Paving the way to single-molecule protein sequencing

Restrepo-Pérez, Laura; Joo, Chirlmin; Dekker, Cees

DOI
10.1038/s41565-018-0236-6

Publication date
2018

Document Version
Accepted author manuscript

Published in
Nature Nanotechnology

Paving the Way to Single-Molecule Protein Sequencing

Laura Restrepo-Pérez, Chirlmin Joo, Cees Dekker**

Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, van der Maasweg 9, 2629 HZ Delft, The Netherlands

*Correspondence: c.joo@tudelft.nl, c.dekker@tudelft.nl

Abstract

Proteins are major building blocks of life. The protein content of a cell and an organism provides key information for the understanding of biological processes and disease. Despite the importance of protein analysis, only a handful of techniques are available to determine protein sequences, and these methods face limitations, e.g. requiring a sizable amount of sample. Single-molecule techniques would revolutionize proteomics research providing ultimate sensitivity for the detection of low-abundance proteins and the realization of single-cell proteomics. In recent years, novel single-molecule protein sequencing schemes have been proposed, using fluorescence, tunnelling currents, and nanopores. Here we present a review of these approaches, together with the first experimental efforts towards their realization. We discuss their advantages and drawbacks, and present our perspective in the development of single-molecule protein sequencing techniques.

Introduction

Proteins are the workhorses in all living cells. Thousands of different proteins sustain all functions of the cell, from copying DNA and catalysing basic metabolism to producing cellular motion. Protein analysis can therefore provide key information for the understanding of biological processes and disease (**Box 1**). Compared to the impressive technical advances in DNA sequencing, the development of highly sensitive, high-throughput protein sequencing techniques lags severely behind. The only methods currently available for protein sequencing are Edman degradation, mass spectrometry, or their combination[1–3] (see **Box 2**).

Box 1. Genomic, transcriptomic, and proteomic analysis in diagnostics

When the human genome project was realized in 2003, sequencing an entire human genome would cost approximately 50 million dollars and would require 100 machines working for ~2500 hours. Today, thanks to the tremendous advances in DNA sequencing technologies, a human genome can be sequenced for only 1000 dollars using one machine working for ~72 hours[4,5]. DNA sequencing is thus becoming a routine technique in clinics allowing the collection of genetic information from patients at reasonable time and cost.

The challenge ahead is the interpretation of the data gathered from DNA sequencing with respect to the health condition of patients. A large gap resides between genotype and phenotype. Transcriptomics studies are often used as a first bridge, which provides information about which genes are actively being expressed. However, the gap still persists as mRNAs levels do not simply correlate to protein levels due to factors such as the variability in translational efficiency of different mRNAs, and the difference between mRNA and protein lifetimes[6]**.** Moreover, protein post-translational modifications further influence the function and structure of proteins.

Proteome analysis is therefore key to understand biological processes and their dynamic nature[7,8]. After all, proteins dictate most biological functions and are directly related to the phenotype of a cell. So, while genomics offers a quick glimpse, much like looking at the menu in a restaurant, proteomics brings you inside the heart of the kitchen, to closely examine what the food looks like and how it tastes.

The current gold standard for protein sequencing is mass spectrometry[9–12]. The technique, however, has fundamental drawbacks in terms of its limit of detection and dynamic range[13]. Human samples are extremely complex, comprising a wide range of protein concentrations. In human plasma, for example, the concentration of proteins can vary from few picograms per millilitre (Interleukin 6) to few milligrams per millilitre (albumin)[14,15]. Therefore an exceedingly high dynamic range (~10^9) is necessary for comprehensive proteome analysis[14,16]. State-of-the-art mass spectrometers are limited to a dynamic range of ~10^4 to 10^5.[14,16] Another drawback of the instrument is its detection limit, which hinders biomarker discovery and translates into the need for large amounts of sample. If we consider a protein that is present in a cell in a low copy-number (less than 1000 molecules per cell)[17], millions of cells are required to reach the limit of detection of the instrument (0.1 to 10 femtomole)[18–20]. Mass spectrometry is thus far away from comprehensive single-cell analysis.

The spectacular advances in DNA sequencing technology, where even single DNA molecules can be sequenced, have inspired dreams of novel technologies for protein sequencing. However, the search for such protein sequencing methods is not trivial due to the complex nature of proteins. Proteins are built from 20 distinctive amino acids, while DNA is comprised of only four different bases. Independent of the read-out method of choice, the detection of 20 distinguishable signals is a tremendous challenge. Moreover, DNA samples with low concentrations of analyte can be amplified using polymerases, whereas protein sequencing platforms cannot benefit from such amplification since there is not PCR-like

amplification method for proteins. Protein sequencing techniques that would read the exact sequence of individual proteins at the single-molecule level could bring a revolution to proteomics, providing the ultimate sensitivity for the detection of low abundance proteins. Moreover, such a method would enable single-cell proteome studies with higher capabilities than current methods[21–25].

In this Review, we present an overview of the exciting nascent field of single-molecule protein sequencing. Several approaches for protein sequencing at the single-molecule level have emerged in the past few years. These new ideas run from renovating Edman degradation and mass spectrometry, through repurposing single-molecule DNA sequencing platforms for protein sequencing, to developing entirely new molecular devices. The proposed methods are based on single-molecule techniques such as nanopores, fluorescence, and tunnelling currents across nanogaps (**Figure 1**). We describe the schemes proposed so far and discuss their advantages and drawbacks. First experimental efforts and proof-of-principle experiments towards their realization are discussed.

Box 2. Current Protein Sequencing Methods

Edman degradation

Invented by Pehr Edman in 1950, Edman degradation allows the ordered identification of the amino acid sequence in a protein from the N- to the C-terminus[26]. It performs cyclic chemical reactions that label, cleave, and identify the amino acid at the terminus of a protein, one at the time (**Figure B1**). In the first step of the reaction, the Edman reagent (phenylisothiocyanate PITC) reacts with the amino group at the N-terminus of the protein under mild basic buffer conditions. The modified N-terminal amino acid is removed as a thiazolinone derivative under acidic conditions. This derivative is then identified using chromatography.

Edman degradation is a useful tool for sequencing, but it is limited to the analysis of purified peptides which are shorter than ~50 amino acids. It cannot be used for the analysis of complex protein mixtures, such as those present in most biological samples. Additionally, each degradation cycle can take approximately 45 minutes[27], making the process extremely time-consuming. N-terminus modifications can also interfere with the process. For example, if the N-terminus of the peptide is acetylated (a common post-translational modification), the reaction cannot take place, prohibiting protein sequencing.

(Box continues on next page)

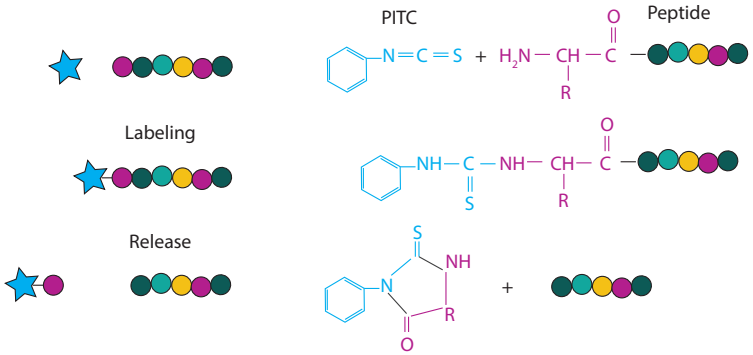

Figure B1. Schematic of Edman degradation reaction showing the process of labelling and cleavage of the amino acid in the N-terminus of the peptide. PITC stands for phenylisothiocyanate.

Mass spectrometry

Since the 1980s, with the discovery of new ionization techniques (MALDI and ESI), mass spectrometry has evolved into an important analytical tool for the life sciences[10]. For deep protein analysis, the introduction of shotgun proteomics marked an important step for the study of samples containing protein mixtures[28]. In a typical experiment proteins are digested into peptides and separated according to hydrophobicity and charge using chromatography (**Figure B2**). As peptides elute from the column, they are ionized and analysed according to their mass-to-charge ratio using tandem mass spectrometry.

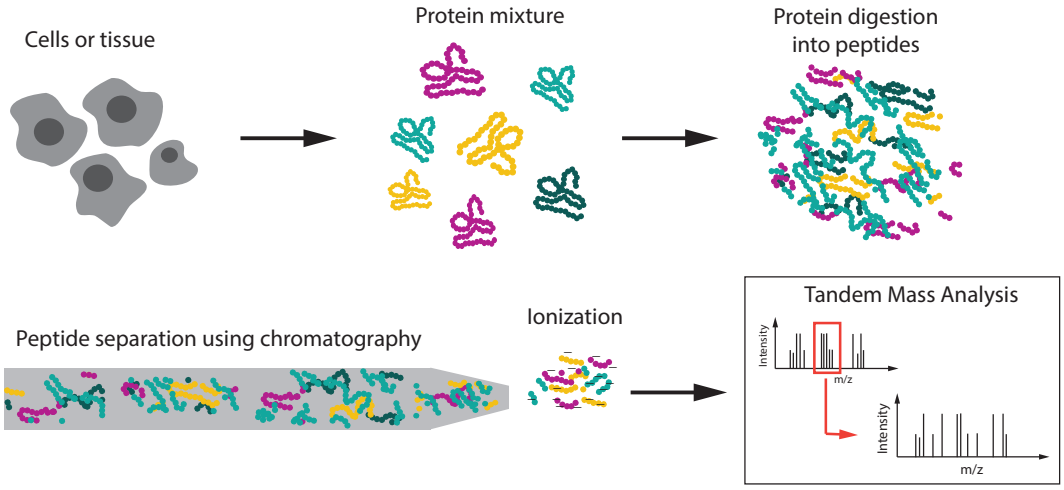

Figure B2. Workflow of proteome analysis with mass spectrometry. Proteins are extracted from cells or tissues and digested into peptides. The peptide mixture is separated using chromatography. Peptides are ionized and analysed using tandem mass spectrometry.

Protein Fingerprinting Using Fluorescence

Fluorescence techniques have been central for the development of high-throughput DNA sequencing devices. In systems such as those of Illumina[29], Pacific Biosciences[30], and Helicos[31], DNA is *de novo* sequenced by monitoring the incorporation of fluorescently labelled nucleotides during strand replication. The development of a *de novo* protein sequencing method based on fluorescence faces enormous challenges. Major constrains are the lack of organic fluorophores for the detection of 20 different amino acids without substantial signal crosstalk, and the absence of a suitable chemistry to specifically label all 20 amino acids[32].

Recently, simplified schemes, in which only a small subset of amino acids is fluorescently labelled and detected, have been proposed. If demonstrated, these could lead to the development of protein identification methods with single-molecule sensitivity[33,34]. These approaches resemble optical mapping of DNA, where partial sequence information is sufficient to identify certain characteristics of a genome or to identify different pathogens[35]. Similar to how optical mapping has served as a complementary lower-resolution technique to DNA sequencing, protein fingerprinting could constitute a complementary technique to *de novo* protein sequencing.

In 2015, Joo and colleagues proposed a fingerprinting scheme based on the detection of two types of amino acids[33]. In their approach, the cysteine (C) and lysine (K) residues of a protein are labelled and sequentially detected. This sequence of C's and K's (or CK sequence) can then be used to identify the protein of interest using a protein database (**Figure 2b)**. To read the CK sequence, an unfoldase called ClpX is immobilized on a single-molecule surface and used as a protein scanner. This molecular motor recognizes tagged polypeptides and unfolds them while translocating them through its internal cavity. If the enzyme is labelled with a donor fluorophore and the substrate contains acceptor dyes in its cysteines and lysines, FRET occurs as each of these amino acids approaches the ClpX constriction, generating a CK read in a string of two different acceptor signals (**Figure 2a**).

The feasibility of this CK fingerprinting approach was computationally assessed using a human protein database containing ~20,000 protein entries[33]. CK sequences were generated computationally taking into consideration the most common errors expected during experimental readings. These generated CK sequences were compared to the database, and the probability of retrieving an original sequence was calculated based on the accuracy of the matches. Considering a 10% error level in the readings, approximately half of the protein sequences could be correctly retrieved. When additional parameters, such as the distance between C's and K's were considered (**Figure 2b**, CK-dist read), the method could accurately identify a major percentage (>70-80%) of proteins even when high error rates (20-30%) were considered (**Figure 2c**).

A proof of concept was experimentally demonstrated by the same group this year[36]. Using a donor-labelled ClpP (the proteolytic chamber that binds ClpX), the authors sequentially read out FRET signals from acceptor-labelled substrates. They could fingerprint 29-, 40-, 51-amino acid long peptides, and a monomeric (119 amino acids) and a dimeric (210 amino acids) titin protein. The repurposed ClpXP showed a constant translocation speed and uni-directionality, features that are suitable for reliable fingerprinting. Note that a similar

fingerprinting system was proposed and experimentally demonstrated by Goldman and colleagues, using a labelled ribosome to monitor the production of specific proteins inside the cell as a way to gain information on protein expression location and levels[37,38].

A different method is pursued by Marcotte and colleagues, in which peptide fingerprinting is accomplished using a single-molecule version of Edman degradation[34]. Unlike conventional Edman degradation methods, the single-molecule detection allows for analysis of mixed populations. In this approach, proteins are digested into peptide fragments (~10-30 amino acids long) and specific amino acids are labelled with fluorophores of distinguishable colours. The labelled peptides are immobilized on a surface, and fluorescence microscopy is used to monitor each cycle of Edman degradation at single-molecule resolution (**Figure 2d)**. Each degradation cycle removes the N-terminal amino acid of the peptide, so that the sequence of labelled amino acids can be detected by monitoring the change of the fluorescence intensity in each cycle. The decrease in fluorescence after a degradation cycle indicates that a labelled amino acid has been cleaved. The cleaved amino acid can be identified using spectral information (**Figure 2e**).

Computer simulations were used to investigate the probability of detecting proteins from the identification of a unique peptide sequence using Marcotte's fingerprinting method[34]. Different immobilization, labelling, and cleavage strategies were evaluated, and it was determined that at least four different labelled amino acids are required to identify 98% of the human proteome[32].

The fingerprinting schemes proposed here take advantage of the fact that proteins can be identified using incomplete sequence information. The approach proposed by Joo and colleagues reads full-length proteins and therefore requires simple two-colour labelling of substrates. The main limitation of this approach is the requirement of a recognition tag in the N- or C-terminus of the substrate for unfoldase recognition. It seems well possible to devise ligation schemes to add such a tag to all proteins in a mixture or to engineer the enzyme to allow recognition of any protein coming from cellular preparations and other biological samples. Marcotte's approach to fingerprinting benefits from an entirely chemical approach, which can be beneficial for commercialization purposes. At the same time, the Edman degradation reaction faces two main challenges. First, the harsh conditions required for the reaction will demand for a careful selection of fluorophores, and a set of adaptations to a conventional TIRF microscope[39]. Second, each cycle of Edman degradation can take approximately 45 minutes, making the sequencing process extremely slow. Havranek and colleagues are currently working on an alternative approach to Edman degradation in which an enzyme has been designed that is capable of cleaving off amino acids, one at the time, from the protein N-terminal[40]. The use of this enzyme, called edmanase, may allow Edman degradation to proceed under physiological conditions, and potentially at a faster pace.

Fluorescence fingerprinting may play a crucial role in the development of fast techniques for parallel protein identification and analysis. Millions to billions of single molecules can be immobilized and monitored together, opening the door to high-throughput assays. Single-molecule protein identification using fluorescence could complement *de novo* protein sequencing methods, improving the sensitivity of current bulk identification techniques such as antibody microarrays or mass-spectrometry protein identification based on peptide

fingerprints. The improved sensitivity of these methods brings important advantages for applications such as biomarker detection for disease diagnosis.

Protein Sequencing Using Tunnelling Currents

The idea of using tunnelling currents to measure on single molecules was first conceived in the 1970s[41]. Tunnelling currents are measured between two metal electrodes separated by a gap that ranges from a few angstroms to a few nanometers (**Figure 3a,d**). When individual molecules pass through the nanoscopic gap, a change in the tunnelling current is measured. This current modulation can be used to determine which molecule is transiently residing in the gap in real time. With the invention of the scanning tunnelling microscope (STM) in the 1980s, the possibility to realize this idea became clear and led to the development of a new field named molecular electronics[42–44]. In recent years, this technique has evolved to study a variety of biomolecules aiming towards DNA and RNA sequencing[45–49] (for a detailed review of these developments see Ref. 47, 48). In a similar way, interest has emerged in the study of amino acids and peptides in an urge towards protein sequencing. In this section, we present a review of these developments.

In 2014, the Lindsay group reported the first measurements of amino acids and short peptides using tunnelling currents[50,51]. They demonstrated the sensitivity of their approach by analysing three sets of amino acids with minor structural differences: glycine vs. its methylated form called sarcosine, the enantiomers of asparagine (L- vs. D- asparagine), and the isobaric amino acids leucine vs. isoleucine. Their experimental set-up consisted of two palladium electrodes, separated by a gap of 2 nm. The electrodes were functionalized with a recognition molecule (4(5)-(2-mercaptoethyl)-1H-imidazole-2-carboxamide), which was covalently bound to the electrodes. The recognition molecule interacted temporarily with the analyte to orient the molecule and thus provided a better defined current path (**Figure 3a)**. When amino acids were introduced, the transient interactions between each amino acid and the recognition molecule were detected as a train of current spikes (**Figure 3b**). Using two-dimensional maps of the current amplitude and the spike shape, the amino acids analysed in each set could be discriminated with an accuracy of 80% or higher (**Figure 3c**).

A subsequent study was reported by the Kawai group in which all 20 amino acids and phosphotyrosine were tested using tunnelling currents with a different experimental set-up [52]. In their study, smaller gaps of 0.70 nm and 0.55 nm were created using gold break junctions. The small size of the gap allowed the detection of amino acids without a recognition molecule (**Figure 3d**). The 0.70-nm gaps produced detectable signals for eight (Y, F, W, H, P, E, D, I) out of the 20 different amino acids, while smaller gaps of 0.55 nm produced signals for nine (P, H, E, D, I, K, C, L, M) amino acids. In total, 12 out of the 20 amino acids could be recognized; the rest did not produce a detectable signal. When one of the detectable amino acids was introduced in the measuring set-up, peaks in the current trace were observed indicating the transient presence of an individual molecule between the electrodes (**Figure 3e**). The amplitude and duration of each peak was used to characterize each amino acid as shown in the scatter diagram in **Figure 3f**. Seven amino acids showed distinctive signals and show potential for their differentiation in complex mixtures; the remaining five produced indistinguishable signals. The detection of post-translational modifications was also demonstrated using 0.70-nm gaps. Tyrosine and phosphotyrosine produced distinctive signals

and mixtures of them yielded two populations in the amplitude histograms. Lastly, using the same approach, short peptides containing tyrosine and phosphotyrosine could be distinguished.

The recognition tunnelling approach used by Lindsay and colleagues shows the remarkable sensitivity of quantum tunnelling currents. This technique can discriminate isomers and molecules with minor structural differences that are indistinguishable by other techniques such as mass spectrometry. The downside of this method is the non-trivial complexity of the data. Each molecule can orient in many different ways within the junction, and exhibits significant translational and rotational fluctuations, leading to considerably different current signals. Therefore machine-learning algorithms may be necessary to distinguish each molecule considering the multiple conformations that can be observed.

The study of the Kawai group presented a systematic characterization of different amino acids and short peptides. Out of the 20 amino acids studied, 7 amino acids generated distinguishable signals. This represents a promising step towards amino acid discrimination for protein sequencing. Arrays containing junctions of different sizes might increase the number of amino acids that are detectable and increase the possibility to distinguish amino acids in a mixture. Technical improvements in the experimental set-ups and fabrication processes would facilitate this task. Taniguchi and colleagues, for example, recently showed that extra coatings on the nanoelectrodes could bring improvements in terms of the signal-to-noise ratio and bandwidth of the measurements[53,54].

To make this proof-of-concept into a sequencing tool, measurements of tunnelling currents should be coupled with a mechanism that threads a polypeptide through the gap in a controlled way. An exopeptidase or other molecular motor could be adapted to translocate the polypeptide through such an electrode gap. Alternatively, electrophoresis, electroosmosis, or a pressure difference could be used as a driving mechanism for molecules if the tunnelling device is coupled to a nanopore. Several groups have reported first experimental efforts in this direction[55–58].

Protein Sequencing Using Nanopores

In 2014, Oxford Nanopore Technologies (ONT) announced the release of the first single-molecule DNA sequencing device based on nanopores[59–62]. These pocket-size devices are revolutionizing DNA sequencing by allowing extremely long reads and *in situ* detection at remote laboratories (even in outer space)[59,63]. In a nanopore experiment, an insulating membrane containing a nanometer-sized pore is placed between two electrolyte-filled compartments. When a voltage is applied across the membrane, an ionic current flows through the nanopore. As individual molecules translocate through the pore, a modulation in ionic currents is observed, which provides structural information about the molecule of interest[64–66]. Using this principle, biopolymers can be sequenced as each individual component of the chain sequentially transverses the nanopore constriction.

Nanopores have proven their potential for DNA sequencing[62,67]. Exploiting nanopores for single-molecule protein sequencing is the next frontier. This is by no means an easy task, as numerous challenges need to be tackled in order to sequence a protein with a nanopore. First, amino acid residues vary widely in charge distribution, unlike DNA that is essentially uniformly charged. Electrophoresis-driven unidirectional translocation of polypeptides through nanopores thus cannot simply be employed. Second, most proteins are folded in their native state. Disruption of their secondary and tertiary structure is necessary to thread them through a nanopore. Third, protein sequencing requires distinction of 20 different amino acids, a five-fold larger number than the four bases in DNA sequencing.

First translocations of polypeptides through nanopores were performed using peptides of only 20 to 30 amino acids[68–72]. Short peptides lack stable tertiary structure and can translocate without the need of denaturing agents. In these studies, peptides containing specific motifs such as β-hairpins, α-helices, or collagen-like helices were analysed using α-hemolysin and aerolysin nanopores. This research elucidated important aspects about the kinetics of polypeptide translocation and emphasized the crucial role of peptide-nanopore interactions during the passage of the molecule. In particular, the detailed work presented by the Bayley group on helical peptides containing the $(AAKAA)_n$ sequence provided key insights into the process of protein capture and partitioning into the nanopore[70].

While the translocation of peptides continues to be a valuable model system to understand basic steps in the complex process of protein translocation[73,74], the final end of a nanopore-based protein sequencer is to read entire proteins, which requires protein denaturation. Multiple chemical and physical methods have been proposed for protein unfolding in nanopore analysis. Several groups have shown the successful unfolding and translocation of proteins through solid-state nanopores using strong denaturants such as urea, sodium dodecyl sulphate (SDS), or guanidine hydrochloride (GdnHCl)[75–77]. Translocation of proteins through biological nanopores using denaturants has also been achieved[78–80]. In this context, solid-state nanopores have an advantage over biological nanopores displaying higher stability when exposed to extreme buffer conditions (8 M urea, 6 M GdnHCl, or 1 % SDS).

Biological channels are more susceptible to denaturing conditions than solid-state devices, but can remarkably withstand concentrations of up to 4 M urea and 1.5 M GdnHCl[81]. These concentrations are sufficient to break the structure of some protein substrates and allow

translocation. For example, the pioneering work of Auvray and colleagues (**Figure 4a,b**), which showed protein unfolding and translocation through alpha hemolysin for the first time, was done using the maltose binding protein (MBP), which could be unfolded at low denaturant concentrations (0.8 M GdnHCl)[78].

Physical methods such as high temperature have been used to unfold proteins in both solid-state and biological nanopores[82,83]. Pelta and colleagues studied the thermal denaturation of an MBP variant in a temperature range from 20ºC to 70ºC in both alpha hemolysin and aerolysin nanopores[83]. Temperature facilitates protein unfolding, but speeds up translocation dynamics, which makes sequencing more challenging. In a similar way, two research groups have shown that high voltages help stretch proteins during the movement through solid-state nanopores[84–86]. These approaches are not compatible with biological nanopores due to the electroporation of the lipid bilayer at high voltages (~0.4V), and also cause an increase in translocation speed.

A major roadblock for the development of a protein sequencer with nanopores is the non-uniform charge distribution of amino acid residues. Unlike DNA that is uniformly charged and moves through a nanopore by electrophoretic forces, proteins carry different local charges. It is therefore not well-defined if electrophoretic or electroosmotic forces on the protein dominate the transport (unless it is set by the electroosmotic force due to ions at the nanopore surface)[87,88]. One way to address this issue is to use SDS as a denaturant. SDS not only unfolds proteins, but also wraps them around with a homogeneous negative charge given by the sulphate groups in the head of the detergent. Timp and colleagues used SDS to enforce proteins through pores with subnanometer diameters, hinting at the potential of using a nanopore for differentiating individual amino acids (**Figure 4c**)[89,90]. A more comprehensive understanding of the effect of SDS on protein unfolding and translocation was presented by our group[77]. Experiments showed that SDS could unfold stably folded proteins such as titin and β-amylase (**Figure 4d**). Additionally, a consistent direction of translocation was induced by the electrophoretic force, thanks to the negative charge conveyed by SDS.

An alternative approach to control the direction of translocation is to attach an oligonucleotide strand to the N- or C-terminus of a protein. The negative charge carried by this lead sequence drags the polypeptide in the direction of the electrophoretic force[91–95]. This principle was first used by Bayley and colleagues to study the translocation of thioredoxin through α-hemolysin[91,92]. In their work, a 30-mer oligonucleotide was attached to the C-terminus of the protein and upon adding the substrate to the *cis* compartment, a repetitive pattern with multiple current levels was observed, which corresponded to the capture of the DNA tag, the local unfolding of the C-terminus, and the unfolding of the remaining of the protein (**Figure 5a**). The partially unfolded intermediate in which the C-terminus of the protein was locally unfolded and translocated through the constriction of the nanopore was further used to discriminate between unphosphorylated, monophosphorylated and diphosphorylated proteins[93]. Other groups have also recently used this approach. Lindsay's group developed a simple and effective click chemistry to facilitate the tagging reaction, while Pelta and colleagues used a DNA lead in a protein to present a direct proof of protein translocation using amplification by PCR[94,95].

In all the studies presented this far, the translocation of proteins occurs at time scales faster than 1 millisecond, which is too fast for sequencing purposes. Indeed, single-protein translocations characteristically occur very fast[96]. Control of the translocation speed will be necessary to guarantee ample time for the accurate reading of different amino acids by a nanopore.

The controlled and unidirectional movement of DNA through a nanopore using helicases or polymerases marked a breakthrough in the development of a nanopore-based DNA sequencer. Akeson and colleagues proposed a similar approach for proteins[97,98]. In their work, a motor enzyme, ClpX, unfolds and pulls the polypeptide chain in a controlled manner through α-hemolysin. ClpX translocates proteins at a speed slow enough for sequencing (80 amino acids per second), with defined step-sizes, and it generates a strong enough force (~20 pN) to unfold proteins[99]. In their experimental scheme (**Figure 5b**) a lipid bilayer containing α-hemolysin separates two compartments. The *cis* side contains a protein known as Smt3, which is modified with a 65-amino acid negatively charged extension and an ssrA tag. The ssrA tag is necessary for ClpX recognition and the 65-amino acid extension is used as an unstructured anchor that orients the protein and allows the ssrA tag to be exposed to the *trans* side where ClpX is added. Time traces showed the process of substrate capture and translocation by ClpX. In a follow-up study[98], a machine-learning algorithm with three parameters (dwell time, average current amplitude, and standard deviation of the current amplitude) was used to distinguish different domains as well as variants of those domains such as mutations or truncations.

This approach overcomes two critical requirements for protein sequencing using nanopores: protein unfolding and controlled translocation of the substrate. The main drawback of this method is the need to add a polypeptide extension in the substrate. This could, however, be overcome by chemically attaching a polypeptide to the N-terminus of proteins. Other approaches have been proposed, but lack experimental proof[100–103]. Sampath proposed the use of a double pore system in which two nanopores are placed in series[100]. As the polypeptide transverses the first pore, it is cleaved by an exopeptidase, and the amino acids released by the enzyme are then analysed with a second nanopore. DiVentra and colleagues proposed the use of perpendicular nanochannels in which a protein is stretched in the longitudinal direction, while ionic current is recorded transversally[101]. Aksimentiev and colleagues proposed the use of graphene to control polypeptide translocation. Graphene and other 2D materials are proposed as attractive nanopore membranes since they can be atomically thin, thereby improving the spatial resolution required to detect individual amino acids[57]. Using molecular dynamics simulations, they showed that proteins and peptides collapsed on top of a graphene membrane by the surface absorption of amino acids, leading to a slow stepwise motion of amino acids into a nanopore[102].

There is also a noticeable attempt of repurposing nanopores for improving mass spectrometry. Stein and colleagues proposed the use of solid-state nanopores to create a renewed version of a mass spectrometer, in which the electrospray ionization, conventionally done with micrometre-sized nozzles, is initiated from a nanopore. This could potentially allow proteins to be sequenced if they are fragmented as they pass through the nanopore and individual amino acids are sequentially ionized and detected[103]. For a more detailed description of efforts in improving the sensitivity of mass spectrometry, we refer to other reviews[19,104].

In summary, great advances have been presented with the nanopore approach towards sequencing peptides and proteins. It is an extremely active field of research, and therefore significant advances are anticipated for the development of a protein sequencer in the coming years. An advantage that a nanopore sequencer could provide is the possibility to perform long reads. Traditional sequencing methods such as Edman degradation and mass spectrometry rely on the digestion of proteins into short peptides, but nanopore devices would allow sequencing of full-length proteins. A major challenge is the control of the polypeptide translocation speed. Different approaches are being explored at the moment, and it is very likely that enzyme-assisted translocations will command this step, as was the case for DNA sequencing. Exploring a pool of unfoldases beyond ClpX will be a critical step to accomplish this aim.

Outlook

The human genome project opened the door to exciting years of genomic research. The coming years will see significant progress in other omics, especially proteomics. In this area, the development of single-molecule approaches will be key for achieving the sensitivity and dynamic range required for protein analysis. Colossal efforts are on-going in the fields of single-molecule fluorescence, tunnelling currents and nanopores. In this Review, we presented the main approaches proposed up to now for single-molecule protein sequencing, with their strengths and limitations. **Table 1** summarizes the different schemes presented, taking into consideration relevant criteria for the development of a protein sequencer, such as read length, and the possibility to perform *de novo* sequencing.

Table 1. Summary of single-molecule protein sequencing approaches taking into consideration their potential to read full-length proteins and perform de novo sequencing.

	Method	**Read length**	**Potential for *de novo* sequencing**	**Labelling required**	**Proof of concept**
Fluorescence	FRET scanning using ClpX	Full length	No	Yes	Computational (Yao et al[33]) Peptide analysis (van Ginkel[36])
	Edman degradation	A few amino acids	No	Yes	Computational (Swaminathan et al[34])
Tunnelling current	Recognition tunnelling	Full length if coupled with a nanopore or enzyme	Yes	No	Single-molecule measurements (Zhao et al[51])
	Sub-nanometer break junctions	Full length if coupled with a nanopore or enzyme	Yes	No	Single-molecule measurements (Ohshiro et al[52])
Nanopore	Solid-state nanopore	Full length	Yes	No	Single-molecule measurements (Li et al[76], Talaga et al[75], Timp et al[89], Restrepo-Perez et al[77])
	Graphene nanopore	Full length	Yes	No	Computational (Wilson et al[102])
	Biological nanopore	Full length	Yes	No	Single molecule measurements (Rodriguez-Larrea et al[91,92])
	Biological nanopore coupled with an enzyme	Full length	Yes	No	Single-molecule measurements showing controlled translocation (Nivala et al[97,98])

We anticipate that first single-molecule protein identification systems may appear as soon as within five years. First systems will most probably rely on a fingerprinting scheme such as those proposed by Marcotte's and Joo's groups. Marcotte's approach has the advantage of relying entirely on chemical reactions, which could lead to a robust device for *in situ* analysis. Major disadvantages of this approach are the complexity of its labelling scheme, its slow speed, and the fact that only short peptides can be analysed. Alternatively, the scheme proposed by the Joo group relies on a simple labelling scheme which can be used for the analysis of full-length proteins, but on the down side, unfoldase engineering or substrate pre-processing need to be worked out for substrate recognition. Both methods need to overcome the challenge of reading multiple fluorophores with minimal error.

Nanopore research is moving fast in the direction of protein analysis and protein sequencing. A nanopore-based protein sequencer has the potential to be commercialized in the next decade. The main challenges revolve around the controlled translocation of proteins through

the nanopore and the read-out. Akeson's approach, in which a ClpX enzyme was used to translocate a polypeptide through an α-hemolysin nanopore, is currently the only system in which a protein is unfolded and transported in a controlled way through the nanopore. The large levels of noise observed in their signals, however, obstructed the identification of specific amino acids. As has become clear from high-resolution DNA sequencing[59,105], alternative configurations schemes and possibly different enzymes should be explored.

A remaining question is whether the measurement of ionic currents will provide the sufficient resolution for the identification of 20 amino acids using nanopores. The experimental results from Lindsay and Kawai indicate that tunnelling currents are extremely sensitive, and can differentiate molecules with minor structural differences. Thereby, the integration of a nanopore system for controlled transport with the sensitive measurement of tunnelling currents is an attractive alternative that would potentially allow single-molecule *de novo* protein sequencing.

A major aim of a single-molecule protein sequencer would be the development of a tool for single-cell analysis. Current attempts to single-cell proteomics[21–24], such as mass cytometry[25], rely on labelled antibodies. The reduced availability of highly specific antibodies and distinguishable labels limits these techniques to the detection of 10 to 40 proteins per cell, a minute fraction of the proteome. Single-molecule detection methods will not require such a preparatory step, and could, in principle, detect thousands of proteins from individual cells. A critical aspect that needs to be resolved is the manipulation and extraction of proteins from single cells without substantial losses or biases[22]. Recent advances in microfluidic devices, where proteins from single cells have been extracted and labelled on chip[17], show first steps towards this goal.

The realization of a single-molecule protein sequencer is technically very challenging. If realized, however, it would revolutionize proteomics research by facilitating the identification of low abundance proteins and achievement of true single-cell proteomics. Low abundance proteins are crucial in biomedical research as they allow the identification of disease-specific biomarkers[106]. Moreover, sensitivity from single-molecule detectors could allow access to the so-called human "dark proteome". The dark proteome comprises approximately 3000 human proteins that have never been directly identified, despite evidence of their existence in genetic or transcriptional information[107]. Besides protein identification, the detection of low abundance proteins can be beneficial for the study of post-translational modifications, reducing the need of complex enrichment processes. Finally, the possibility to perform single-cell proteomic analysis opens the possibility for exciting proteomics research, allowing scientists to study the change in protein expression of individual cells under specific stimuli.

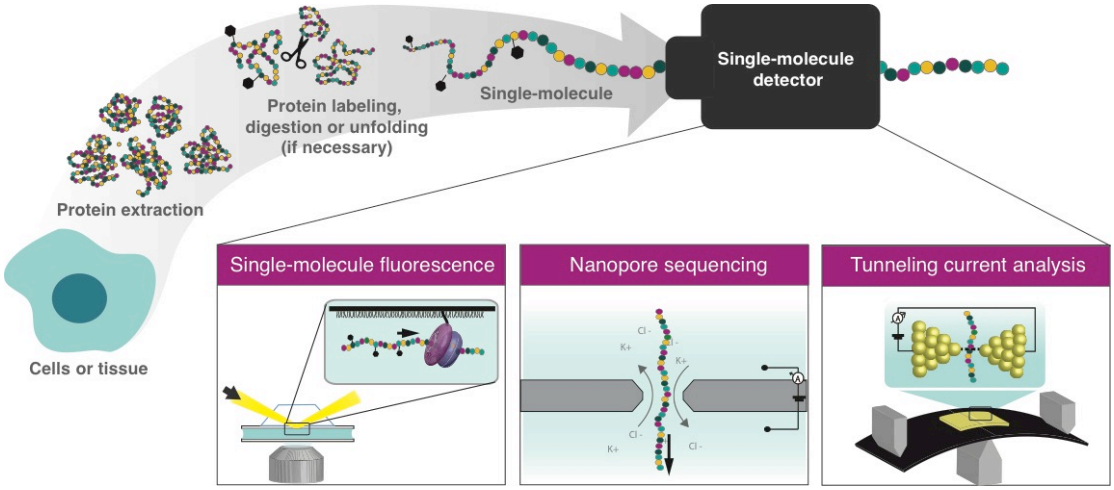

Figure 1. Schematic of the single-molecule protein sequencing workflow with fluorescence, nanopores, or tunnelling currents. In a typical experiment, proteins are extracted from a biological sample or even a single cell, then labelled, unfolded and partly digested (if necessary), and finally, each molecule is sequenced with a single-molecule technique.

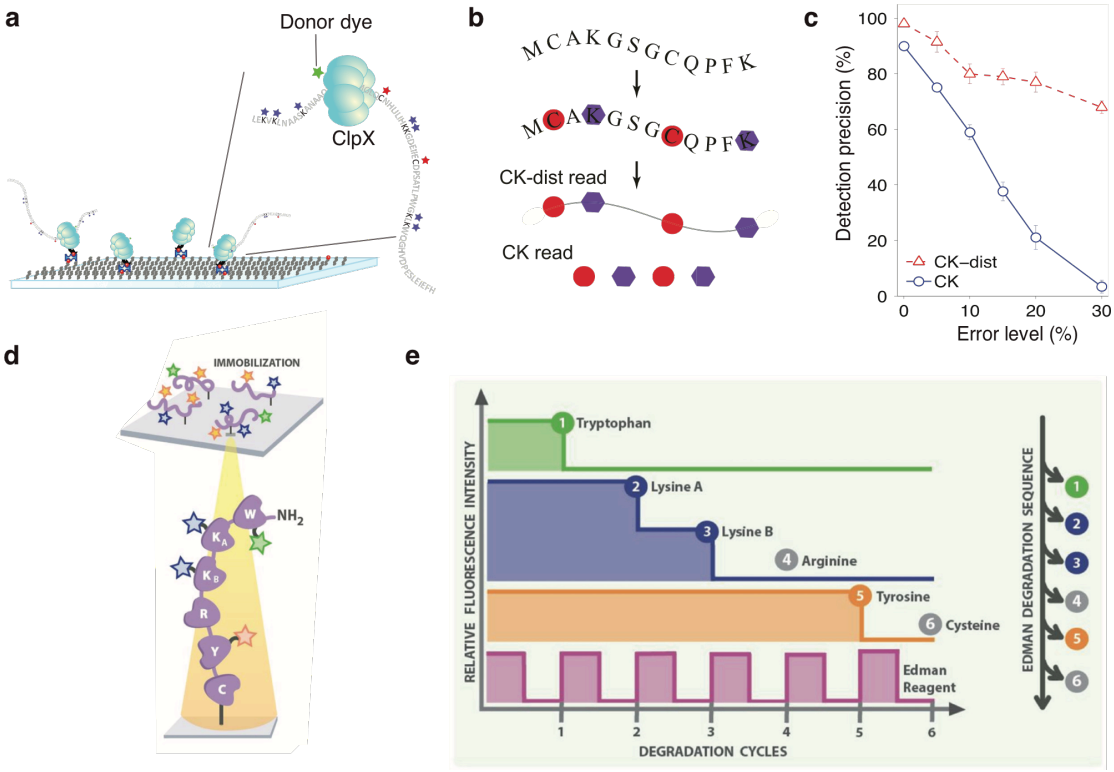

Figure 2. Protein fingerprinting schemes using fluorescence. (a) Scheme proposed in reference [33], in which a labelled unfoldase is immobilized in a surface and used to scan protein substrates. (b) Cysteines and lysines of the protein substrate are labelled and FRET is detected upon the translocation of these residues. The CK sequence is then compared to a protein database. The CK read, corresponds to the sequence of cysteines and lysines residues. The CK-dist read incorporates the distance between these amino acids (c) Graph of the detection precision (number of true positives divided by the number of read-outs returned by the algorithm) vs. error level (number of errors divided by the fingerprint length). (d) Scheme proposed by reference [34]. In this approach, labelled peptides are immobilized and subjected to sequential cycles of Edman degradation. The lost in fluorescence after each cycle is used to determine the sequence. Panels a, b, and c were adapted from ref [33]; panels d and e from ref [34].

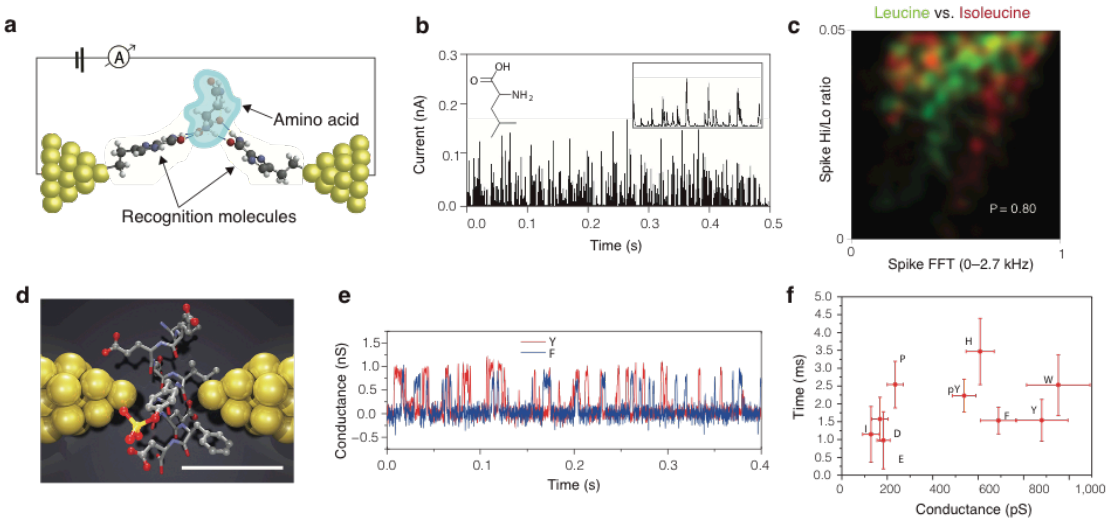

Figure 3. Amino acid and peptide characterization with tunnelling currents. (a) Recognition-tunnelling scheme where STM-coupled palladium electrodes are functionalized with recognition molecules. (b) Typical current vs. time trace obtained for the measurement of an amino acid (here Leucine). (c) Two-dimensional plot of probability density using two different FFT features for Leucine (green) and Isoleucine (red). (d) Schematic of the operating principle: A molecule is sandwiched between two gold nanogap electrodes created using mechanically controlled break junctions (MCJB). Scale bar 1nm. (e) Conductance vs. time traces obtained for measurements of the amino acids Y and F. (f) Scatter plot of time vs. conductance for different amino acids measured in a 0.55nm gap. Panels a, b, and c were adapted from ref [51]; panel d, e and f adapted from ref [52].

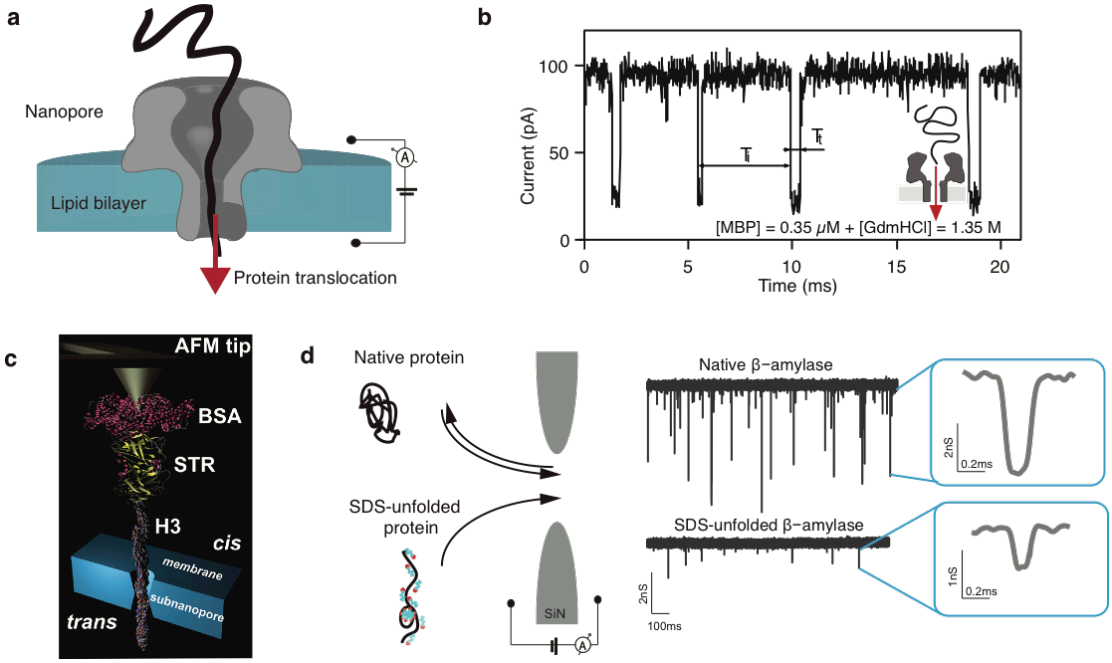

Figure 4. Translocation of peptides and unfolded proteins through nanopores. (a) Schematic representation of a biological nanopore set-up. (b) Representative current traces when GdnHCl was used for unfolding and translocation of a maltose binding protein through an alpha-hemolysin pore. (c) Schematic where a protein is immobilized at an AFM tip and translocated through a nanopore. (d) Schematic of native and SDS-unfolded protein translocation through a solid-state nanopore including typical current traces of native and SDS-unfolded proteins. Panel b from ref [78]; panel c from ref [90],and d from ref [77].

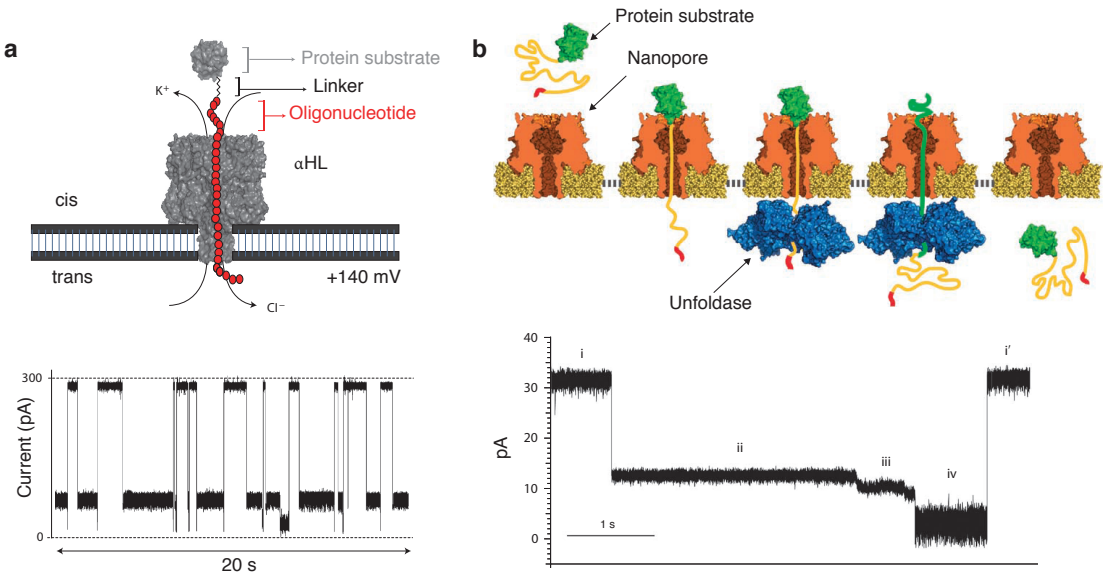

Figure 5. Translocation of unfolded proteins through nanopores using an oligonucleotide linker (left) or an unfoldase (right). (a) Schematic in which a DNA strand is used as lead for protein unfolding and translocation (top). Current traces observed for the translocation of DNA-tagged proteins (bottom) (b) Experimental set-up in which an unfoldase is used to unfold and pull the protein substrate (top). Typical current trace observed during a translocation and unfolding event (bottom). Panel a adapted from ref [91]; panels b from ref [97].

References

1. Miyashita, M. *et al.* Attomole level protein sequencing by Edman degradation coupled with accelerator mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4403–8 (2001).
2. SHIMONISHI, Y. *et al.* Sequencing of Peptide Mixtures by Edman Degradation and Field-Desorption Mass Spectrometry. *Eur. J. Biochem.* **112,** 251–264 (1980).
3. Bradley, C. V, Williams, D. H. & Hanley, M. R. Peptide sequencing using the combination of edman degradation, carboxypeptidase digestion and fast atom bombardment mass spectrometry. *Biochem. Biophys. Res. Commun.* **104,** 1223–30 (1982).
4. KA., W. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata. (Accessed: 2nd July 2018)
5. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).
6. Haider, S. & Pal, R. Integrated Analysis of Transcriptomic and Proteomic Data. *Curr. Genomics* **14,** 91–110 (2013).
7. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405,** 837–846 (2000).
8. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422,** 193–197 (2003).
9. Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5,** 699–711 (2004).
10. Yates III, J. R. A century of mass spectrometry: from atoms to proteomes. *Nat. Methods* **8,** 633–637 (2011).
11. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355 (2016).
12. Walther, T. C. & Mann, M. Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.* **190,** 491–500 (2010).
13. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **28,** 710–721 (2010).
14. A cast of thousands. *Nat. Biotechnol.* **21,** 213 (2003).
15. Anderson, N. L. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. *Mol. Cell. Proteomics* **1,** 845–867 (2002).
16. Zubarev, R. A. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13,** 723–726 (2013).
17. Huang, B. *et al.* Counting Low-Copy Number Proteins in a Single Cell. *Science (80-.).* **315,** 81–84 (2007).
18. Ham, B. M. & MaHam, A. *Analytical Chemistry: A Chemist and Laboratory Technian's Toolkit*. (John Wiley & Sons, 2015).
19. Hawkridge, A. M. Chapter 1 Practical Considerations and Current Limitations in Quantitative Mass Spectrometry-based Proteomics. *Quant. Proteomics* 1–25 (2014). doi:10.1039/9781782626985-00001
20. Pagel, O., Loroch, S., Sickmann, A. & Zahedi, R. P. Current strategies and findings in clinically relevant post-translational modification-specific

proteomics. *Expert Rev. Proteomics* **12,** 235–253 (2015).
21. Heath, J. R., Ribas, A. & Mischel, P. S. Single-cell analysis tools for drug discovery and development. *Nat. Rev. Drug Discov.* **15,** 204–216 (2015).
22. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18,** 345–361 (2017).
23. Su, Y., Shi, Q. & Wei, W. Single cell proteomics in biomedicine: High-dimensional data acquisition, visualization, and analysis. *Proteomics* **17,** 1600267 (2017).
24. Lu, Y., Yang, L., Wei, W. & Shi, Q. Microchip-based single-cell functional proteomics for biomedical applications. *Lab Chip* **17,** 1250–1263 (2017).
25. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165,** 780–791 (2016).
26. Edman, P. Method for determination of the amino acid sequence in peptides. *Acta Chemica Scandinavica* **4,** 283–293 (1950).
27. Li, K. W. & Geraerts, W. P. M. in *Neuropeptide Protocols* 17–26 (Humana Press, 1997). doi:10.1385/0-89603-399-6:17
28. McCormack, A. L. *et al.* Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69,** 767–776 (1997).
29. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. - Supplement. *Nature* **456,** 53–9 (2008).
30. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (80-.).* **323,** 133–138 (2009).
31. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 3960–4 (2003).
32. Hernandez, E. T., Swaminathan, J., Marcotte, E. M. & Anslyn, E. V. Solution-phase and solid-phase sequential, selective modification of side chains in KDYWEC and KDYWE as models for usage in single-molecule protein sequencing. *New J. Chem.* **41,** 462–469 (2017).
33. Yao, Y., Docter, M., van Ginkel, J., de Ridder, D. & Joo, C. Single-molecule protein sequencing through fingerprinting: computational assessment. *Phys. Biol.* **12,** 055003 (2015).
34. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A Theoretical Justification for Single Molecule Peptide Sequencing. *PLOS Comput. Biol.* **11,** e1004080 (2015).
35. Müller, V. & Westerlund, F. Optical DNA mapping in nanofluidic devices: principles and applications. *Lab Chip* **17,** 579–590 (2017).
36. van Ginkel, J. *et al.* Single-molecule peptide fingerprinting. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 3338–3343 (2018).
37. Preminger, M. & Smilansky, Z. Methods for evaluating ribonucleotide sequences. (2009).
38. Stevens, B. *et al.* Fret-based identification of mRNAs undergoing translation. *PLoS One* **7,** e38344 (2012).
39. Swaminathan, J. Single molecule peptide sequencing. (2015). doi:10.15781/T2D21RQ49
40. Borgo, B. & Havranek, J. J. Computer-aided design of a catalyst for Edman degradation utilizing substrate-assisted catalysis. *Protein Sci.* **24,** 571–579 (2015).

41. Aviram, A. & Ratner, M. A. Molecular rectifiers. *Chem. Phys. Lett.* **29,** 277–283 (1974).
42. Dekker, C., Tans, S. J., Oberndorff, B., Meyer, R. & Venema, L. C. STM imaging and spectroscopy of single copperphthalocyanine molecules. *Synth. Met.* **84,** 853–854 (1997).
43. Reed, M. A. Conductance of a Molecular Junction. *Science (80-.).* **278,** 252–254 (1997).
44. Ratner, M. A brief history of molecular electronics. *Nat. Nanotechnol.* **8,** 378–381 (2013).
45. Tsutsui, M., Taniguchi, M., Yokota, K. & Kawai, T. Identifying single nucleotides by tunnelling current. *Nat. Nanotechnol.* **5,** 286–290 (2010).
46. Tanaka, H. & Kawai, T. Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. *Nat. Nanotechnol.* **4,** 518–522 (2009).
47. Shapir, E. *et al.* Electronic structure of single DNA molecules resolved by transverse scanning tunnelling spectroscopy. *Nat. Mater.* **7,** 68–74 (2008).
48. Chang, S. *et al.* Electronic Signatures of all Four DNA Nucleosides in a Tunneling Gap. *Nano Lett.* **10,** 1070–1075 (2010).
49. Feng, J. *et al.* Identification of single nucleotides in MoS2 nanopores. *Nat. Nanotechnol.* **11,** 117–126 (2015).
50. Lindsay, S. *et al.* Recognition tunneling. *Nanotechnology* **21,** 262001 (2010).
51. Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* **9,** 466–73 (2014).
52. Ohshiro, T. *et al.* Detection of post-translational modifications in single peptides using electron tunnelling currents. *Nat. Nanotechnol.* **9,** 835–840 (2014).
53. Morikawa, T., Yokota, K., Tsutsui, M. & Taniguchi, M. Fast and low-noise tunnelling current measurements for single-molecule detection in an electrolyte solution using insulator-protected nanoelectrodes. *Nanoscale* **9,** 4076–4081 (2017).
54. Morikawa, T., Yokota, K., Tanimoto, S., Tsutsui, M. & Taniguchi, M. Detecting Single-Nucleotides by Tunneling Current Measurements at Sub-MHz Temporal Resolution. *Sensors* **17,** 885 (2017).
55. Taniguchi, M., Tsutsui, M., Yokota, K. & Kawai, T. Fabrication of the gating nanopore device. *Appl. Phys. Lett.* **95,** 123701 (2009).
56. Yokota, K., Tsutsui, M. & Taniguchi, M. Electrode-embedded nanopores for label-free single-molecule sequencing by electric currents. *RSC Adv.* **4,** 15886–15899 (2014).
57. Heerema, S. J. & Dekker, C. Graphene nanodevices for DNA sequencing. *Nat. Nanotechnol.* **11,** 127–136 (2016).
58. Ivanov, A. P. *et al.* DNA tunneling detector embedded in a nanopore. *Nano Lett.* **11,** 279–285 (2011).
59. Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics and Bioinformatics* **14,** 265–279 (2016).
60. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12,** 351–356 (2015).
61. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17,** 239 (2016).
62. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing.

Nat. Biotechnol. **34,** 518–524 (2016).
63. Gaskill, M. First DNA Sequencing in Space a Game Changer. (2016). Available at: https://www.nasa.gov/mission_pages/station/research/news/dna_sequencing. (Accessed: 26th April 2017)
64. Waduge, P. *et al.* Nanopore-Based Measurements of Protein Size, Fluctuations, and Conformational Changes. *ACS Nano* **11,** 5706–5716 (2017).
65. Bell, N. A. W. & Keyser, U. F. Digitally encoded DNA nanostructures for multiplexed, single-molecule protein sensing with nanopores. *Nat. Nanotechnol.* **11,** 645–651 (2016).
66. Plesa, C., Ruitenberg, J. W., Witteveen, M. J. & Dekker, C. Detection of individual proteins bound along DNA using solid-state nanopores. *Nano Lett.* **15,** 3153–3158 (2015).
67. Venkatesan, B. M. & Bashir, R. Nanopore sensors for nucleic acid analysis. *Nat. Nanotechnol.* **6,** 615–624 (2011).
68. Stefureac, R., Long, Y.-T., Kraatz, H.-B., Howard, P. & Lee, J. S. Transport of α-Helical Peptides through α-Hemolysin and Aerolysin Pores †. *Biochemistry* **45,** 9172–9179 (2006).
69. Sutherland, T. C. *et al.* Structure of Peptides Investigated by Nanopore Analysis. *Nano Lett.* **4,** 1273–1277 (2004).
70. Movileanu, L., Schmittschmitt, J. P., Martin Scholtz, J. & Bayley, H. Interactions of Peptides with a Protein Pore. *Biophys. J.* **89,** 1030–1045 (2005).
71. Goodrich, C. P. *et al.* Single-Molecule Electrophoresis of β-Hairpin Peptides by Electrical Recordings and Langevin Dynamics Simulations. *J. Phys. Chem. B* **111,** 3332–3335 (2007).
72. Mohammad, M. M. & Movileanu, L. Excursion of a single polypeptide into a protein pore: simple physics, but complicated biology. *Eur. Biophys. J.* **37,** 913–925 (2008).
73. Mahendran, K. R., Romero-Ruiz, M., Schlösinger, A., Winterhalter, M. & Nussberger, S. Protein Translocation through Tom40: Kinetics of Peptide Release. *Biophys. J.* **102,** 39–47 (2012).
74. Ji, Z. *et al.* Fingerprinting of Peptides with a Large Channel of Bacteriophage Phi29 DNA Packaging Motor. *Small* **12,** 4572–4578 (2016).
75. Talaga, D. S. & Li, J. Single-Molecule Protein Unfolding in Solid State Nanopores. *J. Am. Chem. Soc.* **131,** 9287–9297 (2009).
76. Li, J., Fologea, D., Rollings, R. & Ledden, B. Characterization of Protein Unfolding with Solid-state Nanopores. *Protein Pept. Lett.* **21,** 256–265 (2014).
77. Restrepo-Pérez, L., John, S., Aksimentiev, A., Joo, C. & Dekker, C. SDS-assisted protein transport through solid-state nanopores. *Nanoscale* **9,** 11685–11693 (2017).
78. Oukhaled, G. *et al.* Unfolding of Proteins and Long Transient Conformations Detected by Single Nanopore Recording. *Phys. Rev. Lett.* **98,** 158101 (2007).
79. Pastoriza-Gallego, M. *et al.* Dynamics of Unfolded Protein Transport through an Aerolysin Pore. *J. Am. Chem. Soc.* **133,** 2923–2931 (2011).
80. Merstorf, C. *et al.* Wild Type, Mutant Protein Unfolding and Phase Transition Detected by Single-Nanopore Recording. *ACS Chem. Biol.* **7,** 652–658 (2012).
81. Pastoriza-Gallego, M. *et al.* Urea denaturation of α-hemolysin pore inserted in planar lipid bilayer detected by single nanopore recording: Loss of structural asymmetry. *FEBS Lett.* **581,** 3371–3376 (2007).
82. Freedman, K. J. *et al.* Chemical, Thermal, and Electric Field Induced

Unfolding of Single Protein Molecules Studied Using Nanopores. *Anal. Chem.* **83,** 5137–5144 (2011).
83. Payet, L. *et al.* Thermal unfolding of proteins probed at the single molecule level using nanopores. *Anal. Chem.* **84,** 4071–4076 (2012).
84. Cressiot, B. *et al.* Protein Transport through a Narrow Solid-State Nanopore at High Voltage: Experiments and Theory. *ACS Nano* **6,** 6236–6243 (2012).
85. Oukhaled, A. *et al.* Dynamics of Completely Unfolded and Native Proteins through Solid-State Nanopores as a Function of Electric Driving Force. *ACS Nano* **5,** 3628–3638 (2011).
86. Freedman, K. J., Haq, S. R., Edel, J. B., Jemth, P. & Kim, M. J. Single molecule unfolding and stretching of protein domains inside a solid-state nanopore by electric field. *Sci. Rep.* **3,** 1638 (2013).
87. Firnkes, M., Pedone, D., Knezevic, J., Döblinger, M. & Rant, U. Electrically facilitated translocations of proteins through silicon nitride nanopores: Conjoint and competitive action of diffusion, electrophoresis, and electroosmosis. *Nano Lett.* **10,** 2162–2167 (2010).
88. Huang, G., Willems, K., Soskine, M., Wloka, C. & Maglia, G. Electro-osmotic capture and ionic discrimination of peptide and protein biomarkers with FraC nanopores. *Nat. Commun.* **8,** 935 (2017).
89. Kennedy, E., Dong, Z., Tennant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm3 resolution using a subnanometre-diameter pore. *Nat. Nanotechnol.* **11,** 968–976 (2016).
90. Dong, Z., Kennedy, E., Hokmabadi, M. & Timp, G. Discriminating Residue Substitutions in a Single Protein Molecule Using a Sub-nanopore. *ACS Nano* **11,** 5440–5452 (2017).
91. Rodriguez-Larrea, D. & Bayley, H. Multistep protein unfolding during nanopore translocation. *Nat. Nanotechnol.* **8,** 288–295 (2013).
92. Rodriguez-Larrea, D. & Bayley, H. Protein co-translocational unfolding depends on the direction of pulling. *Nat. Commun.* **5,** 4841 (2014).
93. Rosen, C. B., Rodriguez-Larrea, D. & Bayley, H. Single-molecule site-specific detection of protein phosphorylation with a nanopore. *Nat. Biotechnol.* **32,** 179–181 (2014).
94. Biswas, S., Song, W., Borges, C., Lindsay, S. & Zhang, P. Click Addition of a DNA Thread to the N-Termini of Peptides for Their Translocation through Solid-State Nanopores. *ACS Nano* **9,** 9652–9664 (2015).
95. Pastoriza-Gallego, M. *et al.* Evidence of Unfolded Protein Translocation through a Protein Nanopore. *ACS Nano* **8,** 11350–11360 (2014).
96. Plesa, C. *et al.* Fast translocation of proteins through solid state nanopores. *Nano Lett.* **13,** 658–63 (2013).
97. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* **31,** 247–250 (2013).
98. Nivala, J., Mulroney, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among Protein Variants Using an Unfoldase-Coupled Nanopore. *ACS Nano* **8,** 12365–12375 (2014).
99. Aubin-Tam, M.-E., Olivares, A. O., Sauer, R. T., Baker, T. a & Lang, M. J. Single-molecule protein unfolding and translocation by an ATP-fueled proteolytic machine. *Cell* **145,** 257–67 (2011).
100. Sampath, G. Amino acid discrimination in a nanopore and the feasibility of sequencing peptides with a tandem cell and exopeptidase. *RSC Adv.* **5,** 30694–

30700 (2015).
101. Boynton, P. & Di Ventra, M. Sequencing proteins with transverse ionic transport in nanochannels. *Sci. Rep.* **6,** 25232 (2016).
102. Wilson, J., Sloman, L., He, Z. & Aksimentiev, A. Graphene Nanopores for Protein Sequencing. *Adv. Funct. Mater.* **26,** 4830–4838 (2016).
103. Maulbetsch, W., Wiener, B., Poole, W., Bush, J. & Stein, D. Preserving the Sequence of a Biopolymer's Monomers as They Enter an Electrospray Mass Spectrometer. *Phys. Rev. Appl.* **6,** 054006 (2016).
104. Keifer, D. Z. & Jarrold, M. F. Single-molecule mass spectrometry. *Mass Spectrometry Reviews* **36,** 715–733 (2016).
105. Manrao, E. A. *et al.* Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* **30,** 349–353 (2012).
106. Millioni, R. *et al.* High abundance proteins depletion vs low abundance proteins enrichment: Comparison of methods to reduce the plasma proteome complexity. *PLoS One* **6,** e19603 (2011).
107. Baker, M. S. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **8,** 14271 (2017).

Acknowledgements

We thank S. Pud, S. Schmid, S. Caneva, J. van Ginkel and M. Filius for useful discussions. We acknowledge funding received from the Netherlands Organisation for Scientific Research (NWO/OCW) as a part of the Frontiers of Nanoscience program. The C.D. lab was further supported by the ERC Advanced Grant SynDiv (No. 669598) and by the National Human Genome Research Institute of the National Institute of Health under Award Number R01-HG007406. C.J. was funded by the Foundation for Fundamental Research on Matter (12PR3029 and SMPS).