Berg Sletfjerding, Simon Bo Jensen, Stefano Stella, Bijoya Paul, Mette Galsgaard Malle, Guillermo Montoya, Troels Christian Petersen, and Nikos S Hatzakis. Deepfret, a software for rapid and automated single-molecule fret data classification using deep learning. *Elife*, 9:e60404, 2020.

[36] Ciro Cecconi, Elizabeth A Shank, Carlos Bustamante, and Susan Marqusee. Direct observation of the three-state folding of a single protein molecule. *Science*, 309(5743):2057–2060, 2005.

[37] Prakash Shrestha, Darren Yang, Toma E Tomov, James I MacDonald, Andrew Ward, Hans T Bergal, Elisha Krieg, Serkan Cabi, Yi Luo, Bhavik Nathwani, et al. Single-molecule mechanical fingerprinting with dna nanoswitch calipers. *Nature Nanotechnology*, 16(12):1362–1370, 2021.

[38] Darren Yang, Andrew Ward, Ken Halvorsen, and Wesley P Wong. Multiplexed single-molecule force spectroscopy using a centrifuge. *Nature communications*, 7(1):1–7, 2016.

[39] Julien Chaste, A Eichler, J Moser, G Ceballos, R Rurali, and A Bachtold. A nanomechanical mass sensor with yoctogram resolution. *Nature Nanotechnology*, 7(5):301–304, 2012.

[40] M. Hanay, S. Kelber, A.K. Naik, D. Chi, S. Hentz, E.C. Bullard, E. Colinet, L. Duraffourg, and M.L. Roukes. Single-protein nanomechanical mass spectrometry in real time. *Nature Nanotechnology*, 7(9):602–608, 2012.

[41] MJJ Dijkstra, WJ Fokkink, J Heringa, E van Dijk, and S Abeln. The characteristics of molten globule states and folding pathways strongly depend on the sequence of a protein. *Molecular Physics*, 116(21-22):3173–3180, 2018.

[42] Rebecca Victoria Bowen, Clive Gavin Brown, Mark Bruce, Andrew John Heron, Jayne Elizabeth Wallace, James White, Joseph Hargreaves Lloyd, David Antoni Alves, Domenico Caprotti, Lakmal Jayasinghe, Luke McNeil, John Milton, Antonino Puglisi, and Szabolcs Soeroes. Method for controlling the movement of a polynucleotide through a transmembrane pore, January 5 2017. US Patent App. 15/113,174.

[43] Kate R Lieberman, Gerald M Cherf, Michael J Doody, Felix Olasagasti, Yvette Kolodji, and Mark Akeson. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *Journal of the American Chemical Society*, 132(50):17961–17972, 2010.

[44] James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[45] Harrison Specht, Edward Emmott, Aleksandra A Petelski, R Gray Huffman, David H Perlman, Marco Serra, Peter Kharchenko, Antonius Koller, and Nikolai Slavov. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology*, 22(1):1–27, 2021.

[46] Yasset Perez-Riverol, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, et al. The pride database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, 50(D1):D543–D552, 2022.

[47] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H Spaaks, Faruk Diblen, Simon Rogers, and Justin JJ Van Der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):e1008724, 2021.

[48] Douglas D Axe. Estimating the prevalence of protein sequences adopting functional enzyme folds. *Journal of Molecular Biology*, 341(5):1295–1315, 2004.

[49] Yao Yao, Margreet Docter, Jetty Van Ginkel, Dick de Ridder, and Chirlmin Joo. Single-molecule protein sequencing through fingerprinting: computational assessment. *Physical Biology*, 12(5):055003, 2015.

[50] Florian Leonardus Rudolfus Lucas, Kumar Sarthak, Erica Mariska Lenting, David Coltan, Nieck Jordy van der Heide, Roderick Corstiaan Abraham Versloot, Aleksei Aksimentiev, and Giovanni Maglia. The manipulation of the internal hydrophobicity of FraC nanopores augments peptide capture and recognition. *ACS Nano*, 2021.

[51] Henry Brinkerhoff, Albert CW Kang, Jingqian Liu, Aleksei Aksimentiev, and Cees Dekker. Infinite re-reading of single proteins at single-amino-acid resolution using nanopore sequencing. *bioRxiv*, DOI: 10.1101/2021.07.13.452225(published: 14 July), 2021.

[52] Jagannath Swaminathan, Alexander A Boulgakov, and Edward M Marcotte. A theoretical justification for single molecule peptide sequencing. *PLOS computational biology*, 11(2):e1004080, 2015.

[53] Yanan Zhao, Brian Ashcroft, Peiming Zhang, Hao Liu, Suman Sen, Weisi Song, JongOne Im, Brett Gyarfas, Saikat Manna, Sovan Biswas, et al. Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nature Nanotechnology*, 9(6):466–473, 2014.

[54] Brent Ewing, LaDeana Hillier, Michael C Wendl, and Phil Green. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome Research*, 8(3):175–185, 1998.

[55] Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Research*, 8(3):186–194, 1998.

[56] Muhammad Haseeb and Fahad Saeed. High performance computing framework for tera-scale database search of mass spectrometry data. *Nature Computational Science*, 1(8):550–561, 2021.

[57] Erwin L van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, 2014.

[58] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, 2016.

[59] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338, 2018.

[60] Jared T Simpson, Rachael Workman, Philip C Zuzarte, Matei David, Lewis Jonathan Dursi, and Winston Timp. Detecting DNA methylation using the Oxford Nanopore Technologies MinION sequencer. *bioRxiv*, page 047142, 2016.

[61] Benjamin N Miles, Aleksandar P Ivanov, Kerry A Wilson, Fatma Doğan, Deanpen Japrung, and Joshua B Edel. Single molecule sensing with solid-state nanopores: novel materials, methods, and applications. *Chemical Society Reviews*, 42(1):15–28, 2013.

[62] Parveen Goyal, Petya V Krasteva, Nani Van Gerven, Francesca Gubellini, Imke Van den Broeck, Anastassia Troupiotis-Tsaïlaki, Wim Jonckheere, Gérard Péhau-Arnaudet, Jerome S Pinkner, Matthew R Chapman, et al. Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature*, 516(7530):250–253, 2014.

[63] Tom Z Butler, Mikhail Pavlenok, Ian M Derrington, Michael Niederweis, and Jens H Gundlach. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):20647–20652, 2008.

[64] Giovanni Maglia, Marcela Rincon Restrepo, Ellina Mikhailova, and Hagan Bayley. Enhanced translocation of single DNA molecules through α-hemolysin nanopores by manipulation of internal charge. *Proceedings of the National Academy of Sciences*, 105(50):19720–19725, 2008.

[65] David Stoddart, Andrew J Heron, Jochen Klingelhoefer, Ellina Mikhailova, Giovanni Maglia, and Hagan Bayley. Nucleobase recognition in ssDNA at the central constriction of the α-hemolysin pore. *Nano Letters*, 10(9):3633–3637, 2010.

[66] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of DNA in a nanopore at 5-å precision. *Nature Biotechnology*, 30(4):344–348, 2012.

[67] Elizabeth A Manrao, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology*, 30(4):349–353, 2012.

[68] Andrew John Heron, David Antoni Alves, James Anthony Clarke, Marion Louise Crawford, Daniel Ryan Garalde, Graham Hall, Daniel John Turner, and James White. Enzyme stalling method, January 14 2016. US Patent App. 14/773,164.

[69] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, 12(4):351–356, 2015.

[70] Camilla LC Ip, Matthew Loose, John R Tyson, Mariateresa de Cesare, Bonnie L Brown, Miten Jain, Richard M Leggett, David A Eccles, Vadim Zalunin, John M Urban, et al. MinION analysis and reference consortium: Phase 1 data release and analysis. *F1000Research*, 4, 2015.

[71] Ruby White, Christophe Pellefigues, Franca Ronchese, Olivier Lamiable, and David Eccles. Investigation of chimeric reads using the minion. *F1000Research*, 6, 2017.

[72] Marcus Stoiber and James Brown. BasecRAWller: Streaming nanopore basecalling directly from raw signal. *bioRxiv*, page 133058, 2017.

[73] Tamas Szalay and Jene A Golovchenko. *De novo* sequencing and variant calling with nanopores using PoreSeq. *Nature Biotechnology*, 33(10):1087–1091, 2015.

[74] Felix Weninger, Johannes Bergmann, and Björn W Schuller. Introducing CURRENNT: the Munich open-source CUDA recurrent neural network toolkit. *Journal of Machine Learning Research*, 16(3):547–551, 2015.

[75] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015.

[76] Hans J Jansen, Michael Liem, Susanne A Jong-Raadsen, Sylvie Dufour, Finn-Arne Weltzien, William Swinkels, Alex Koelewijn, Arjan P Palstra, Bernd Pelster, Herman P Spaink, et al. Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. *Scientific Reports*, 7(1):7213, 2017.

[77] Marcus H Stoiber, Joshua Quick, Rob Egan, Ji Eun Lee, Susan E Celniker, Robert Neely, Nicholas Loman, Len Pennacchio, and James B Brown. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv*, page 094672, 2016.

[78] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv*, 1506.00019, 2015.

[79] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[80] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv*, 1211.3711, 2012.

[81] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[82] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, 25(11):1750–1756, 2015.

[83] Dmitry Antipov, Anton Korobeynikov, Jeffrey S McLean, and Pavel A Pevzner. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, 32(7):1009–1015, 2015.

[84] Mohammed-Amin Madoui, Stefan Engelen, Corinne Cruaud, Caroline Belser, Laurie Bertrand, Adriana Alberti, Arnaud Lemainque, Patrick Wincker, and Jean-Marc Aury. Genome assembly using nanopore-guided long and error-free DNA reads. *BMC Genomics*, 16(1):327, 2015.

[85] Minh Duc Cao, Son Hoang Nguyen, Devika Ganesamoorthy, Alysha Elliott, Matthew Cooper, and Lachlan JM Coin. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *bioRxiv*, page 054783, 2016.

[86] Chengxi Ye, Christopher M Hill, Shigang Wu, Jue Ruan, and Zhanshan Sam Ma. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports*, 6:31900, 2016.

[87] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE*, 9(11):e112963, 2014.

[88] John R Tyson, Nigel J O'Neil, Miten Jain, Hugh E Olsen, Philip Hieter, and Terrance P Snutch. Whole genome sequencing and assembly of a *Caenorhabditis elegans* genome with complex genomic rearrangements using the MinION sequencing device. *bioRxiv*, page 099143, 2017.

[89] Erwin Datema, Raymond JM Hulzink, Lisanne Blommers, Jose Espejo Valle-Inclan, Nathalie Van Orsouw, Alexander HJ Wittenberg, and Martin De Vos. The megabase-sized fungal genome of *Rhizoctonia solani* assembled from nanopore reads only. *bioRxiv*, page 084772, 2016.

[90] Francesca Giordano, Louise Aigrain, Michael A Quail, Paul Coupland, James K Bonfield, Robert M Davies, German Tischler, David K Jackson, Thomas M Keane, Jing Li, et al. *De novo* yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific Reports*, 7(1):3935, 2017.

[91] Benjamin Istace, Anne Friedrich, Léo d'Agata, Sébastien Faye, Emilie Payen, Odette Beluche, Claudia Caradec, Sabrina Davidas, Corinne Cruaud, Gianni Liti, et al. *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *GigaScience*, 6(2):1–13, 2017.

[92] Yesesri Cherukuri and Sarath Chandra Janga. Benchmarking of *de novo* assembly algorithms for nanopore data reveals optimal performance of OLC approaches. *BMC Genomics*, 17(7):507, 2016.

[93] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017.

[94] Govinda M Kamath, Ilan Shomorony, Fei Xia, Thomas Courtade, and N Tse David. Hinge: Long-read assembly achieves optimal repeat resolution. *Genome Research*, pages gr–216465, 2017.

[95] Yu Lin, Jeffrey Yuan, Mikhail Kolmogorov, Max W Shen, Mark Chaisson, and Pavel A Pevzner. Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America*, 113(52):E8396–E8405, 2016.

[96] Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016.

[97] Kim Judge, Martin Hunt, Sandra Reuter, Alan Tracey, Michael A Quail, Julian Parkhill, and Sharon J Peacock. Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. *Microbial Genomics*, 2(9), 2016.

[98] Sergey Koren, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W Richard McCombie, Erich D Jarvis, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7):693–700, 2012.

[99] Eugene W Myers, Granger G Sutton, Art L Delcher, Ian M Dew, Dan P Fasulo, Michael J Flanigan, Saul A Kravitz, Clark M Mobarry, Knut HJ Reinert, Karin A Remington, et al. A whole-genome assembly of *Drosophila*. *Science*, 287(5461):2196–2204, 2000.

[100] Jason R Miller, Arthur L Delcher, Sergey Koren, Eli Venter, Brian P Walenz, Anushka Brownley, Justin Johnson, Kelvin Li, Clark Mobarry, and Granger Sutton. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24):2818–2824, 2008.

[101] Aleksey Zimin. MSR-CA–efficient *De Novo* genome assembler for long and short read data. In *Plant and Animal Genome XXI Conference*. Plant and Animal Genome, 2013.

[102] Steven L Salzberg, Adam M Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J Treangen, Michael C Schatz, Arthur L Delcher, Michael Roberts, et al. Gage: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3):557–567, 2012.

[103] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630, 2015.

[104] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569, 2013.

[105] Maximilian H-W Schmidt, Alexander Vogel, Alisandra K Denton, Benjamin Istace, Alexandra Wormit, Henri van de Geest, Marie E Bolger, Saleh Alseekh, Janina Maß, Christian Pfaff, et al. *De novo* assembly of a new *Solanum pennellii* accession using nanopore sequencing. *The Plant Cell*, 29(10):2336–2348, 2017.

[106] Hengyun Lu, Francesca Giordano, and Zemin Ning. Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics & Bioinformatics*, 14(5):265–279, 2016.

[107] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome research*, 27(5):737–746, 2017.

[108] Chen-Shan Chin, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, 2016.

[109] Michael Börsch and Thomas M. Duncan. Spotlighting motors and controls of single FoF1-ATP synthase. *Biochemical Society transactions*, 41(5):1219–1226, 2013.

[110] Matthew D Newton, Benjamin J Taylor, Rosalie PC Driessen, Leonie Roos, Nevena Cvetesic, Shenaz Allyjaun, Boris Lenhard, Maria Emanuela Cuomo, and David S Rueda. DNA stretching induces Cas9 off-target activity. *Nature Structural and Molecular Biology*, 26(3):185–192, 2019.

[111] Viktorija Globyte, Seung Hwan Lee, Taegeun Bae, Jin-Soo Kim, and Chirlmin Joo. CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. *EMBO Journal*, 38(4):e99466, 2019.

[112] Björn Hellenkamp, Sonja Schmid, Olga Doroshenko, Oleg Opanasyuk, Ralf Kühnemuth, Soheila Rezaei Adariani, Benjamin Ambrose, Mikayel Aznauryan, Anders Barth, Victoria Birkedal, et al. Precision and accuracy of single-molecule fret measurements—a multi-laboratory benchmark study. *Nature Methods*, 15(9):669–676, 2018.

[113] Stephan Uphoff, Seamus J Holden, Ludovic Le Reste, Javier Periz, Sebastian Van De Linde, Mike Heilemann, and Achillefs N Kapanidis. Monitoring multiple distances within a single molecule using switchable FRET. *Nature Methods*, 7(10):831, 2010.

[114] Max Greenfeld, Dmitri S. Pavlichin, Hideo Mabuchi, and Daniel Herschlag. Single Molecule Analysis Research Tool (SMART): an integrated approach for analyzing single molecule data. *PLOS ONE*, 7(2):30024, 2012.

[115] Sebastian König, Mélodie Hadzic, Erica Fiorini, Richard Börner, Danny Kowerko, Wolf Blanckenhorn, and Roland Sigel. BOBA FRET: Bootstrap-Based Analysis of Single-Molecule FRET Data. *PLOS ONE*, 8(12):1–17, 2013.

[116] Jan Willem van de Meent, Jonathan E. Bronson, Chris H. Wiggins, and Ruben L. Gonzalez. Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. *Biophysical Journal*, 106(6):1327–1337, 2014.

[117] Søren Preus, Sofie L Noer, Lasse L Hildebrandt, Daniel Gudnason, and Victoria Birkedal. isms: single-molecule fret microscopy software. *Nature Methods*, 12(7):593–594, 2015.

[118] Manuel Juette, Daniel Terry, Michael Wasserman, Roger Altman, Zhou Zhou, Hong Zhao, and Scott Blanchard. Single-molecule imaging of non-equilibrium molecular ensembles on the millisecond timescale. *Nature Methods*, 13(4):341–344, 2016.

[119] Melodie C. A. S. Hadzic, Richard Börner, Sebastian L. B. König, Danny Kowerko, and Roland K. O. Sigel. Reliable state identification and state transition detection in fluorescence intensity-based single-molecule förster resonance energy-transfer data. *Journal of Physical Chemistry B*, 122(23):6134–6147, 2018.

[120] Antonino Ingargiola, Ted Laurence, Robert Boutelle, Shimon Weiss, and Xavier Michalet. Photon-HDF5: An open file format for timestamp-based single-molecule fluorescence experiments. *Biophysical Journal*, 110(1):26–33, 2016.

[121] Bokeh Development Team. *Bokeh: Python library for interactive visualization*, 2019.

[122] Jacob Schreiber. Pomegranate: fast and flexible probabilistic modeling in python. *Journal of Machine Learning Research*, 18(164):1–6, 2018.

[123] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[124] Donald J Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

[125] Jongjin Lee, Sangjun Park, Wooyoung Kang, and Sungchul Hohng. Accelerated super-resolution imaging with FRET-PAINT. *Molecular Brain*, 10(1):63, 2017.

[126] Mike Filius, Tao Ju Cui, Adithya N Ananth, Margreet W Docter, Jorrit W Hegge, John van der Oost, and Chirlmin Joo. High-speed super-resolution imaging using protein-assisted dna-paint. *Nano Letters*, 20(4):2264–2270, 2020.

[127] Damla Senol Cali, Jeremie S Kim, Saugata Ghose, Can Alkan, and Onur Mutlu. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Briefings in Bioinformatics*, 20(4):1542–1559, 2019.

[128] Michael Schmid, Daniel Frei, Andrea Patrignani, Ralph Schlapbach, Juerg E Frey, Mitja NP Remus-Emsermann, and Christian H Ahrens. Pushing the limits of *de novo* genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *bioRxiv*, 2018. doi:10.1101/300186.

[129] Todd P Michael, Florian Jupe, Felix Bemm, S Timothy Motley, Justin P Sandoval, Christa Lanz, Olivier Loudet, Detlef Weigel, and Joseph R Ecker. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nature communications*, 9(1):541, 2018.

[130] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[131] Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack: visualizing and processing long read sequencing data. *bioRxiv*, 2018. doi:10.1101/237180.

[132] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[133] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

[134] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

[135] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

[136] Eric D Green, James D Watson, and Francis S Collins. Human genome project: Twenty-five years of big biology. *Nature*, 526(7571):29–31, 2015.

[137] Brigitte Bruijns, Roald Tiggelaar, and Han Gardeniers. Massively parallel sequencing techniques for forensics: A review. *Electrophoresis*, 39(21):2642–2654, 2018.

[138] Elizabeth A Normand, Alicia Braxton, Salma Nassef, Patricia A Ward, Francesco Vetrini, Weimin He, Vipulkumar Patel, Chunjing Qu, Lauren E Westerfield, Samantha Stover, et al. Clinical exome sequencing for fetuses with ultrasound abnormalities and a suspected mendelian disorder. *Genome medicine*, 10(1):1–14, 2018.

[139] Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C Bruhm, Sarah Østrup Jensen, Jamie E Medina, Carolyn Hruban, James R White, et al. Genome-wide cell-free dna fragmentation in patients with cancer. *Nature*, 570(7761):385–389, 2019.

[140] Aaron M Newman, Scott V Bratman, Jacqueline To, Jacob F Wynne, Neville CW Eclov, Leslie A Modlin, Chih Long Liu, Joel W Neal, Heather A Wakelee, Robert E Merritt, et al. An ultrasensitive method for quantitating circulating tumor dna with broad patient coverage. *Nature Medicine*, 20(5):548–554, 2014.

[141] Nuno Rodrigues Faria, Ester C Sabino, Marcio RT Nunes, Luiz Carlos Junior Alcantara, Nicholas J Loman, and Oliver G Pybus. Mobile real-time surveillance of zika virus in brazil. *Genome medicine*, 8(1):1–4, 2016.

[142] Jacqueline Goordial, Ianina Altshuler, Katherine Hindson, Kelly Chan-Yam, Evangelos Marcolefas, and Lyle G Whyte. *In situ* field sequencing and life detection in remote (79°26'N) Canadian high arctic permafrost ice wedge microbial communities. *Frontiers in Microbiology*, page 2594, 2017.

[143] Aaron Pomerantz, Nicolás Peñafiel, Alejandro Arteaga, Lucas Bustamante, Frank Pichardo, Luis A Coloma, César L Barrio-Amorós, David Salazar-Valenzuela, and Stefan Prost. Real-time dna barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *GigaScience*, 7(4):giy033, 2018.

[144] Carlos de Lannoy, Dick de Ridder, and Judith Risse. The long reads ahead: de novo genome assembly using the minion. *F1000Research*, 6, 2017.

[145] Vladimír Boža, Peter Perešíni, Broňa Brejová, and Tomáš Vinař. Deepnano-blitz: a fast base caller for minion nanopore sequencers. *Bioinformatics*, 36(14):4191–4192, 2020.

[146] Matthew Loose, Sunir Malla, and Michael Stout. Real-time selective sequencing using nanopore technology. *Nature Methods*, 13(9):751–754, 2016.

[147] Henri van Kruistum, Reindert Nijland, David N Reznick, Martien AM Groenen, Hendrik-Jan Megens, and Bart JA Pollux. Parallel genomic changes drive repeated evolution of placentas in live-bearing fish. *Molecular Biology and Evolution*, 38(6):2627–2638, 2021.

[148] Dylan A Mistry, Jenny Y Wang, Mika-Erik Moeser, Thomas Starkey, and Lennard YW Lee. A systematic review of the sensitivity and specificity of lateral flow devices in the detection of sars-cov-2. *BMC Infectious Diseases*, 21(1):1–14, 2021.

[149] Marie-Madlen Pust, Colin Francis Davenport, Lutz Wiehlmann, and Burkhard Tümmler. Direct rna nanopore sequencing of pseudomonas aeruginosa clone c transcriptomes. *Journal of Bacteriology*, 204(1):e00418–21, 2021.

[150] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[151] Sujeevan Ratnasingham and Paul DN Hebert. Bold: The barcode of life data system (http://www. barcodinglife. org). *Molecular ecology notes*, 7(3):355–364, 2007.

[152] Kathryn Doroschak, Karen Zhang, Melissa Queen, Aishwarya Mandyam, Karin Strauss, Luis Ceze, and Jeff Nivala. Rapid and robust assembly and decoding of molecular tags with dna-based nanopore signatures. *Nature Communications*, 11(1):1–8, 2020.

[153] Roman A Zubarev. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics*, 13(5):723–726, 2013.

[154] Ruedi Aebersold, Jeffrey N Agar, I Jonathan Amster, Mark S Baker, Carolyn R Bertozzi, Emily S Boja, Catherine E Costello, Benjamin F Cravatt, Catherine Fenselau, Benjamin A Garcia, et al. How many human proteoforms are there? *Nature Chemical Biology*, 14(3):206, 2018.

[155] Laura Restrepo-Pérez, Chirlmin Joo, and Cees Dekker. Paving the way to single-molecule protein sequencing. *Nature Nanotechnology*, 13(9):786–796, 2018.

[156] Javier Antonio Alfaro, Peggy Bohländer, Mingjie Dai, Mike Filius, Cecil J Howard, Xander F van Kooten, Shilo Ohayon, Adam Pomorski, Sonja Schmid, Aleksei Aksimentiev, et al. The emerging landscape of single-molecule protein sequencing technologies. *Nature methods*, 18(6):604–617, 2021.

[157] Charles Gawad, Winston Koh, and Stephen R Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, 2016.

[158] Jeff Nivala, Logan Mulroney, Gabriel Li, Jacob Schreiber, and Mark Akeson. Discrimination among protein variants using an unfoldase-coupled nanopore. *ACS Nano*, 8(12):12365–12375, 2014.

[159] Jetty Van Ginkel, Mike Filius, Malwina Szczepaniak, Pawel Tulinski, Anne S Meyer, and Chirlmin Joo. Single-molecule peptide fingerprinting. *Proceedings of the National Academy of Sciences*, 115(13):3338–3343, 2018.

[160] Sungchul Hohng, Chirlmin Joo, and Taekjip Ha. Single-molecule three-color fret. *Biophysical Journal*, 87(2):1328–1337, 2004.

[161] Jean-Pierre Clamme and Ashok A Deniz. Three-color single-molecule fluorescence resonance energy transfer. *ChemPhysChem*, 6(1):74–77, 2005.

[162] Mike Filius, Sung Hyun Kim, Ivo Severins, and Chirlmin Joo. High-resolution single-molecule FRET via DNA exchange (FRET X). *Nano Letters*, 21(7):3295–3301, 2021.

[163] Sung Hyun Kim, Hyunwoo Kim, Hawoong Jeong, and Tae-Young Yoon. Encoding multiple virtual signals in dna barcodes with single-molecule fret. *Nano Letters*, 21(4):1694–1701, 2021.

[164] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[165] Jinbo Xu, Matthew Mcpartlon, and Jin Li. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence*, pages 1–9, 2021.

[166] Andrzej Kolinski and Jeffrey Skolnick. Reduced models of proteins and their applications. *Polymer*, 45(2):511–524, 2004.

[167] Sanne Abeln, Michele Vendruscolo, Christopher M Dobson, and Daan Frenkel. A simple lattice model that captures protein folding, aggregation and amyloid formation. *PLOS ONE*, 9(1):e85185, 2014.

[168] I Coluzza, HG Muller, and D Frenkel. Designing refoldable model molecules. *Physical Review E*, 68(4):046703, 2003.

[169] Valentino Bianco, Neus Pagès-Gelabert, Ivan Coluzza, and Giancarlo Franzese. How the stability of a folded protein depends on interfacial water properties and residue-residue interactions. *Journal of Molecular Liquids*, 245:129–139, 2017.

[170] Juami Hermine Mariama Van Gils, Erik Van Dijk, Alessia Peduzzo, Alexander Hofmann, Nicola Vettore, Marie P Schützmann, Georg Groth, Halima Mouhib, Daniel E Otzen, Alexander K Buell, et al. The hydrophobic effect characterises the thermodynamic signature of amyloid fibril growth. *PLOS Computational Biology*, 16(5):e1007767, 2020.

[171] Sergei F Chekmarev. How the dyes affect folding of small proteins in single-molecule fret experiments: A simulation study. *Biophysical chemistry*, 254:106243, 2019.

[172] Lucy R Allen and Emanuele Paci. Simulation of fluorescence resonance energy transfer experiments: effect of the dyes on protein folding. *Journal of Physics: Condensed Matter*, 22(23):235103, 2010.

[173] Arumay Pal and Yaakov Levy. Structure, stability and specificity of the binding of ssdna and ssrna with proteins. *PLOS Computational Biology*, 15(4):e1006768, 2019.

[174] Morten Källberg, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu, and Jinbo Xu. Template-based protein structure modeling using the raptorx web server. *Nature protocols*, 7(8):1511–1522, 2012.

[175] Morten Källberg, Gohar Margaryan, Sheng Wang, Jianzhu Ma, and Jinbo Xu. Raptorx server: a resource for template-based protein structure modeling. In *Protein structure prediction*, pages 17–27. Springer, 2014.

[176] Justin Kale, Elizabeth J Osterlund, and David W Andrews. Bcl-2 family proteins: changing partners in the dance towards death. *Cell Death & Differentiation*, 25(1):65–80, 2018.

[177] Nobuko Shiraiwa, Naohiro Inohara, Seiji Okada, Michisuke Yuzaki, Shin-ichi Shoji, and Shigeo Ohta. An additional form of rat bcl-x, bcl-xβ, generated by an unspliced rna, promotes apoptosis in promyeloid cells. *Journal of Biological Chemistry*, 271(22):13258–13265, 1996.

[178] Mariano A Garcia-Blanco, Andrew P Baraniak, and Erika L Lasda. Alternative splicing in disease and therapy. *Nature biotechnology*, 22(5):535–546, 2004.

[179] Omar Boutureira and Gonçalo JL Bernardes. Advances in chemical protein modification. *Chemical reviews*, 115(5):2174–2195, 2015.

[180] Nicolas Abello, Huib AM Kerstjens, Dirkje S Postma, and Rainer Bischoff. Selective acylation of primary amines in peptides and proteins. *Journal of Proteome Research*, 6(12):4770–4776, 2007.

[181] Darren A Thompson, Raymond Ng, and Philip E Dawson. Arginine selective reagents for ligation to peptides and proteins. *Journal of Peptide Science*, 22(5):311–319, 2016.

[182] Harold P Erickson. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biological procedures online*, 11(1):32–51, 2009.

[183] Shixian Lin, Xiaoyu Yang, Shang Jia, Amy M Weeks, Michael Hornsby, Peter S Lee, Rita V Nichiporuk, Anthony T Iavarone, James A Wells, F Dean Toste, et al. Redox-based reagents for chemoselective methionine bioconjugation. *Science*, 355(6325):597–602, 2017.

[184] Dimitri Alvarez Dorta, David Deniaud, Mathieu Mével, and Sébastien Gouin. Tyrosine conjugation methods for protein labelling. *Chemistry-A European Journal*, pages Online–ahead, 2020.

[185] Aneika C Leney and Albert JR Heck. Native mass spectrometry: what is in the name? *Journal of the American Society for Mass Spectrometry*, 28(1):5–13, 2016.

[186] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[187] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

[188] Stanley D Chandradoss, Anna C Haagsma, Young Kwang Lee, Jae-Ho Hwang, Jwa-Min Nam, and Chirlmin Joo. Surface passivation for single-molecule protein studies. *JoVE (Journal of Visualized Experiments)*, 86:e50549, 2014.

[189] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[190] Sanzo Miyazawa and Robert L Jernigan. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins: Structure, Function, and Bioinformatics*, 34(1):49–68, 1999.

[191] Fredrik Pontén, Karin Jirström, and Matthias Uhlen. The Human Protein Atlas—a tool for pathology. *The Journal of Pathology*, 216(4):387–393, 2008.

[192] Hasmik Keshishian, Michael W Burgess, Harrison Specht, Luke Wallace, Karl R Clauser, Michael A Gillette, and Steven A Carr. Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. *Nature Protocols*, 12(8):1683, 2017.

[193] Luke F Vistain and Savaş Tay. Single-cell proteomics. *Trends in Biochemical Sciences*, 46(8):661–672, 2021.

[194] Pier Giorgio Righetti and Egisto Boschetti. Introducing Low-Abundance Species in Proteome Analysis. In *Low-Abundance Proteome Discovery*, pages 1–11. Elsevier, jan 2013.

[195] Jagpreet S Nanda and Jon R Lorsch. Labeling of a protein with fluorophores using maleimide derivitization. *Methods in Enzymology*, 536:79–86, 2014.

[196] Zheng-Li Hu, Ming-Zhu Huo, Yi-Lun Ying, and Yi-Tao Long. Biological nanopore approach for single-molecule protein sequencing. *Angewandte Chemie*, 2021.

[197] Todd C Sutherland, Yi-Tao Long, Radu-Ioan Stefureac, Irene Bediako-Amoa, Heinz-Bernhard Kraatz, and Jeremy S Lee. Structure of peptides investigated by nanopore analysis. *Nano Letters*, 4(7):1273–1277, 2004.

[198] Gang Huang, Kherim Willems, Misha Soskine, Carsten Wloka, and Giovanni Maglia. Electro-osmotic capture and ionic discrimination of peptide and protein biomarkers with FraC nanopores. *Nature Communications*, 8(1):1–11, 2017.

[199] Fabien Piguet, Hadjer Ouldali, Manuela Pastoriza-Gallego, Philippe Manivet, Juan Pelta, and Abdelghani Oukhaled. Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. *Nature Communications*, 9(1):1–13, 2018.

[200] Jared Houghtaling, Cuifeng Ying, Olivia M Eggenberger, Aziz Fennouri, Santoshi Nandivada, Mitu Acharjee, Jiali Li, Adam R Hall, and Michael Mayer. Estimation of shape, volume, and dipole moment of individual proteins freely transiting a synthetic nanopore. *ACS Nano*, 13(5):5231–5242, 2019.

[201] Hadjer Ouldali, Kumar Sarthak, Tobias Ensslen, Fabien Piguet, Philippe Manivet, Juan Pelta, Jan C Behrends, Aleksei Aksimentiev, and Abdelghani Oukhaled. Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nature Biotechnology*, 38(2):176–181, 2020.

[202] Luciano Brocchieri and Samuel Karlin. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, 33(10):3390–3400, 2005.

[203] Shengli Zhang, Gang Huang, Roderick Versloot, Bart Marlon Herwig, Paulo Cesar Telles de Souza, Siewert-Jan Marrink, and Giovanni Maglia. Bottom-up fabrication of a multi-component nanopore sensor that unfolds, processes and recognizes single proteins. *bioRxiv*, doi: 10.1101/2020.12.04.411884(published: 6 December), 2020.

[204] Gang Huang, Arnout Voet, and Giovanni Maglia. FraC nanopores with adjustable diameter identify the mass of opposite-charge peptides with 44 dalton resolution. *Nature Communications*, 10(1):1–10, 2019.

[205] Pradeep Waduge, Rui Hu, Prasad Bandarkar, Hirohito Yamazaki, Benjamin Cressiot, Qing Zhao, Paul C Whitford, and Meni Wanunu. Nanopore-based measurements of protein size, fluctuations, and conformational changes. *ACS Nano*, 11(6):5706–5716, 2017.

[206] Rui Hu, João V Rodrigues, Pradeep Waduge, Hirohito Yamazaki, Benjamin Cressiot, Yasmin Chishti, Lee Makowski, Dapeng Yu, Eugene Shakhnovich, Qing Zhao, et al. Differential enzyme flexibility probed using solid-state nanopores. *ACS Nano*, 12(5):4494–4502, 2018.

[207] Giovanni Di Muccio, Aldo Eugenio Rossini, Daniele Di Marino, Giuseppe Zollo, and Mauro Chinappi. Insights into protein sequencing with an α-hemolysin nanopore by atomistic simulations. *Scientific Reports*, 9(1):1–8, 2019.

[208] James Wilson, Kumar Sarthak, Wei Si, Luyu Gao, and Aleksei Aksimentiev. Rapid and accurate determination of nanopore ionic current using a steric exclusion model. *ACS Sensors*, 4(3):634–644, 2019.

[209] Sarah Zernia, Nieck Jordy van der Heide, Nicole Stéphanie Galenkamp, Giorgos Gouridis, and Giovanni Maglia. Current blockades of proteins inside nanopores for real-time metabolome analysis. *ACS Nano*, 14(2):2296–2307, 2020.

[210] Zongmin Li, Lisette Arnaud, Patricia Rockwell, and Maria E Figueiredo-Pereira. A single amino acid substitution in a proteasome subunit triggers aggregation of ubiquitinated proteins in stressed neuronal cells. *Journal of neurochemistry*, 90(1):19–28, 2004.

[211] Galen A Collins, Tara Adele Gomez, Raymond J Deshaies, and William P Tansey. Combined chemical and genetic approach to inhibit proteolysis by the proteasome. *Yeast*, 27(11):965–974, 2010.

[212] Erik C Yusko, Brandon R Bruhn, Olivia M Eggenberger, Jared Houghtaling, Ryan C Rollings, Nathan C Walsh, Santoshi Nandivada, Mariya Pindrus, Adam R Hall, David Sept, et al. Real-time shape approximation and fingerprinting of single proteins using a nanopore. *Nature Nanotechnology*, 12(4):360–367, 2017.

[213] Natalie Lisa Mutter, Gang Huang, Nieck Jordy van der Heide, Florian Leonardus Rudolfus Lucas, Nicole Stéphanie Galenkamp, Giovanni Maglia, and Carsten Wloka. Preparation of fragaceatoxin c (FraC) nanopores. *Nanopore Technology*, pages 3–10, 2021.

[214] Bengt Bjellqvist, Graham J Hughes, Christian Pasquali, Nicole Paquet, Florence Ravier, Jean-Charles Sanchez, Séverine Frutiger, and Denis Hochstrasser. The focusing positions of polypeptides in immobilized ph gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14(1):1023–1031, 1993.

[215] John J Kasianowicz and Sergey M Bezrukov. On'three decades of nanopore sequencing'. *Nature biotechnology*, 34(5):481–482, 2016.

[216] Randall E Burton, Samia M Siddiqui, Yong-In Kim, Tania A Baker, and Robert T Sauer. Effects of protein stability and structure on substrate processing by the clpxp unfolding and degradation machine. *The EMBO journal*, 20(12):3092–3100, 2001.

[217] Stefan Canzar and Steven L Salzberg. Short read mapping: an algorithmic tour. *Proceedings of the IEEE*, 105(3):436–458, 2015.

[218] Tilman Schlotter, Sean Weaver, Csaba Forró, Dmitry Momotenko, János Vörös, Tomaso Zambelli, and Morteza Aramesh. Force-controlled formation of dynamic nanopores for single-biomolecule sensing and single-cell secretomics. *ACS nano*, 14(10):12993–13003, 2020.

[219] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[220] Christopher Maffeo, Han-Yi Chou, and Aleksei Aksimentiev. Single-molecule biophysics experiments in silico: Toward a physical model of a replisome. *iScience*, 25(5):104264, 2022.

[221] Brandon Ho, Anastasia Baryshnikova, and Grant W Brown. Unification of protein abundance datasets yields a quantitative saccharomyces cerevisiae proteome. *Cell systems*, 6(2):192–205, 2018.

[222] Annette Granéli, Caitlyn C Yeykal, Tekkatte Krishnamurthy Prasad, and Eric C Greene. Organized arrays of individual dna molecules tethered to supported lipid bilayers. *Langmuir*, 22(1):292–299, 2006.

[223] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022.

[224] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[225] Dan Cireşan and Ueli Meier. Multi-column deep neural networks for offline handwritten chinese character classification. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–6. IEEE, 2015.

[226] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv*, 1301.3781, 2013.

[227] Janani Durairaj, Mehmet Akdel, Dick de Ridder, and Aalt DJ van Dijk. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics*, 36(Supplement_2):i718–i725, 2020.

[228] Sem Tamara, Maurits A den Boer, and Albert JR Heck. High-resolution native mass spectrometry. *Chemical Reviews*, 122(8):7269–7326, 2021.

[229] Suji Lee, Kavyasree Chintalapudi, and Abraham K Badu-Tawiah. Clinical chemistry for developing countries: Mass spectrometry. *Annual Review of Analytical Chemistry*, 14:437–465, 2021.

[230] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[231] Jun Li, Haoqiu Wang, Lingfeng Mao, Hua Yu, Xinfen Yu, Zhou Sun, Xin Qian, Shi Cheng, Shuchang Chen, Junfang Chen, et al. Rapid genomic characterization of sars-cov-2 viruses from clinical specimens using nanopore sequencing. *Scientific Reports*, 10(1):1–10, 2020.

[232] Meriem Laamarti, Mohammed Walid Chemao-Elfihri, Souad Kartti, Rokia Laamarti, Loubna Allam, Mouna Ouadghiri, Imane Smyej, Jalila Rahoui, Houda Benrahma, Idrissa Diawara, et al. Genome sequences of six sars-cov-2 strains isolated in morocco, obtained using oxford nanopore minion technology. *Microbiology Resource Announcements*, 9(32):e00767–20, 2020.

[233] Sully Márquez, Belén Prado-Vivar, Juan José Guadalupe, Mónica Becerra-Wong, Bernardo Gutierrez, Juan Carlos Fernández-Cadena, Derly Andrade-Molina, Gabriel Morey, Veronica Barragan, Patricio Rojas-Silva, et al. Sars-cov-2 genome sequencing from covid-19 in ecuadorian patients: a whole country analysis. *medRxiv*, 2021.

[234] Muaaz Gul Awan, Abdullah Gul Awan, and Fahad Saeed. Benchmarking mass spectrometry based proteomics algorithms using a simulated database. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 10(1):1–8, 2021.

[235] Niranjan Nagarajan, Timothy D Read, and Mihai Pop. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics*, 24(10):1229–1235, 2008.

Summary

Although heterogeneity in properties and behavior of biomolecules at the single-molecule (SM) level is crucial for cellular functions, this subtlety has long been obscured by the bulk nature of common methods in biological research. With the advent of SM identification and sequencing methods for biological heteropolymers, the veil on this heterogeneity is beginning to lift. In particular, SM sequencing of nucleic acid chains has recently seen widespread adoption. Although SM protein analysis has yet to reach a similar state of maturity, the current pace of development indicates that this technology will likely catch up in the coming years. This thesis shows how bioinformatics and computational biology may steer this catch up process and the development of SM analysis of heteropolymers in general, over the coming years.

First, the concept of SM analysis of heteropolymers is introduced and the landscape of readout methods that is currently being explored is described. Special attention is given to the complexities of analyzing proteins using electrical and optical readouts, and why the properties of proteins make their analysis more difficult than that of NAs. This is followed in **chapter 2** by a closer look at electrical readout methods – currently one of the best developed methods for NA analysis – discussing the physical basis, implementation and a range of dedicated data analysis tools.

The subsequent three chapters describe several software tools that exemplify the ways in which bioinformatics can streamline development of new SM methodology in different stages of maturity. In the proof-of-concept stage, analysis software should be adaptable to a wide variety of experimental setups, while remaining intuitive to use by researchers without extensive instruction. FRETboard, the highly modular semi-supervised classification software described in **chapter 3** fulfills both these requirements for optical readouts.

Once a method has sufficiently matured to see adoption by a research community, it is likely to go through rapid iterations of improvements, making it difficult to track the current state-of-art. poreTally, the tool described in **chapter 4**, was designed to mitigate this issue for DNA nanopore sequencers, by making the repeated benchmarking of read and assembly quality and the sharing of results straightforward. Ultimately, a method that is developed to a stage of broad adoption requires analysis software that is tailored to a given purpose. For some applications this means that accuracy may be sacrificed to gain computational efficiency. This principle is implemented by baseLess, the analysis tool for electrical readouts of NAs described in **chapter 5**, which drastically reduces computational

hardware requirements by focusing on detection of specific sequences.

Bioinformatics and computational biology can also steer the development of experimental methodology, by simulating methods under development and indicating under which conditions the method is predicted to yield usable results. I describe two such simulations in chapters 6 and 7. In **chapter 6** coarse-grained molecular dynamics simulations are used to evaluate a novel SM protein fingerprinting analysis method employing fluorescence-based intermolecular distance measurements. Similarly, **chapter 7** describes a simulation study for Chop-n-Drop, a recently proposed method utilizing a proteasome-nanopore construct to digest a protein and estimate the molecular weight of the fragments to obtain a fingerprint.

Finally, this thesis concludes with a perspective on the future of SM protein analysis. Lessons are taken from SM NA analysis that may be applied to SM protein analysis, and recommendations to streamline future development from the perspective of computational biology and bioinformatics are provided.

Publications

De Lannoy, C.V.; De Ridder, D.; Risse, J. The long reads ahead: de novo genome assembly using the MinION. *F1000Research 2017, 6, 1083.*

De Lannoy, C.V.; Risse, J.; De Ridder, D. poreTally: run and publish *de novo* nanopore assembler benchmarks. *Bioinformatics 2018, 35(15), 2663–2664.*

De Lannoy, C.V.; Filius, M.; Kim, S.H.; Joo C.; De Ridder, D. FRETboard: Semisupervised classification of FRET traces. *Biophysical Journal 2021, 120(16), 3253–3260.*

De Lannoy, C.V.; Lucas, F.L.R.; Maglia, G.; De Ridder, D. *In silico* assessment of a novel single-molecule protein fingerprinting method employing fragmentation and nanopore detection. *iScience 2021, 24(10), 103202.*

De Lannoy, C.V.; Filius, M.; Van Wee, R.; Joo, C.; De Ridder, D. Evaluation of FRET X for single-molecule protein fingerprinting. *iScience 2021, 24(11), 103239.*

Noordijk, B.; Nijland, R.; Carrion, V.; Raaijmakers, J.; **De Lannoy, C.** baseLess: Lightweight detection of sequences in raw MinION data. *bioRxiv 2022.*

Götz, M.; Barth, A.; Bohr, S.S.R.; Börner, R.; Chen, J.; Cordes, T.; Erie, D.A.; Gebhardt, C.; Hadzic, M.C.A.S.; Hamilton, G.L.; Hatzakis, N.S.; Hugel, T.; Kisley, L.; Lamb, D.C.; **De Lannoy, C.V.**;Mahn, C.; Dunakara, D.; De Ridder, D.; Sanabria, H.; Schimpf, J.; Seidel, C.A.M.; Sigel, R.K.O.; Sletfjerding, M.B.; Thomsen, J.; Vollmar, L.; Wanninger, S.; Weninger, K.R.; Xu, P.; Schmid, S. A blind benchmark of analysis tools to infer kinetic rate constants from single-molecule FRET trajectories. *Nature Communications 2022, 13(1), 1–12.*

Acknowledgments

This is the part where I admit that nothing of this work, not the smallest step in this journey, would have been possible without the many supportive people around me. Here I will try to acknowledge all of these people by name, so if you are the type to read a book from cover to cover, I suggest you strap in.

A PhD only fares well by the grace of good mentors and frankly I could not have wished for better. Dick, Salima once mentioned that you are so impossibly present for everyone that you secretly must be triplets, acting as one person. The fact that I do not dare to fully refute this theory indicates what a force of nature you are. Thank you for all the guidance over the years. Judith, I must thank you as well, for guiding me at the start of my journey into single-molecule land with grounded advice and common sense. Being part of a consortium between three universities has been a blessing, as I got to know the very best people of each place. Chirlmin, thank you for introducing me to the wonders of single-molecule spectroscopy in Delft, and for your flavor of leadership; soft-spoken and visionary. Giovanni, you completed the consortium trifecta from Groningen. Thank you for getting me into single-molecule analysis in the first place; you set it all off with your lecture at KU Leuven on nanopore sequencing six years ago, and let me mess around in your lab to boot.

No less important have been the many peers along whom I had the pleasure of working, I thank you all for the helpful discussions and random banter, which often merged seamlessly. Raùl, your excellent organizational skills are only paralleled by your deadpan line delivery. Don't ever change either. Jay and Mehmet, both of you are equal parts ridiculously intelligent and kind, I'm grateful to know you. Miguel, our own personal Jesus, Vittorio, you _____ mongrel, Ronald, forever DM/eloquent curser, Barbara, blessed with endless imagination, Eef, prime purveyor of dry wit, Lotte, fellow cat herder and conceiver of cool tool names, honorary bioinformaticians Chiara and Elvira, Sina, Margi, Roven, Dirk-Jan, Kumar, David, Nike, Hannah, Joris, Victoria, Zach, Jorge, Mohammad, Farooq, Ehsan, Rens, Siavash, AJ, Justin, Harm, Sandra, Anne, Maria and Marie-Jose, thank you all for making the stay with the bioinformatics group so memorable. Jennifer, you left us too soon, but I will always remember your kindness to a rookie who tortured your servers from day one. Thanks for geeking out with us. Joo Group of TU Delft, as an adoptive member I owe you much as well. Mike and Raman, masters of the microscope, your relentless experimenting and quick troubleshooting ability made it all happen. Thanks for sharing the woes and victories! Sung Hyun, Sungchul, Laura, Victorija, Carolien, Ivo, Ilja, Margreet,

Dong Hoon, Bhagyashree, Kijun, Koushik, Moon Hyeok, thanks for accepting this bioinformatician into the fold. Students who came to me for thesis work, thank you for all your dedication. Thijs, Marijke, Virgil, Ben, Bart and Ben (again), you were all sharp as razors. I hope I was able to teach you some things, but I certainly learned much from you as well.

Outside academia, I must thank several people that helped me preserve sanity over the years, which is arguably just as important as writing a thesis. Charlie, camping master and DM par excellence, thanks for the mental health checks and just generally being you. Erwin and Conrad, fellow occupants of the inappropriately named Bioscoop app group, thanks for being there for me since Bachelor times in Leiden.

Pap en Mam, dankzij jullie onophoudelijke steun kon ik gaan waar ik wilde in het leven, in voor- en tegenspoed. Ik zeg het niet vaak genoeg, maar jullie zijn mijn voorbeelden, bedankt voor alles.

Thanks also to Nala, Jovi, Kara and Zeno, and the many other cats that make life great with their weird ways.

Infine, grazie Salima, grazie mille per essere con me. Grazie per aver illuminato le mie giornate con il tuo senso dell'umorismo unico e il tuo senso dell'avventura. non siamo stati molto lontani da Lovanio, eppure mi mostri ancora modi diversi di vivere. Questa tesi è finita, ma non vedo l'ora di scrivere nuovi capitoli e storie diverse con te.

This work originates as part of the research programme of the Foundation for Fundamental Research on Matter (FOM), and falls as of April 1, 2017 under the responsibility of Foundation for Nederlandse Wetenschappelijk Onderzoek Instituten (NWO-I), which is part of the Dutch Research Council (NWO).

Cover design by Carlos Victor de Lannoy
Printed by ProefschriftenMaken, Wageningen