The emerging landscape of single-molecule protein sequencing technologies

Bohländer, Peggy; Filius, Mike; van Kooten, Xander F.; Pomorski, Adam; Schmid, Sonja; Dekker, Cees; Eelkema, Rienk; Kim, Sung Hyun; Joo, Chirlmin; More Authors

DOI
10.1038/s41592-021-01143-1

Publication date
2021

Document Version
Accepted author manuscript

Published in
Nature Methods

Citation (APA)
Bohländer, P., Filius, M., van Kooten, X. F., Pomorski, A., Schmid, S., Dekker, C., Eelkema, R., Kim, S. H., Joo, C., & More Authors (2021). The emerging landscape of single-molecule protein sequencing technologies. *Nature Methods*, *18*(6), 604-617. https://doi.org/10.1038/s41592-021-01143-1

The emerging landscape of single-molecule protein sequencing technologies

Javier Alfaro[1,‡,†,*], Peggy Bohländer[2,†], Mingjie Dai[3,4,†], Mike Filius[5,†], Cecil J. Howard[6,†], Xander F. van Kooten[7,†], Shilo Ohayon[7,†], Adam Pomorski[5,†], Sonja Schmid[8,†], Aleksei Aksimentiev[9], Eric V. Anslyn[6], Georges Bedran[1], Cao Chan[10], Mauro Chinappi[11], Etienne Coyaud[12], Cees Dekker[5], Gunnar Dittmar[13], Nicholas Drachman[14], Rienk Eelkema[2], David Goodlett[15], Sebastien Hentz[16], Umesh Kalathiya[1], Neil L. Kelleher[17], Ryan T. Kelly[18], Zvi Kelman[19,20], Sung Hyun Kim[5], Bernhard Kuster[21,22], David Rodriguez-Larrea[23], Stuart Lindsey[24], Giovanni Maglia[25], Edward M. Marcotte[26], John P. Marino[19], Christophe Masselon[27], Michael Mayer[28], Patroklos Samaras[21], Kumar Sarthak[9], Lusia Sepiashvili[29], Derek Stein[14], Meni Wanunu[30,31], Mathias Wilhelm[21], Peng Yin[3,4], Amit Meller[7,32‡,*] and Chirlmin Joo[5,‡,*].

[1] International Centre for Cancer Vaccine Science, University of Gdańsk, Gdańsk, Poland
[2] Faculty of Applied Sciences, Delft University of Technology, Van der Maasweg 9, 2629 HZ, Delft, The Netherlands
[3] Wyss Institute for Biologically Inspired Engineering, Harvard University, 3 Blackfan circle, Boston, MA, 02115 USA
[4] Department of Systems Biology, Harvard Medical School, Boston, MA, 02115 USA
[5] Department of BioNanoScience, Kavli Institute of Nanoscience, Delft University of Technology, van der Maasweg 9, 2629 HZ Delft, The Netherlands
[6] Department of Chemistry, The University of Texas at Austin, Austin, Texas 78712, United States
[7] Department of Biomedical Engineering, Technion – Israel Institute of Technology, Haifa, 32000, Israel
[8] NanoDynamicsLab, Biophysics Chair, Wageningen University, Stippeneng 4, 6708WE Wageningen, The Netherlands.
[9] Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801
[10] Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[11] Dipartimento di Ingegneria Industriale, Università di Roma Tor Vergata, via del Politecnico 1, 00133 Roma, Italy
[12] Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France
[13] Proteomics of cellular signalling, Quantitative Biology Unit, Luxembourg Institute of Health, Strassen, Luxembourg.
[14] Department of Physics, Brown University, Providence, RI 02912, USA.
[15] Department of Microbial Pathogenesis, School of Dentistry, University of Maryland, Baltimore, Maryland 21201, United States
[16] Université Grenoble Alpes, CEA, LETI, 38000 Grenoble, France.
[17] Departments of Chemistry, Molecular Biosciences, and the Feinberg School of Medicine, Northwestern University, 2145 Sheridan Road, Box #101, Evanston, IL 60208-3113, United States
[18] Department of Chemistry and Biochemistry, Brigham Young University, Provo, UT 84602, United States
[19] Institute for Bioscience and Biotechnology Research, National Institute of Standards and Technology, University of Maryland, 9600, Gudelsky Drive, Rockville, MD, USA.
[20] Biomolecular Labeling Laboratory, IBBR, Rockville, MD, USA
[21] Chair of Proteomics and Bioanalytics, Technische Universität München, Emil-Erlenmeyer Forum 5, 85354 Freising, Germany
[22] Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany
[23] Biofisika Institute (CSIC, UPV/EHU), Department of Biochemistry and Molecular Biology (UPV/EHU), Leioa 48940, Spain
[24] Biodesign Institute, School of Molecular Sciences, Department of Physics, Arizona State University, Tempe, Arizona 85287, United States
[25] Groningen Biomolecular Sciences & Biotechnology Institute, University of Groningen, 9747 AG Groningen, The Netherlands
[26] Department of Molecular Biosciences, Center for Systems and Synthetic Biology, University of Texas, Austin, TX 78712, USA
[27] Université Grenoble Alpes, CEA, Inserm, BGE U1038, Grenoble, France
[28] University of Fribourg, Adolphe Merkle Institute, Chemin des Verdiers 4, CH-1700 Fribourg, Switzerland
[29] University of Toronto, Hospital for Sick Children, 555 University Ave., Rm 3606 Atrium, Toronto Ontario M5G 1X8
[30] Department of Physics, Northeastern University, Boston, Massachusetts 02115, United States
[31] Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States
[32] Russell Berrie Nanotechnology Institute, Technion − IIT, Haifa 32000, Israel
[†] These authors contributed equally
[‡] These authors jointly directed this work.
[*] Corresponding Authors

* Correspondence should be addressed to: J.A. Alfaro- Javier Alfaro javier.alfaro@proteogenomics.ca, A. Meller - Amit Meller ameller@technion.ac.il, and C. Joo - Chirlmin Joo c.joo@tudelft.nl

Abstract

Proteins are involved in the majority of structures and biochemical reactions of living cells. New single-molecule protein sequencing and identification technologies alongside innovations in mass spectrometry and antibody-based methods will eventually enable broad sequence coverage in single-cell profiling. The ultimate precision and sensitivity of proteomes promised by these technologies will create new directions in research and biomedical applications, from global proteomics of single cells and bodily fluids to sensing and classifying low-abundance protein biomarkers for disease screening and precision diagnostics.

Introduction

The emergence of Next Generation Sequencing (NGS) and single-molecule DNA sequencing technologies have revolutionized genomics. Proteomics awaits a similar transformative wave of protein-sequencing techniques that will allow for the examination of proteins at the single-cell and ultimately single-molecule levels, even with low-abundant proteins. Such techniques would allow routine global proteome profiling, like today's single-cell RNA sequencing studies, creating opportunities for single-cell proteomics and potentially permitting real-time testing for on-site medical diagnostics and disease screening. Importantly, whole proteome sequencing and profiling of the vast repertoire of cell types is expected to fundamentally enhance our understanding of all living systems.

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

MS remains a staple of protein identification and continues to develop towards single cell methods (**Box 2**). Alongside, a diverse range of protein sequencing and identification techniques have emerged that aim to increase the sensitivity of proteomics to the single-molecule level. Many of these techniques rely on fluorescence and nanopores for single-molecule sensing as an alternative means to sequence or identify proteins (**Figure 1).** The landscape of emerging proteomics is already vast, with different approaches at various stages of development, some of which have already secured industry investment[3,4], an important step towards broad dissemination to the research community. Other technologies have shown great promise and gained popularity among the single-molecule biophysics communities and some are available as proofs of concept at just one or a few laboratories.

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

A Renaissance of Classical Techniques

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

Massively parallel Edman degradation

Edman degradation[7] was the first method to determine the amino acid sequence of a purified peptide. The method relied on chemically modifying the N-terminal amino acid, cleaving it from the peptide, and finally determining the sequence of the cleaved labeled amino acid using high-performance liquid chromatography. Until recently, attempts to conduct sequencing of this sort in a massively parallel fashion were not possible as the method relied on highly purified peptides. However, recent multiplex strategies that employ peptide arrays and either sequence chemically labeled peptides (“fluorosequencing”), or successively detect the N-terminal amino acid are making breakthroughs (**Box 3**).

Fluorosequencing combines Edman chemistry, single-molecule microscopy, and stable synthetic fluorophore chemistry (**Figure 2a)**. Millions of individual fluorescently labeled peptides can be visualized in parallel, while changing fluorescence intensities are monitored as N-terminal amino acids are sequentially removed. The resulting fluorescence signatures serve to uniquely identify individual peptides[8]. This method allows for millions of distinct peptide molecules to be sequenced in parallel, identified, and digitally quantified on a zeptomole-scale[9]. However, the technology is not without challenges, as the reagents used for Edman degradation chemistry lead to increased rates of fluorescent dye destruction, which in turn limits the read length. These reagents include slightly basic structures such as pyridine, strong acids such as trifluoroacetic acid, and the electrophile phenyl isothiocyanate. Furthermore, the reliance on chemical labelling leads to partial sequencing of the peptide, with

the unidentified remainder inferred by comparison to a reference proteome. In addition, the inefficient labeling can lead to errors that must be modeled into the reference proteome comparison, spurring the development of new protocols to increase yields[10]. Exciting new proposals could add the dimension of protonation-based sequencing. The pK_a of the N-terminal amino acid could be used for identification by observing and interpreting the protonation-deprotonation signal of the peptide at fixed pH through the Edman Degredation process[11]. Much like fluorosequencing, the signal observed would be for the whole peptide and the decay pattern would be interpreted to derive a pK_a of each N-Terminal aminoacid.

Several natural proteins and RNA molecules recognize specific amino acids either as free amino acids or as a part of a polypeptide chain[12]. These proteins and nucleic acids provide different solutions for N-terminal amino acid recognition. Each N-terminal amino acid binder (NAAB) probe selectively identifies a specific N-terminal amino acid or an N-terminal amino acid derivative. With each cycle, another amino acid is revealed in the sequence of the peptide. However, further directed evolution and engineering of the NAAB probe is required to meet the stringent affinity, selectivity and stability needed for error-free sequencing applications. In addition, such probes would need to discriminate among all amino acids, including the same amino acid in alternative positions in the peptide sequence. Probes that bind a class of N-terminal amino acids (e.g., short aliphatic residues) could also be useful, but would introduce ambiguities in the sequencing process. Different probes could also be designed to recognize short N-terminal k-mers, which would increase the number of probes needed, but reduce the ambiguity in the resulting sequencing information. To circumvent this limitation, it may be possible to sequence the N-terminal amino acid by selective recognition using a plurality of probes in each cycle of Edman degradation[13,14] (**Figure 2b)**.

Single-molecule mass spectrometry

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

widely used to directly derive the charge states of single proteins and even their fragment ions[20]. Orbitraps are particularly useful since the readout of individual ions can be multiplexed by 100-1000 fold in the Orbitrap-based CDMS[20]. Individual ion MS has already shown resolution of mixtures with ~1000 proteoforms that provided no data when standard MS was used[20,21]. This has greatly expanded the top-down approach to confirm DNA-inferred sequences of whole proteins, with localization of their post-translational modifications[20–22]. Without extensive alteration, Orbitraps can therefore measure tens of thousands of proteins in a matter of minutes. With these rapidly evolving technologies, the charting of the full human proteoform atlas has already begun[23], making strides towards a comprehensive Human Proteoform Project. However, a critical requirement for MS of proteins and peptides is ionization, and not all ions are efficiently transmitted through the mass spectrometer. This might restrict some of the proteoform mapping efforts providing a niche for the other technologies in **Figure 1**.

For higher molecular weight species, the ionization of proteins and complexes yields a mixture of macro ions with variable charge states, resulting in a net reduction of sensitivity, as the signal distributes over multiple peaks in the mass-to-charge dimension. Moreover, charge state distributions may overlap above a certain mass or in the case of mixtures, challenging the species identification. Since their inception[24], nano-mechanical mass sensors have made tremendous progress towards protein characterization[25]. Such devices, which take the shape of cantilevers or beams with lateral dimensions in the hundreds of nanometers, can detect individual particles accreting onto their active surface through the changes in their vibration frequency. Importantly, as the particle's inertial mass is determined directly from the frequency change, these devices are insensitive to charge states[26]. This realization prompted the development of new MS instrument designs devoid of ion guides, which no longer depend on electromagnetic fields to collect and transmit the analytes (**Figure 2c**). Such nano-mechanical resonator-based MS system has recently shown the ability to characterize large protein assemblies such as individual viral capsids above 100 MDa[27]. Outside of proteomics, 1 Da resolution has been demonstrated with carbon nanotubes[28]. Moreover, recent reports suggested the possibility to determine other physical parameters like the stiffness or shape of the analyte by monitoring multiple vibrational modes[29,30]. These previously inaccessible metrics may open new avenues to discriminate peptides, proteins and their complexes. Nonetheless, one of the challenges of the nano-resonator-MS lies in devising efficient ways to bring individual proteins onto the resonator's active surface for mass sensing.

Ionization is commonly achieved by electrospray ionization of a solution containing the compound(s) of interest. The use of ever-smaller electrospray ion source apertures has led to significant improvements in the sensitivity of mass spectrometry[31,32]. Mass spectrometers with a nanopore ion source have been developed for the purpose of sequencing single proteins[33] (**Figure 2d**). A nanopore electrospray can potentially deliver individual amino acid ions directly into a high-vacuum gas phase, where the ions can be efficiently detected by their mass-to-charge ratios. This opens a path to sequencing peptides one amino acid at a time. The concept makes use of the nanopore to guide the protein into a linear configuration so that its monomers can be delivered into the mass spectrometer sequentially[34]. Individual amino acids must be cleaved from the protein molecule as they transit the nanopore, and this could potentially be accomplished using photodissociation[35] or chemical digestion methods. The 100 MHz bandwidth of the channeltron single ion detectors used in this setup is also sufficient to resolve the arrival order of the ions. The high mass resolution makes this technique promising for identifying post-translational modifications (PTMs), which change the masses of particular amino acid residues by predictable amounts. One challenge on the path for this technology will be achieving high throughput, which might require a strategy for parallelizing the mass analysis.

Tunneling conductance measurements

The appearance of the scanning tunneling microscope in the 1980s opened a new way to analyze molecules. Small organic molecules can be transiently trapped between two metal electrodes with sub-nanometer separation, with the tunneling currents between the electrodes reporting on the molecular signature of the analyte. Recently, several technical advances have been made towards single-molecule amino acid and protein analysis. Extracting insightful information from electron tunneling is complicated by the noise due to water and contaminants reaching the electrode surfaces. To overcome these problems, recognition tunneling has been developed. The electrodes are covalently modified with adaptor molecules that form transient, but well-defined links to the target molecule[36]. The rapidly fluctuating tunnel-current signals are processed using machine learning algorithms, which makes it possible to distinguish individual amino acids and small peptides[37]. Moreover, smaller electrode gaps have been made to obtain distinct signals from different amino acids and PTMs[38]. Further development of the technology will depend on a reliable source of tunnel junctions with a defined gap to replace the cumbersome scanning tunneling microscopy, but it is clear that both the sequence and PTMs of small peptides can be determined[37]. Currently, tunneling conductance is a proof-of-concept technology for fully sequencing short peptides that could one day be used for the analysis of protein digests and expanded to PTM analysis (**Figure 1**).

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

DNA Nanotechnologies for Protein Sequencing

DNA nanotechnologies, which utilize the ability to custom-design a large number of sequences with prescribed pairing interaction and dynamic properties, have facilitated developments in fields ranging from synthetic biology to diagnostics and drug delivery[41]. For example, the programmable, transient binding between short DNA strands is central to the super-resolution technique, DNA-based Point Accumulation for Imaging in Nanoscale Topography (DNA-PAINT[42–44]) (**Box 4**). Here we describe the application of DNA-PAINT and DNA-based local and global pairwise distance measurement methods for single-molecule protein detection and identification.

Fingerprinting via DNA PAINT

DNA-PAINT uses the repetitive binding between designed docking and imager DNA strands to allow for imaging with molecular-level resolution (**Box 4**). This method provides a promising way to fingerprint proteins on the level of single molecules. A simple way for characterizing proteins could be through amino acid counting using quantitative DNA-PAINT (qPAINT[44]). In this technique, the total blinking rate of a region of interest is measured, which linearly reflects the number of molecular targets in the region. It has been proposed that high-efficiency DNA labeling of specific amino acids (**Figure 3a**) followed by qPAINT could lead to single-molecule protein fingerprinting of intact proteins (**Figure 3b**)[45]. More than 75% of the human proteome can be identified (≤5 degeneracy) if the error in counting is less than 5% from detecting three kinds of specific amino acids.

The recent development of DNA-PAINT has allowed discrete molecular imaging (DMI) of individual molecular targets with <5 nm spatial resolution[43]. Therefore, protein identification by

the fingerprinting of amino acids along an extended protein backbone is a possibility. DMI was achieved by combining a systematic analysis and optimization of DNA-PAINT super-resolution workflow and a high-accuracy (<1 nm) drift correction method. To effectively unfold and extend the protein backbone, N- and C-terminal specific modifications should be used to attach surface and microbead anchors. The protein can then be subjected to mechanical or electromagnetic extension force (**Figure 3c**). Proposals to combine protein extension methods with high-resolution DMI[45] indicate that with lysine labelling alone and 5-nm effective imaging resolution, more than 50% of the human proteome could be uniquely identified, even with up to 20% amino acid imaging error. Labeling lysine and cysteine would allow the proteome coverage to increase to more than 75%.

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

DNA proximity recording

An alternative method for DNA-based protein identification attaches DNA probes to specific amino-acids on a protein and uses enzymatic DNA-amplification between nearby probes to generate DNA 'records' that vary in length according to pairwise distances within a protein. An example is auto-cycling proximity recording[46] (APR) (**Figure 3d**). The distribution of lengths of these molecular records is then analysed to decode the pairwise distance between the two DNA tags. It is possible to use unique molecular identifier barcoding and repetitive enzymatic recording, such that each lysine and cysteine residue can be studied and a pairwise distance map can be constructed among them, allowing for single-molecule protein identification[47,48]. DNA proximity recording takes advantage of high-throughput next generation DNA sequencing methods for efficient protein fingerprinting analysis, and will be useful for both the analysis of purified proteins and complex protein mixtures.

Protein fingerprinting using FRET

A different approach that allows for global pairwise distance measurement is combining DNA technology with single-molecule Förster resonance energy transfer (FRET)[49]. The current state of the art of single-molecule FRET analysis allows us to deal with only one or two FRET

pairs[50]. The new high-resolution FRET using transient binding between DNA tags allows for probing one FRET pair at a time when many of them are collectively present on a single protein[49]. Similar to the above-mentioned approaches, specific amino acids (e.g. lysines, cysteines, etc) required for fingerprinting have to be labelled with a set of different DNA docking strands. Furthermore, a fixed position on the protein (either the N or the C terminus) is labelled with the acceptor fluorophore. Only a single FRET pair forms at a time by using DNA strands that are complementary to only a single docking strand. The measurements are then repeated to probe the remaining docking strands and thus the amino acids. The output of this approach will be a FRET histogram containing information on the position (referred to as FRET fingerprints) of each detected amino acid relative to one of the reference points. This information is compared to a database consisting of predicted 'FRET fingerprints' and allows for the identification of the protein species (**Figure 3e**). The proposed high-resolution FRET approach (named high resolution FRET using DNA eXchange, or FRET X) benefits from the immobilization of the protein molecules, allowing users to probe each protein multiple times to obtain fingerprints with a high resolution. FRET X will be a particularly promising tool for targeted proteomics or proteoform analysis as it is able to distinguish small structural changes.

Biological and Solid-State Nanopores

Nanopore-based DNA and direct RNA sequencing technologies have become key players in the sequencing field, offering unprecedented read-lengths and portability. Since its first demonstration as a single-biomolecule sensor[51], nanopore sensing has progressively matured reaching the goal of single-molecule, long-read DNA sequencing[52]. Many of the nanopore sequencing applications to date have materialized using an ultra-small device[53], which features vast arrays of biological nanopores, each coupled to its own current amplifier, allowing readout of hundreds of DNA strands simultaneously. Nanopore sequencing involves drawing biomolecules through the nanopore in a single file manner, hence partially blocking the ionic current flowing through the pore, leading to time-dependent and sequence-specific electrical signals. In the past two decades a variety of synthetic nanopore biosensors have significantly progressed and are currently used in diverse applications beyond sequencing, including applications in detecting epigenetic variations and enabling ultra-sensitive mRNA expression[54], to name a few.

Just like gel electrophoresis, nanopores may serve as a generic tool to analyze biomolecules. Therefore, as nanopore-based DNA sequencing continues to advance, this technique is poised to extend to proteins, metabolites and to other analytes. But despite the remarkable advances in DNA and RNA sequencing, nanopore-based protein sensing is still in its infancy,

facing challenges unique to proteins and proteomics. In particular, proteins span a large range of sizes and have a stable three-dimensional folded structure. In contrast to nucleic acids, peptides' backbones are not naturally charged, complicating the possibility of single-file electrokinetic threading into nanopores. In addition, proteins are composed of 20 amino acids instead of 4 nucleobases, further complicating the task of relating the ionic current signals to the amino acid sequence.

While a significant progress in nanopore-based protein sensing have been made, to date the development of full protein sequencers, or single-protein identification based on nanopores remains to be a topic of intense focus. Here, we focus on three of the principal directions in this field (**Figure 4**): (i) Single-file threading and direct sensing of the sequence of the polypeptide's amino acids, analogous to the nanopore DNA sequencing principle. In this approach, either the translocation of full-length proteins or shorter polypeptide digests of the proteins may be targeted. (ii) Protein identification methods based on sensing unique fingerprints in linearized proteins, without *de novo* amino acid sequencing. (iii) Protein identification of folded proteins, based on specific patterns in their nanopore current blockades. In the following sections, we provide short overviews of the current state of these approaches and refer to additional methods.

Reading the amino acid sequence of linearized peptides

In this proposed approach, a single protein or peptide is linearized and threaded through a nanopore and the resulting ion current interpreted to an amino-acid sequence (**Figure 4a**). Theoretical work using all-atom MD simulations on alpha-hemolysin pores has demonstrated a global correlation between the volume of an amino acid and the current blockade in homo-polymers[55]. Computationally efficient predictions using course-grained models have also performed well compared to all-atom MD simulations for both solid-state and biological pores[56].

Discrimination among peptides differing by one amino acid substitution (alanine to glutamate) have been demonstrated using an engineered Fragaceatoxin C (FraC) nanopores[57]. Moreover, Piguet *et al.* resolved single amino acid differences within short poly-arginine peptides with superb resolution, using the aerolysin protein pore in its wild-type conformation[58]. Combining MD simulations and single channel experiments, Cao *et al.* have rationally determined specific point mutations in aerolysin to fine-tune the charge and diameter of the pore, which enhanced its sensitivity and selectivity as showcased experimentally using DNA and peptides[59]. Notably, protein pore sensors were used for analysis of bodily fluids

(blood, sweat, etc.), indicating significant potential for applications in diagnostics[60]. As an alternative to nanopore sequencing of intact polypeptide chains, smaller digested fragments can also be analyzed and minute differences in the amino acid composition can be detected[61]. Even post-translational modifications can be detected including individual phosphorylations and glycosylations using the protein pore FraC[62].

An essential step in the development of nanopore based DNA sequencing, came with the application of an enzymatic stepping motor (e.g. a helicase) that produces a nucleotide-by-nucleotide progression of the DNA through the nanopore. A similar system is pursued for single-molecule protein sequencing: Molecular motors of the Type II secretion system[63] (SecY) and the AAA family[64] (ClpX) are known to unfold and pull protein substrates through pores in an ATP-dependent way. Nivala et al.[65,66] employed ClpXP (or ClpX alone) to unfold and translocate a multi-domain fusion protein through the hemolysin pore using the energy derived from ATP hydrolysis. In this approach the motor is at the exit of the nanopore, and therefore the step size of translocation is caused by the stable structural motifs that resist translocation – rather than being controlled by the enzyme. This approach is currently being expanded by several groups, who conjugated ClpXP covalently to alpha-hemolysin at the entrance of the nanopore to form a combined sensor as well as a substrate delivery machine. The Maglia lab genetically introduced a nanopore directly into an archaeal proteasome and found that the assisted transport across the nanopore is not influenced by the unfolding of the protein. These nanoscale constructs would also allow a *cut-and-drop* approach, in which single proteins are recognized by the pattern of peptide fragments as they are sequentially cleaved by the peptidase above the nanopore[67]. Knyazev *et al.* introduced a protein-secreting ATPase as an additional natural choice for a potential peptide translocating motor[68,69]. Other proteins have the potential to control the protein translocation through nanopores, beyond secretases and unfoldases, including chaperones (Hsp70), via processes resembling protein translocation into the mitochondrial matrix[70]. Recently, Rodriguez-Larrea's group has discussed how protein refolding at the entry or exit compartment can oppose or promote protein translocation, respectively[71,72] and the use of deep learning networks to analyze the raw ionic current signals for accurate classification of single-point mutations in a translocating protein. In addition, Cardozo *et al.* built a library of ~20 proteins that are orthogonally barcoded with an intrinsic peptide sequence, and successfully read them by nanopore sensors[73].

Fingerprinting linearized proteins

An accurate quantification of different protein species in the proteome with single-molecule resolution would already be a highly significant achievement. This can be realized by single-

molecule fingerprinting, i.e. by the identification of individual protein molecules based on prior knowledge of their amino-acid sequence, or based on the specific signal patterns, recognized by machine learning[8,74,75] (**Figure 4b**). To that end several nanopore approaches have been pursued: Restrepo-Pérez *et al.*[76] established a fingerprinting approach using six chemical tags, which were placed on a dipolar peptide[77]. Additionally, Wang *et al.* reported the ability to distinguish individual lysine and cysteine residues in short polypeptides, coupled specifically to fluorescent tags using a solid-state nanopore with low fluorescence background[78]. In all these approaches, separating the proteins by mass, prior to single molecule sensing may highly facilitate the identification of proteins in complex samples containing many different proteins[79].

Nanopore protein fingerprinting can make extensive use of advanced deep-learning artificial intelligence (AI) strategies to identify patterns in noisy signals. Ohayon *et. al.* has recently shown computationally that >95% of all the proteins in the human proteome can be identified with high confidence, based on the labelling of three amino acids (lysine, cysteine and methionine) and linear threading through a solid state nanopore[75]. These simulations predict that even partial labelling of the proteins will be sufficient to achieve a high degree of accurate whole proteome identification, due to the ability of AI functions to correctly recognize partial protein patterns. This identification method involves the incorporation of subwavelength light localization in the proximity of the nanopore using plasmonic nanostructures[80]. The work in this field benefits from recent advances in nanofabrication and nanopatterning technologies, allowing for the formation of complex metallic nanostructures to induce light localization and plasmonics[81].

Characterization and identification of folded proteins

To date, nanopores have been successfully employed to detect specific sets of folded proteins and protein oligomers[82] (**Figure 4c**) such as large globular proteins, various cytokines and even low molecular weight proteins, such as Ubiquitin. Holding the proteins in their folded state inside the nanopore for sufficiently long times is a key requirement. Early studies have shown that globular proteins of about 5 nm in size can only be detected for a few tens of microseconds or less[83], which is too short for characterization. Several approaches to overcome this challenge have been devised. A lipid bilayer coating of a solid-state nanopore can be used to tether the proteins for extended periods of time[84]. Lipid tethered proteins[84], and more recently also freely diffusing proteins (using a higher bandwidth sensing system)[85] have been characterized based on their size, shape, charge, dipole, and rotational diffusion coefficient[86]. Various strategies are being pursued to ‘trap’ proteins in a nanopore. One

strategy is to use plasmonics to hold a protein in a nanopore for seconds or even minutes[87,88]. More recently single proteins have been demonstrated to be held at the nanopore's most sensitive region for minutes to hours using the nanopore electro-osmotic trap (NEOtrap) that exploits strong electro-osmotic water flows created in-situ by a charged, permeable objects, such as a DNA origami structures[89]. Another approach for slowing down the translocation of proteins involves the use of smaller nanopores compared to earlier studies, in order to increase the hydrodynamic drag, thus resulting in longer translocation dwell-times that are easier to measure[90,91]. In addition, high bandwidth measurements can resolve differential conformational flexibility within folded proteins[90–93], and even changes in conformational flexibility[94]. Biological nanopores with a diameter of 5.5 or 10 nm[95] can also be used to measure folded proteins, including protein conformations[96] and post-translational modifications[97] such as ubiquitination. Lastly, Aramesh *et al*.[98] used a combination of atomic force microscopy and nanopore technology, to make the first steps at nanopore sensing directly inside cells. Altogether, protein detection, identification, and even sequencing using single nanopore approaches has become a highly active, thriving research field, with great potential to revolutionize proteomics, medical diagnostics, and also fundamental biosciences.

Chemistry for Next-Generation Proteomics Technologies

Single-molecule protein fingerprinting has underlined the need for innovative approaches for attaching various functional groups onto peptides, such as fluorescent moieties. A high degree of chemical specificity is required to avoid down-stream misidentification of amino acids, which could lead to sequencing errors. Chemists are making headway on a suite of selective and high-yield methods for labeling specific amino acid side chains, amino acid termini, and post-translational modifications with minimal cross reactivity (**Box 5**).

Labelling stability and efficiency is paramount to the success of sequencing technologies, but is also a challenge the chemists face. First, modification of most or all individual residues of one amino acid type is desired for an explicit identification of a peptide sequence, which requires selective and highly efficient reactions. Second, error-free sequence prediction requires multiple chemical labels, but the stability of the chemical labels has been an issue in some sequencing techniques. These issues have been best characterized for fluorosequencing (**Box 5a**).

For many of the sequencing techniques, amino acids must be labeled with a chemical tag to allow for the differentiation between the amino acids. While it is theoretically possible to get a broad coverage of the proteome with a minimal set of amino acid labeling, specific identification of peptides or broader sequence coverage requires a larger suite of labels. Overall, there are twelve distinct side chain types in peptides ranging from highly reactive amino acids like lysine and cysteine, to functional groups that are more challenging to modify, such as amides (Gln/Asn) and alkanes (Ala, Gly, Ile, Leu, Pro, and Val). There are a large number of methods to label amino acids, however some chemistries do not provide sufficiently stable bonds for some single-molecule sequencing approaches. To date, only eight (Lys, Cys, Glu/Asp, Tyr, Trp, His, and Arg) have thus far been shown to be stable, selective, and reactive enough for the single-molecule fluorosequencing approach[9,99]. Research is ongoing to test a wide variety of other labeling conditions to cover all of the proteinogenic amino acids (**Box 5b**).

Chemical modification of protein termini is highly desired for several sequencing techniques like fluorosequencing, nanopores and DNA-PAINT approaches where end labeling or ligation is required (**Figure 1**). The terminus provides an attachment point for surface immobilization and can offer a simple way to remove excess chemical reagents during procedures that require multiple labeling steps. Two terminus-specific methods have found great promise for single-molecule sequencing, C-terminal labeling using decarboxylative alkylation (**Box 5c**) and modification of the N-terminus with 2-pyridinecarboxaldehyde (**Box 5d**).

The long-term goal of characterizing proteoforms requires methods to detect and differentiate PTMs. They can be recognized by mass spectrometry through the mass shifts they cause on a protein, peptide and their fragments[100,101] and databases of the expected mass shifts like Unimod are used to support the identification[102]. However, these databases show that there can be significant overlaps between PTMs of the same or similar mass suggesting that orthogonal methods are needed. Single-molecule protein sequencing methods rely on either site-specific labeling or elimination and replacement chemistries (**Box 5e**).

Discussion: a spectrum of opportunities

An emerging landscape of single-molecule protein sequencing and fingerprinting technologies is being developed **(Figure 1)** with the promise to resolve the full proteome of single cells with single-protein resolution, opening up unprecedented opportunities in fundamental science and medical diagnostics. Cellular tissues' composition could then be resolved with single-cell resolution, opening up new research avenues from embryonic development to cancer

research. Diagnostics could benefit from the ultimate single-molecule resolution by resolving very low amounts of protein in bodily samples. The detection of rare proteins with copy numbers as low as one or a few may uncover new molecular regulatory networks within cells. Some of the emerging technologies described here are still at their early proof-of-concept stages of development, whereas others like sequencing by Edman degradation and nanopore sequencing technologies have already attracted industry funding. Additional single-molecule approaches are also promoted by commercial entities, and are out of scope for this review.

A real-world application of a technology that is not MS- or antibody-based for whole proteome characterization is yet to be achieved. In the meantime, MS will continue to improve in its capacity, to support single-ion detection[22] and single-cell proteomics[103]. Similarly, antibody-based strategies such as immunoassays that rely on specific antigen-antibody interactions have served as the standard methods for protein identification and quantification for the last few decades. Single Molecule Array technologies (Simoa[104]) commercialized by Quanterix is one of the most sensitive single protein sensing antibody-based methods used for the analysis of small analytical samples and clinical studies down to attomolar concentration level[105]. The SARS-Cov-2 pandemic has accelerated the development of a high-throughput serological tests of clinical samples utilizing Simoa[106] based on ultrasmall blood samples. These and other antibody based protein sensing method are likely to take greater share in the biomedical sensing industry, in parallel to the emergence of other single molecule techniques that will further permit comprehensive proteoform inference or differentiation.

The emerging landscape of alternative protein sequencing and fingerprinting technologies in **Figure 1** could one day help to sequence human proteoforms in a more complete way. High-throughput Edman degradation could pair with bottom-up MS strategies to improve current sequence coverage limitations (**Box 1**). These bottom-up methods could benefit from nanopore sequencing and DNA fluorescence-based methods that aim for long read sequencing and structural fingerprinting of whole proteins. The integration of both existing and emerging technologies promises to iteratively reveal an atlas of full length proteoforms, which could itself assist these up-and-coming technologies to infer what cannot be directly measured in terms of protein primary sequence and structure.

The far-reaching vision in single-molecule proteomics is in their applications for the analysis of protein-protein interactions. A map covering a wide range of proteoforms and their PPIs is an unmet milestone needed to finely understand protein networks in normal tissues and in disease. Bottom-up MS-based approaches, such as cross-linking[107,108] or affinity-purification

are implemented to identify physical[109] and proximal interactomes[110]. However, these techniques present either biochemical or sample processing yield limitations, which brings problems, including intra-protein cross-link over-representation, PPIs loss upon solubilization and limitations inherent to MS analysis, hindering single-cell interactome analysis. As of today, single-molecule analysis of PPIs has not reached main-stream proteomics, and single-cell interactomics even more so. Achieving this goal would be of outstanding interest for accurately defining *e.g.* protein organization within highly dynamic membraneless organelles[111], such as resolving protein condensates, spatial and temporal organization at a single organelle or single cell scale, which will provide an unprecedented resolution of PPIs organization.

Challenges for next-generation protein sequencing

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

To gain utility in both, research and clinical settings, emerging technologies will be evaluated in terms of their sensitivity, specificity, proteome coverage (number of proteins in the sample covered), sequence coverage (average fraction of a protein sequence covered), peptide read length (number of amino-acids covered by a single read), accuracy (error in calling an amino-acid) and cost. In this regard, additional research and validation will be required to demonstrate the benefits of these orthogonal technologies. The formation of a dedicated global academic/scientific community in single protein sequencing may catalyze further development and implementation of these technologies for more widespread use. Multidisciplinary conferences that bring together experts in chemistry, physics, biochemistry,

industry, computation, and other relevant areas of expertise (e.g. pathologists, clinicians) with a clear vision of the most relevant problems and unmet needs, will need to be embraced.

Acknowledgements

The authors thank all the presenting delegates of the 2019 Single Molecule Protein Sequencing conference (Jerusalem). Authors thank the PL-Grid Infrastructure, Poland for providing their hardware and software resources. S.S. acknowledges the Postdoc Mobility fellowship no. P400PB 180889 by the Swiss National Science Foundation. E.M.M. and E.V.A. acknowledge funding from the NIH (R35 GM122480, R01 DK110520 to EMM), Welch Foundation (F1515 to EMM and F-0046 to EVA), Army Research Office grant W911NF-12-1-0390, and Erisyon, Inc. E.M.M. and E.V.A. are co-founders and shareholders of Erisyon. RTK acknowledges funding from NIGMS (R01 GM138931). C.D. acknowledges the ERC Advanced Grant Looping DNA (no. 883684) and the NWO programs NanoFront and Basyc. E.M.M., E.V.A., and C.J.H. are co-inventors on patents relevant to this work. S.O acknowledges the support of the Azrieli fellowship foundation. NLK acknowledges funding from Paul G. Allen Frontiers Program (11715), the NIH HuBMAP program (UH3 CA246635) and NIGMS (P41 GM108569). J.P.M. and Z.K acknowledge internal funding from NIST and are co-inventors on patents relevant to this work. M.W. acknowledges funding from the NIH (HG009186). K.S. and A.A. acknowledge funding from NSF (PHY-1430124). C. Joo, C.D, and R.E. acknowledge funding from NWO-I (SMPS). Joo acknowledges funding from HFSP (RGP0026/2019). A.P. acknowledges the Bekker fellowship no. PPN/BEK/2018/1/00296 from the Polish National Agency for Academic Exchange. .M. and S.H. acknowledge funding from the European Research Council (ERC “Enlightened”, GA 616251) and the CEA Transverse Program “Instrumentation and Detection” (PTC-ID VIRIONEMS). Support from the Proteomics French Infrastructure (PROFI) is also gratefully acknowledged. GD acknowledges funding from FNR (C17/BM/11642138). M.M. acknowledges funding from the Adolphe Merkle Foundation, the Michael J. Fox Foundation for Parkinson’s Research (Grant ID: 17924) and the Swiss National Science Foundation (grant no. 200021-169304). A.M acknowledges funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 833399-ERC NanoProt-ID, and ISF award 3485/19. MC acknowledges the computational resources from CINECA (NATWE project), and the Swiss National Super-computing Centre (CSCS), projects ID sm11 and s865. E.C. acknowledges funding from I-Site Lille, Région Hauts-de-France and European Union’s Horizon 2020 Marie Skłodowska-Curie No 843052. The study was supported by the project “International Centre for Cancer Vaccine Science” that is carried out within the International Agendas Programme of the Foundation for Polish

Science co-financed by the European Union under the European Regional Development Fund. We thank Viktorija Globyte for critical reading.

Author Contributions

J.A. Alfaro., C. Joo, and A. Meller conceived of and initiated, coordinated, and supervised the project. The first draft of the manuscript was written by J.A. Alfaro, C. Joo, A. Meller, P. Bohländer, M. Filius, X. F. van Kooten , S. Ohayon, A. Pomorski, S. Schmid, C. J. Howard, M. Dai, P. Samaras, G. Bedran, M. Wilhelm and L. Sepiashvili. Subsequently, the manuscript was revised and approved by all authors.

Competing Interests:

S. Hentz and C. Masselon are co-inventors of the patent application EP14158255. E. Marcotte is a co-inventor on patent 9625469. D. Stein is sponsored by Oxford Nanopore for his work on nanotip mass spectrometry. E.M. Marcotte and E.V. Anslyn are co-founders and shareholders of Erisyon, Inc. Some authors may be bound by confidentiality agreements that prevent them from disclosing their competing interests in this work.

References:

1. Breuza, L. *et al.* The UniProtKB guide to the human proteome. *Database* **2016**, 1–10 (2016).
2. Smith, L. M. *et al.* Proteoform: a single term describing protein complexity. *Nat. Methods* **10**, 186–187 (2013).
3. Seattle Times business staff. Seattle biotech startup Nautilus to get $350 million, stock listing in blank-check deal. *The Seattle Times.* (2021). https://www.seattletimes.com/business/seattle-biotech-startup-nautilus-to-get-350-million-stock-listing-in-blank-check-deal/
4. Reuters Staff. Protein sequencing firm Quantum-Si to go public via $1.46 billion SPAC merger. *Reuters* (2021). https://www.reuters.com/article/us-quantum-si-m-a-highcape-capital-idUSKBN2AI1HT
5. Cohen, L. & Walt, D. R. Single-Molecule Arrays for Protein and Nucleic Acid Analysis. *Annu. Rev. Anal. Chem.* **10**, 345–363 (2017).
6. Aggarwal, V. & Ha, T. Single-molecule fluorescence microscopy of native macromolecular complexes. *Curr. Opin. Struct. Biol.* **41**, 225–232 (2016).
7. Edman, P. A method for the determination of the amino acid sequence in peptides. *Arch. Biochem.* **22**, 475–476 (1949).
8. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **11**, 1076–1082 (2015).
9. Swaminathan, J. *et al.* Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* **36**, 1076–1082 (2018).
10. Howard, C. J. *et al.* Solid-phase peptide capture and release for bulk and single-molecule proteomics. *ACS Chem. Biol.* **15**, 1401–1407 (2020).
11. Miclotte, G., Martens, K. & Fostier, J. Computational assessment of the feasibility of protonation-based protein sequencing. *PLoS One* **15**, e0238625 (2020).

12. Tullman, J., Marino, J. P. & Kelman, Z. Leveraging nature's biomolecular designs in next-generation protein sequencing reagent development. *Appl. Microbiol. Biotechnol.* 1–11 (2020).
13. Rodriques, S. G., Marblestone, A. H. & Boyden, E. S. A theoretical analysis of single molecule protein sequencing via weak binding spectra. *PLoS One* **14**, e0212868 (2019).
14. Tullman, J., Callahan, N., Ellington, B., Kelman, Z. & Marino, J. P. Engineering ClpS for selective and enhanced N-terminal amino acid binding. *Appl. Microbiol. Biotechnol.* **103**, 2621–2633 (2019).
15. Smith, R. D., Cheng, X., Brace, J. E., Hofstadler, S. A. & Anderson, G. A. Trapping, detection and reaction of very large single molecular ions by mass spectrometry. *Nature* **369**, 137–139 (1994).
16. Keifer, D. Z. & Jarrold, M. F. Single-molecule mass spectrometry. *Mass Spectrom. Rev.* **36**, 715–733 (2017).
17. Rose, R. J., Damoc, E., Denisov, E., Makarov, A. & Heck, A. J. High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. *Nat. Methods* **9**, 1084–1086 (2012).
18. Makarov, A. & Denisov, E. Dynamics of ions of intact proteins in the Orbitrap mass analyzer. *J. Am. Soc. Mass Spectr.* **20**, 1486–1495 (2009).
19. Kafader, J. O. *et al.* Measurement of individual ions sharply increases the resolution of orbitrap mass spectra of proteins. *Anal. Chem.* **91**, 2776–2783 (2019).
20. Kafader, J. O. *et al.* Multiplexed mass spectrometry of individual ions improves measurement of proteoforms and their complexes. *Nat. Methods* **17**, 391–394 (2020).
21. Wörner, T. P. *et al.* Resolving heterogeneous macromolecular assemblies by Orbitrap-based single-particle charge detection mass spectrometry. *Nat. Methods* **17**, 395–398 (2020).

22. Kafader, J. O. *et al.* Individual ion mass spectrometry enhances the sensitivity and sequence coverage of top down mass spectrometry. *J. Proteome Res.* **19**, 1346–1350 (2020).
23. Smith, L. *et al.* The Human Proteoform Project: A Plan to Define the Human Proteome. *Preprints* (2020) doi:10.20944/preprints202010.0368.v1.
24. Ekinci, K. L., Huang, X. M. H. & Roukes, M. L. Ultrasensitive nanoelectromechanical mass detection. *Appl. Phys. Lett.* **84**, 4469–4471 (2004).
25. Hanay, M. S. *et al.* Single-protein nanomechanical mass spectrometry in real time. *Nat. Nanotechnol.* **7**, 602-608 (2012).
26. Sage, E. *et al.* Neutral particle mass spectrometry with nanomechanical systems. *Nat. Commun.* **6**, 1–5 (2015).
27. Dominguez-Medina, S. *et al.* Neutral mass spectrometry of virus capsids above 100 megadaltons with nanomechanical resonators. *Science* **362**, 918–922 (2018).
28. Chaste, J. *et al.* A nanomechanical mass sensor with yoctogram resolution. *Nat. Nanotechnol.* **7**, 301–304 (2012).
29. Hanay, M. S. *et al.* Inertial imaging with nanomechanical systems. *Nat. Nanotechnol.* **10**, 339–334 (2015).
30. Malvar, O. *et al.* Mass and stiffness spectrometry of nanoparticles and whole intact bacteria by multimode nanomechanical resonators. *Nat. Commun.* **7**, 1–8 (2016).
31. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68**, 1–8 (1996).
32. El-Faramawy, A., Siu, K. M. & Thomson, B. A. Efficiency of nano-electrospray ionization. *J. Am. Soc. Mass Spectr.* **16**, 1702–1707 (2005).
33. Bush, J. *et al.* The nanopore mass spectrometer. *Rev. Sci. Instrum.* **88**, 113307 (2017).
34. Maulbetsch, W., Wiener, B., Poole, W., Bush, J. & Stein, D. Preserving the sequence of a biopolymer's monomers as they enter an electrospray mass spectrometer. *Phys. Rev. Appl.* **6**, 054006-1–054006-9 (2016).

35. Brodbelt, J. S. Photodissociation mass spectrometry: new tools for characterization of biological molecules. *Chem. Soc. Rev.* **43**, 2757–2783 (2014).
36. Chang, S. *et al.* Tunnelling readout of hydrogen-bonding-based recognition. *Nat. Nanotechnol.* **4**, 297–301 (2009).
37. Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* **9**, 466–473 (2014).
38. Ohshiro, T. *et al.* Detection of post-translational modifications in single peptides using electron tunnelling currents. *Nat. Nanotechnol.* **9**, 835 (2014).
39. Zhang, B. *et al.* Observation of giant conductance fluctuations in a protein. *Nano Futures* **1**, 035002 (2017).
40. Zhang, B. *et al.* Engineering an enzyme for direct electrical monitoring of activity. *ACS Nano* **14,** 1360-1368 (2020).
41. Seeman, N. C. & Sleiman, H. F. DNA nanotechnology. *Nat. Rev. Mat.* **3**, 1–23 (2017).
42. Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with DNA-PAINT. *Nat. Protoc.* **12**, 1198–1228 (2017).
43. Dai, M., Jungmann, R. & Yin, P. Optical imaging of individual biomolecules in densely packed clusters. *Nat. Nanotechnol.* **11**, 798–807 (2016).
44. Jungmann, R. *et al.* Quantitative super-resolution imaging with qPAINT. *Nat. Methods* **13**, 439–442 (2016).
45. Dai, M. & Yin, P. Methods and compositions relating to super-resolution imaging and modification. (2018).
46. Schaus, T. E., Woo, S., Xuan, F., Chen, X. & Yin, P. A DNA nanoscope via auto-cycling proximity recording. *Nat. Commun.* **8**, 1–9 (2017).
47. Kishi, J. Y., Schaus, T. E., Gopalkrishnan, N., Xuan, F. & Yin, P. Programmable autonomous synthesis of single-stranded DNA. *Nat. Chem.* **10**, 155–164 (2018).
48. Gopalkrishnan, N., Punthambaker, S., Schaus, T. E., Church, G. M. & Yin, P. A DNA nanoscope that identifies and precisely localizes over a hundred unique molecular

features with nanometer accuracy. *bioRxiv* 2020.08.27.271072 (2020) doi:10.1101/2020.08.27.271072.

49. Filius, M., Kim, S. H., Severins, I. & Joo, C. High-resolution single-molecule FRET via DNA eXchange (FRET X). *bioRxiv* 2020.10.15.340885 (2020) doi:10.1101/2020.10.15.340885.

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

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

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

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

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

55. Di Muccio, G., Rossini, A. E., Di Marino, D., Zollo, G. & Chinappi, M. Insights into protein sequencing with an α-Hemolysin nanopore by atomistic simulations. *Sci. Rep.* **9**, 1–8 (2019).

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

57. Huang, G., Voet, A. & Maglia, G. FraC nanopores with adjustable diameter identify the mass of opposite-charge peptides with 44 dalton resolution. *Nat. Commun.* **10**, 1–10 (2019).

58. Piguet, F. *et al.* Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. *Nat. Commun.* **9**, 1–13 (2018).

59. Cao, C. *et al.* Single-molecule sensing of peptides and nucleic acids by engineered aerolysin nanopores. *Nat. Commun.* **10**, 1–11 (2019).
60. Galenkamp, N. S., Soskine, M., Hermans, J., Wloka, C. & Maglia, G. Direct electrical quantification of glucose and asparagine from bodily fluids using nanopores. *Nat. Commun.* **9**, 1–8 (2018).
61. Ouldali, H. *et al.* Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat. Biotechnol.* **38**, 176–181 (2020).
62. Restrepo-Pérez, L., Wong, C. H., Maglia, G., Dekker, C. & Joo, C. Label-free detection of post-translational modifications with a nanopore. *Nano Lett.* **19**, 7957–7964 (2019).
63. Korotkov, K. V., Sandkvist, M. & Hol, W. G. The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat. Rev. Microbiol.* **10**, 336–351 (2012).
64. Olivares, A. O., Baker, T. A. & Sauer, R. T. Mechanical protein unfolding and degradation. *Annu. Rev. Physiol.* **80**, 413–429 (2018).
65. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* **31**, 247–250 (2013).
66. 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).
67. Zhang, S. *et al.* Bottom-up fabrication of a multi-component nanopore sensor that unfolds, processes and recognizes single proteins. *bioRxiv* 2020.12.04.411884 (2020) doi:10.1101/2020.12.04.411884.
68. Sachelaru, I. *et al.* YidC and SecYEG form a heterotetrameric protein translocation channel. *Sci. Rep.* **7**, 1–15 (2017).
69. Knyazev, D. G., Kuttner, R., Zimmermann, M., Sobakinskaya, E. & Pohl, P. Driving forces of translocation through bacterial translocon SecYEG. *J. Membrane Biol.* **251**, 329–343 (2018).

70. Backes, S. & Herrmann, J. M. Protein translocation into the intermembrane space and matrix of mitochondria: mechanisms and driving forces. *Front. Mol. Biosci.* **4**, 83 1-11 (2017).
71. Feng, J. *et al.* Transmembrane protein rotaxanes reveal kinetic traps in the refolding of translocated substrates. *Commun. Biol.* **3**, 1–11 (2020).
72. Rosen, C. B., Bayley, H. & Rodriguez-Larrea, D. Free-energy landscapes of membrane co-translocational protein unfolding. *Commun. Biol.* **3**, 1–9 (2020).
73. Cardozo, N. *et al.* Multiplexed direct detection of barcoded protein reporters on a nanopore array. *bioRxiv* 837542 (2019) doi:10.1101/837542.
74. 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).
75. Ohayon, S., Girsault, A., Nasser, M., Shen-Orr, S. & Meller, A. Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. *PLoS Comput. Biol.* **15**, e1007067 (2019).
76. Restrepo-Pérez, L. *et al.* Resolving chemical modifications to a single amino acid within a peptide using a biological nanopore. *ACS Nano* **13**, 13668-13676 (2019).
77. Asandei, A. *et al.* Placement of oppositely charged aminoacids at a polypeptide termini determines the voltage-controlled braking of polymer transport through nanometer-scale pores. *Sci. Rep.* **5**, 10419 (2015).
78. Wang, R. *et al.* Single-molecule discrimination of labeled DNAs and polypeptides using photoluminescent-free TiO2 nanopores. *ACS Nano* **12**, 11648–11656 (2018).
79. Zrehen, A., Ohayon, S., Huttner, D. & Meller, A. On-chip protein separation with single-molecule resolution. *Sci. Rep.* **10**, 1–12 (2020).
80. Assad, O. N. *et al.* Light-enhancing plasmonic-nanopore biosensor for superior single-molecule detection. *Adv. Mater.* **29**, 1605442 (2017).

81. Spitzberg, J. D., Zrehen, A., van Kooten, X. F. & Meller, A. Plasmonic-nanopore biosensors for superior single-molecule detection. *Adv. Mater.* **31**, 1900422 (2019).
82. Houghtaling, J., List, J. & Mayer, M. Nanopore-based, rapid characterization of Individual amyloid particles in solution: concepts, challenges, and prospects. *Small* **14**, 1802412 (2018).
83. Plesa, C. *et al.* Fast translocation of proteins through solid state nanopores. *Nano Lett.* **13**, 658–663 (2013).
84. Yusko, E. C. *et al.* Controlling protein translocation through nanopores with bio-inspired fluid walls. *Nat. Nanotechnol.* **6**, 253 (2011).
85. Houghtaling, J. *et al.* Estimation of shape, volume, and dipole moment of individual proteins freely transiting a synthetic nanopore. *ACS Nano* **13**, 5231–5242 (2019).
86. Yusko, E. C. *et al.* Real-time shape approximation and fingerprinting of single proteins using a nanopore. *Nat. Nanotechnol.* **12**, 360 (2017).
87. Pang, Y. & Gordon, R. Optical trapping of a single protein. *Nano Lett.* **12**, 402–406 (2012).
88. Verschueren, D., Shi, X. & Dekker, C. Nano-optical tweezing of single proteins in plasmonic nanopores. *Small Methods* **3**, 1800465 (2019).
89. Schmid, S. & Dekker, C. Nanopores–a versatile tool to study protein dynamics. Essays Biochem. EBC20200020 (2020).
90. Larkin, J., Henley, R. Y., Muthukumar, M., Rosenstein, J. K. & Wanunu, M. High-bandwidth protein analysis using solid-state nanopores. *Biophys. J.* **106**, 696–704 (2014).
91. Nir, I., Huttner, D. & Meller, A. Direct sensing and discrimination among ubiquitin and ubiquitin chains using solid-state nanopores. *Biophys. J.* **108**, 2340–2349 (2015).
92. Waduge, P. *et al.* Nanopore-based measurements of protein size, fluctuations, and conformational changes. *ACS Nano* **11**, 5706–5716 (2017).

93. Varongchayakul, N., Hersey, J. S., Squires, A., Meller, A. & Grinstaff, M. W. A Solid-state hard microfluidic–nanopore biosensor with multilayer fluidics and on-chip bioassay/purification chamber. *Adv. Funct. Mater.* **28**, 1804182 (2018).
94. Hu, R. *et al.* Differential enzyme flexibility probed using solid-state nanopores. *ACS Nano* **12**, 4494–4502 (2018).
95. Huang, G. *et al.* Electro-osmotic vortices promote the capture of folded proteins by PlyAB nanopores. *Nano Lett.* **20**, 3819-3827 (2020).
96. Soskine, M., Biesemans, A. & Maglia, G. Single-molecule analyte recognition with ClyA nanopores equipped with internal protein adaptors. *J. Am. Chem. Soc.* **137**, 5793–5797 (2015).
97. Wloka, C. *et al.* Label-free and real-time detection of protein ubiquitination with a biological nanopore. *ACS Nano* **11**, 4387–4394 (2017).
98. Aramesh, M. *et al.* Localized detection of ions and biomolecules with a force-controlled scanning nanopore microscope. *Nat. Nanotechnol.* **14**, 791–798 (2019).
99. 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).
100. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513 (2017).
101. Zhong, J. *et al.* Proteoform characterization based on top-down mass spectrometry. *Brief. Bioinformatics*, bbaa015 (2020).
102. Creasy, D. M. & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **4**, 1534–1536 (2004).
103. Marx, V. A dream of single-cell proteomics. *Nat. Methods* **16**, 809–812 (2019).

104. Rissin, D. M. *et al.* Single-molecule enzyme-linked immunosorbent assay detects serum proteins at subfemtomolar concentrations. *Nat. Biotechnol.* **28**, 595–599 (2010).
105. Wu, C., Garden, P. M. & Walt, D. R. Ultrasensitive detection of attomolar protein concentrations by dropcast single molecule assays. *J. Am. Chem. Soc.* **142**, 12314–12323 (2020).
106. Norman, M. *et al.* Ultrasensitive high-resolution profiling of early seroconversion in patients with COVID-19. *Nat. Biomed. Eng.* **4**, 1180–1187 (2020).
107. Liu, F., Rijkers, D. T., Post, H. & Heck, A. J. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* **12**, 1179–1184 (2015).
108. Iacobucci, C., Götze, M. & Sinz, A. Cross-linking/mass spectrometry to get a closer view on protein interaction networks. *Curr. Opin. Biotechnol.* **63**, 48–53 (2020).
109. Dunham, W. H., Mullin, M. & Gingras, A.-C. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics* **12**, 1576–1590 (2012).
110. Gentzel, M., Pardo, M., Subramaniam, S., Stewart, A. F. & Choudhary, J. S. Proteomic navigation using proximity-labeling. *Methods* **164**, 67–72 (2019).
111. Zhao, Y. G. & Zhang, H. Phase separation in membrane biology: the interplay between membrane-bound organelles and membraneless condensates. *Dev. Cell* (2020).
112. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
113. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Anal. Chem.* **9**, 499–519 (2016).
114. Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat. Commun.* **9**, 1–10 (2018).
115. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 161 (2018).

116. Zhu, Y. *et al.* Proteomic analysis of single mammalian cells enabled by microfluidic nanodroplet sample preparation and ultrasensitive NanoLC-MS. *Angew. Chem. Int. Ed.* **57**, 12370–12374 (2018).
117. Kelly, R. T. Single-cell proteomics: Progress and prospects. *Mol. & Cell. Proteom.* **19**, 1739–1748 (2020).
118. Gavrilyuk, J., Ban, H., Nagano, M., Hakamata, W. & Barbas, C. F. Formylbenzene diazonium hexafluorophosphate reagent for tyrosine-selective modification of proteins and the introduction of a bioorthogonal aldehyde. *Bioconjugate Chem.* **23**, 2321–2328 (2012).
119. Ban, H., Gavrilyuk, J. & Barbas III, C. F. Tyrosine bioconjugation through aqueous ene-type reactions: a click-like reaction for tyrosine. *J. Am. Chem. Soc.* **132**, 1523–1525 (2010).
120. Bach, K., Beerkens, B. L., Zanon, P. R. & Hacker, S. M. Light-activatable, 2, 5-disubstituted tetrazoles for the proteome-wide profiling of aspartates and glutamates in living bacteria. *ACS Cent. Sci.* **6**, 546–554 (2020).
121. Taylor, M. T., Nelson, J. E., Suero, M. G. & Gaunt, M. J. A protein functionalization platform based on selective reactions at methionine residues. *Nature* **562**, 563–568 (2018).
122. Lin, S. *et al.* Redox-based reagents for chemoselective methionine bioconjugation. *Science* **355**, 597–602 (2017).
123. Christian, A. H. *et al.* A physical organic approach to tuning reagents for selective and stable methionine bioconjugation. *J. Am. Chem. Soc.* **141**, 12657–12662 (2019).
124. Jia, S., He, D. & Chang, C. J. Bioinspired thiophosphorodichloridate reagents for chemoselective histidine bioconjugation. *J. Am. Chem. Soc.* **141**, 7294–7301 (2019).
125. Bloom, S. *et al.* Decarboxylative alkylation for site-selective bioconjugation of native proteins via oxidation potentials. *Nat. Chem.* **10**, 205 (2018).

126. Rosen, C. B. & Francis, M. B. Targeting the N terminus for site-selective protein modification. *Nat. Chem. Biol.* **13**, 697–705 (2017).
127. Busch, G. K. *et al.* Specific N-terminal protein labelling: use of FMDV 3C pro protease and native chemical ligation. *Chem. Commun.* **29**, 3369–3371 (2008).
128. Bandyopadhyay, A., Cambray, S. & Gao, J. Fast and selective labeling of N-terminal cysteines at neutral pH via thiazolidino boronate formation. *Chem. Sci.* **7**, 4589–4593 (2016).
129. Agten, S. M., Dawson, P. E. & Hackeng, T. M. Oxime conjugation in protein chemistry: from carbonyl incorporation to nucleophilic catalysis. *J. Pept. Sci.* **22**, 271–279 (2016).
130. MacDonald, J. I., Munch, H. K., Moore, T. & Francis, M. B. One-step site-specific modification of native proteins with 2-pyridinecarboxyaldehydes. *Nat. Chem. Biol.* **11**, 326 (2015).
131. Matheron, L. *et al.* Improving the selectivity of the phosphoric acid β-elimination on a biotinylated phosphopeptide. *J. Am. Soc. Mass Spectr.* **23**, 1981–1990 (2012).
132. Du, J. *et al.* Metabolic glycoengineering: Sialic acid and beyond. *Glycobiology* **19**, 1382–1401 (2009).
133. Tommasone, S. *et al.* The challenges of glycan recognition with natural and artificial receptors. *Chem. Soc. Rev.* **48**, 5488–5505 (2019).

Boxes:

Box 1: mass spectrometry-based global proteomics

The last decade saw the maturation of mass spectrometry use in global proteomics. The typical proteomics workflow is "bottom-up" in nature and involves digesting a protein sample using a protease and characterizing the resulting peptides by mass spectrometry (MS)[112]. Two types of measurements are typically made in succession: (1) MS^1 spectra survey the masses of a set of peptides present in the mass-spectrometer at a given moment and (2) MS^2 spectra probe the structures of peptide ion species identified in the MS^1 survey by isolating, fragmenting and measuring the fragment masses of one or a few of them. Peptides identified from the MS^2 spectra are then mapped back to the proteins they come from to infer their overall protein abundances.

Current mass spectrometers have drawbacks in terms of their dynamic range, the read-length (peptide-length) of "sequenced" peptides, and biases in detectability arising from the ionization mechanism, transmission and the mass analyzer used. Consequently, although "top-down" proteomics methods capable of analyzing intact proteins exist[113], most state-of-the-art proteomics approaches characterize the proteome with high numbers of proteins but on average the proteins are characterized with low unambiguous sequence coverage and low sequencing depth. Different sample-preparation strategies, instruments and elution profiles can improve the numbers and average sequence coverage of proteins identified in an experiment. Summarizing the best single-sample run from 47 experiments (a summary of over 1000 distinct samples) in ProteomicsDB[49] (**Figure Box1**) reveals that even with complex sample preparation, the average sequence coverage for a single sample reaches just 33%.

Box 2: MS-based single-cell proteomics

The dream of extending mass spectrometry (MS)-based proteomics to the single-cell level eluded researchers for decades. Even as the sensitivity of MS instrumentation improved to provide single-cell-compatible detection limits, samples comprising at least thousands of cells were in practice required to obtain an in-depth proteome profile. Two recent advances have made single-cell proteomics a reality. Miniaturized sample processing workflows such as nanoPOTS[114] (Nanodroplet Processing in One pot for Trace Samples) have dramatically increased the efficiency of single cell sample preparation. NanoPOTS utilizes a robotic nanopipettor to interface with a microfabricated nanowell plate. The reduced surface contact and increased protein concentrations within the nanoliter droplets dramatically enhance

digestion kinetics and increase sample recovery for single cells and other trace samples. Concurrently, multiplexed strategies (i.e., Single Cell ProtEomics by Mass Spectrometry; SCoPE-MS)[115] have been developed in which proteins from single cells are labeled with unique isobaric tags, and several cells are analyzed together in the presence of a larger carrier sample. Single cells and carrier provide a combined MS signal for each protein, and unique reporter ions released upon fragmentation enable protein quantification for each cell. While nanoPOTS and SCoPE-MS originally enabled quantification of hundreds of proteins[115,116], the combination of the two techniques, as well as advances in miniaturized liquid chromatography and gas-phase separations, now enable >1000 proteins to be quantified from single mammalian cells[117].

Box 3 High-throughput Edman-Sequencing.

In high-throughput Edman fluorosequencing, proteins are digested to shorter peptides and immobilized on a glass surface using the C-terminus[8]. Multiple rounds of Edman degradation coupled to fluorescence microscopy are used for sequencing. Specific amino acids are covalently labeled with spectrally distinguishable fluorophores, and the peptide fingerprint comes from measuring the decrease in fluorescence of peptides following Edman degradation[9]. Much like in mass spectrometry, the partial sequence is mapped back to a reference proteome within a probabilistic framework. In another method, NAAB probes specifically recognize each N-terminal amino acid of an unlabeled peptide for more complete amino acid identification[12].

Box 4: DNA-PAINT

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

Box 5: Chemistry concepts in protein sequencing

(a) labelling efficiency and stability. The challenges in labeling efficiency and stability are well characterized in fluorosequencing, which uses harsh conditions (including neat trifluoroacetic acid) which can lead to the reversal of maleimide-labeled cysteine residues. To circumvent this reversal, fluorosequencing instead utilizes the iodoacetamide chemistry which generates a more stable bond. Another point of complexity is that full conversion is dictated by solvent accessibility of targeted amino acid side chains and has an influence on the labeling efficiency. However, modeling suggests labeling efficiencies and stabilities significantly less than 100% can be compensated for computationally, at least to some degree, during the reference database matching process[8].

(b) Labelling side chains. The most widely accessible labels are those that target lysine and cysteine residues using NHS esters, maleimide, and iodoacetamide reactive groups, respectively, (**Figure 5a** and **b**). Additionally, the phenol ring of tyrosine can be labeled using benzyl diazo groups[118] (**Figure 5c**), however, the attachment of fluorescent molecules generally requires a two-step labeling procedure due to the cross-reactivity with fluorescent molecules. Another robust bioconjugation method to selectively target tyrosine side chains is an ene-like reaction with cyclic diazodicaboxamides in aqueous buffer[119]. Carboxylic acids have also been labeled on peptides, but due to the similar reactivities between Asp, Glu, and the C-terminus, this has primarily been used on synthetic peptides. The method makes use of a standard technique (EDC-coupling) for binding amines covalently to carboxylic acids, forming an amide bond (**Figure 5d**). A recently reported promising bioconjugation approach has shown that light-activated 2,5-disubstituted tetrazoles are able to convert glutamic and aspartic acid residues in high yields[120]. Finally, tryptophan can be labeled at the C-2 position using sulfenyl chlorides (**Figure 5e**). However, this comes with limitations that the reaction is extremely water sensitive and the reactive group must be made *in situ*[99]. There are also promising new methods that allow for chemical modifications of other amino acids. Methionine, for example, can either be elegantly labeled with hypervalent iodine reagents[121] or by the use of urea-derived oxaziridines[122,123]. Recently, a histidine-selective conjugation methodology was reported where thiophosphorodichloridates selectively form a covalent bond with histidines in proteins[124].

(c) C-terminal labeling. Labeling of the C-terminus brings a challenge in that it must be separated from aspartic and glutamic acid, which carry the same functionality. A photoredox reaction on the C-terminus of peptides and proteins by de-carboxylation of the C-terminal carboxylic acid followed by an alkylation step by a Michael acceptor has been recently

reported[125],. Due to their higher oxidation potential, the carboxylates of internal amino acid chains are less prone to this modification, making the method highly site-selective. This technique has been applied to a variety of peptide substrates as well as the C-terminus-specific alkylation of human insulin A (**Figure 5f**).

(d) N-terminal labeling. Several methods exist for modifying the N-terminus[126]. Classic approaches like reductive amination with aldehydes or acylation with NHS-esters, which rely on pH control to increase the selectivity, are not sufficiently specific. Other strategies involve the side chain of the N-terminal amino acid. Native chemical ligation[127] or condensation reactions with aldehydes[128], could be used to label N-terminal cysteine, serine, threonine or tryptophan residues. Furthermore, oxidizing N-terminal serine or threonine residues to their corresponding aldehydes allows oxime conjugation with hydrazides or hydroxylamines[129]. A more general methodology has emerged where the N-terminal amine condenses with the 2-pyridinecarboxaldehyde (2PCA), forming an imine structure, which further reacts in a cyclisation with the nearby amide nitrogen of the second amino acid to form the stable imidazolidinone product[130]. This reaction has recently been shown to be useful for single-molecule peptide sequencing as a method for the immobilization of peptides onto a solid-phase resin, multiple chemical derivatization steps without purification, and subsequent traceless release prior to fluorosequencing[10].

(e) PTMs. As an example of elimination replacement chemistries, phospho-serine and phospho-threonine residues can be labeled by β-elimination followed by Michael addition (BEMA). In mass-spectrometry-based phospho-proteomics, it is used to introduce an additional trypsin cleavage site at the phosphorylated amino acid[131], whereas at the single molecule level it can be utilized to site-specifically attach a fluorescent label. Such approach has been established for Edman degradation described above[9].

Protein glycosylation can be complex, featuring many different types of monomeric units bound in possibly branching polymer structures. Their full structural characterization often requires derivatization and is done on glycans that are released from the protein. Therefore, schemes for understanding site-specific and simple glycosylation events should be the current focus. N-glycan anchoring asparagine residue can be converted to aspartate by glycan removal with PNGase F enzyme practically for all protein sequencing approaches, reducing the complexity to the detection of this mutation. Another possibility to introduce ite-selective labels is the incorporation of azide-tagged glycans by adding modified carbohydrates to the cell medium[132]. In other detection schemes the location could be also inferred using glycan-specific reporter molecules such as lectins, engineered proteins or aptamers[133].

Figure Legends:

Figure Box1: Sequence coverage in global proteomics studies. MS-based global proteomics studies identify and quantify the proteins with variable sequence coverage. The single best run from 47 publications present in proteomicsDB shows how sample-specific protein sequence coverage improves with sample preparation methods. Sequence coverage generally decreases with sample complexity and increases with time (cost) dedicated to studying the sample.

Figure 1: The emerging landscape of single-molecule protein sequencing and fingerprinting technologies. The new proteomics landscape can be understood in terms of the type of analyte that is being studied, the method of protein identification, and the target niches in proteomics. Various techniques, particularly those involving complex readout signals, are suitable to characterize short peptide sequences, while others are primed to characterize full-length proteins or larger complexes. Technologies may specialize in short peptides (Peptides in the figure), whole proteins (Proteins) or macomolecular complexes (Complexes). The method of protein identification may fingerprint certain classes of amino acids (aa-fingerprinting), reveal each amino acid down to its physiochemical class or better (aa-sequencing). Much like mass spectrometry, technologies might characterize proteins by their masses and/or the masses of their fragments (Mass spectrum). Other methods aim to characterize properties of folded proteins (Structural fingerprint). The target niches could include the study of specific PTMs or deciphering whole proteoforms (PTM/proteoform inference), analyzing purified proteins or complex mixtures of proteins (Complex mixtures). Other applications can include protein interaction inference (PPI-studies) or glimpsing insights into protein structure (Structure).

Figure 2: The renaissance of classic techniques. High-throughput fluorosequencing by Edman degradation featuring **(a)** amino acid-specific chemical modification of peptides with fluorophores and **(b)** N-terminal amino acid recognition using a plurality of probes. **(c)** Neutral particle mass spectrometry is a promising technique to characterize proteoforms. Currently, the technology can be used to characterize large megadalton-scale complexes using Si-based nanosensors. Graphene-nanosensors and further developments may push the technology towards smaller and smaller proteins and potentially lead to increased sequence coverage in global proteomics. Electrospray Ionization (ESI) **(d)** Nanopore electrospray is a marriage of

nanopores, classical electrospray, and single-particle detection techniques to sequence single proteins by measuring the amino acids one at a time.

Figure 3: DNA facilitated protein sequencing. (a) Schematic of specific amino acid labelling on a denatured protein with DNA strands. Each DNA strands contains both a barcode for the specific amino acid, and (optionally) a unique molecular identifier (UMI). **(b-e)** Illustration of various readout strategies of DNA-labelled samples, for protein identification. **(b)** Protein kinetic fingerprinting using quantitative DNA-PAINT. **(c)** Protein linear barcoding using molecular-resolution DNA-PAINT. **(d)** DNA Proximity Recording. **(e)** Protein structural fingerprinting using DNA-FRET-PAINT.

Figure 4: Three strategies of Nanopore-based protein sequencing and sensing. In all cases, an electrical force is used to translocate either a linearized or a folded protein through a nanoscale aperture (red arrow). **(a)** Reading unlabeled proteins or peptides using a biological nanopore. **(b)** Identification of whole proteins and peptides by fingerprinting with deep learning algorithms. Residue-specific fluorescent labels (e.g. at K, C, M) can be used to fingerprint proteins and peptides alongside electrical current sensing. **(c)** Identification of folded proteins using lipid tethering. Other tethers might include DNA carriers, DNA origami anchors, or plasmonic trapping.

Figure 5: Chemistry for protein sequencing. (a) Lysine labeling with NHS esters **(b)** Cysteine labeling with iodoacetamide reactive groups **(c)** Strategies for labeling the phenol ring of tyrosine **(d)** Aspartate/Glutamate labeling **(e)** Tryptophan Labeling with sulfenyl chlorides**. (f)** C-terminal derivatization through Monoalkylation of A chain (41%).

Figure Box 1: Sequence coverage in global proteomics studies with MS

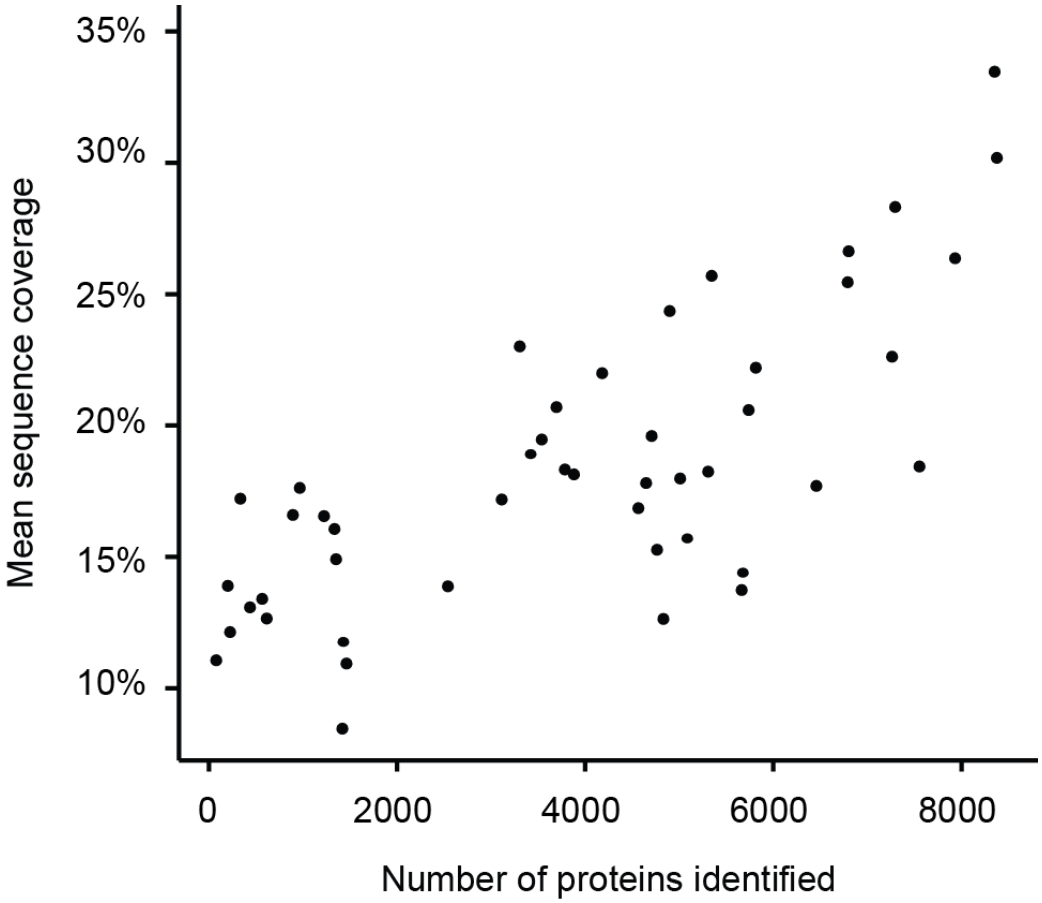

Figure 1: The emerging landscape of single-molecule protein sequencing and fingerprinting technologies.

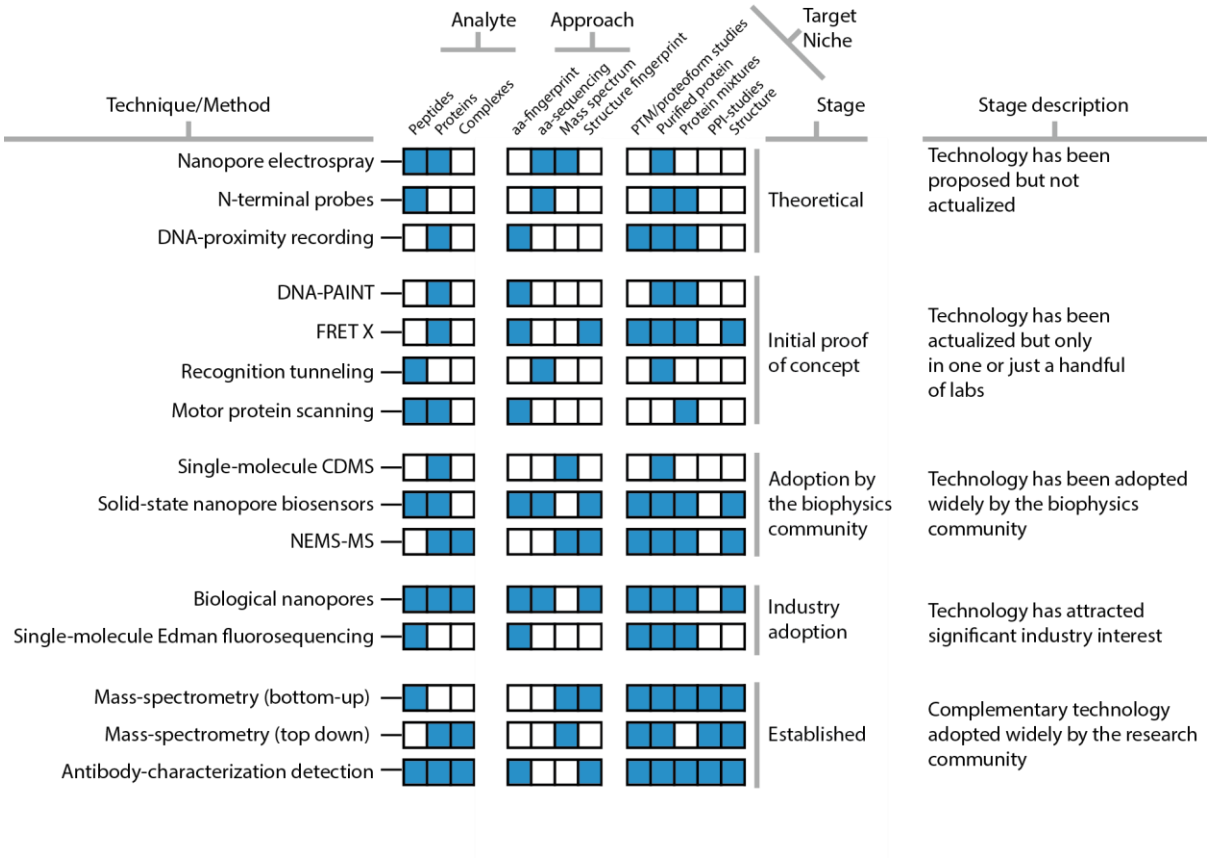

Figure 2: The renaissance of classical techniques

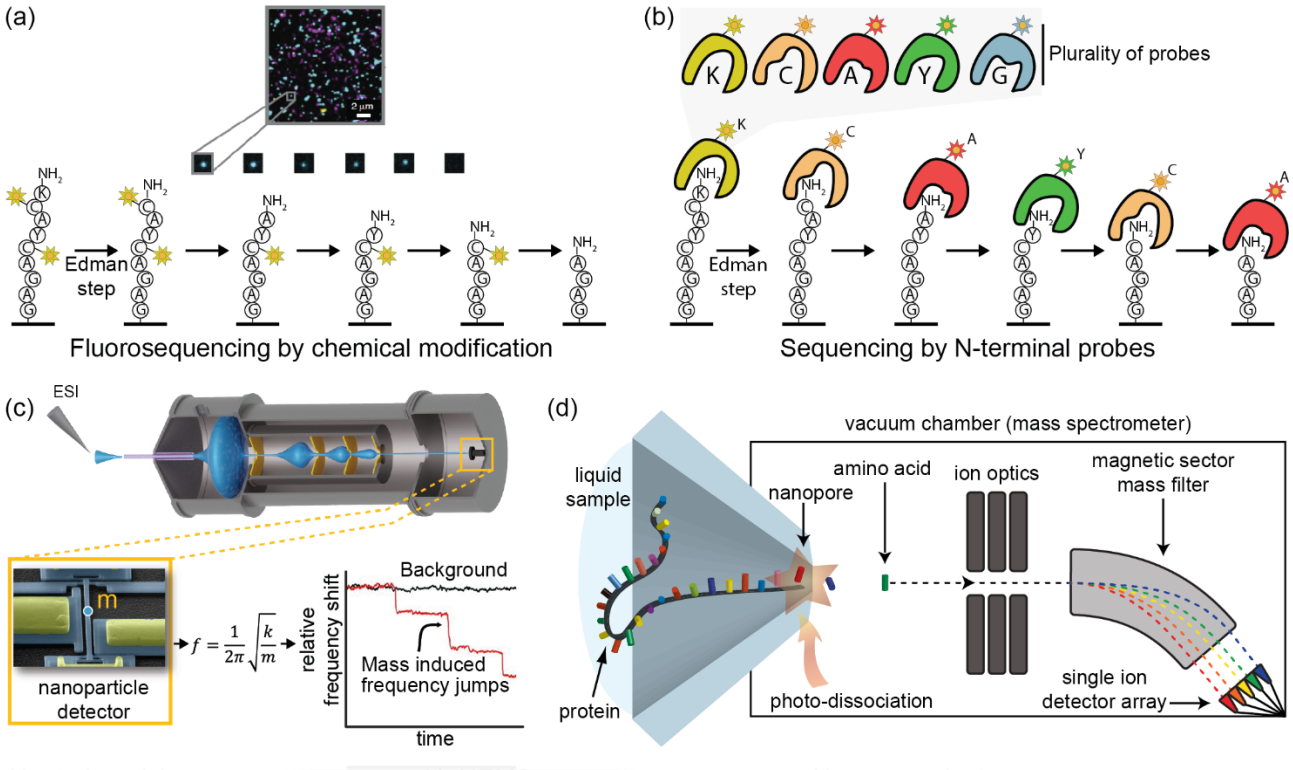

Figure 3: DNA-facilitated protein sequencing

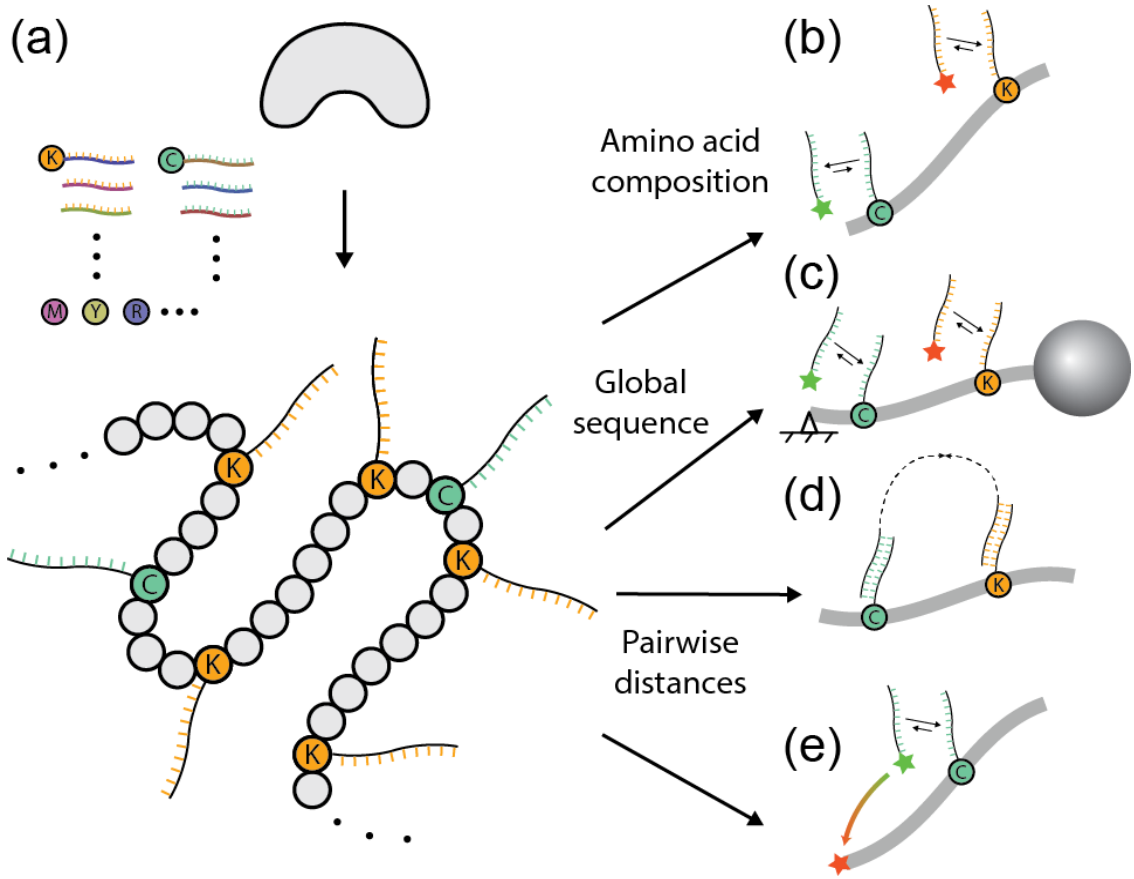

Figure 4: Nanopore-based protein sequencing

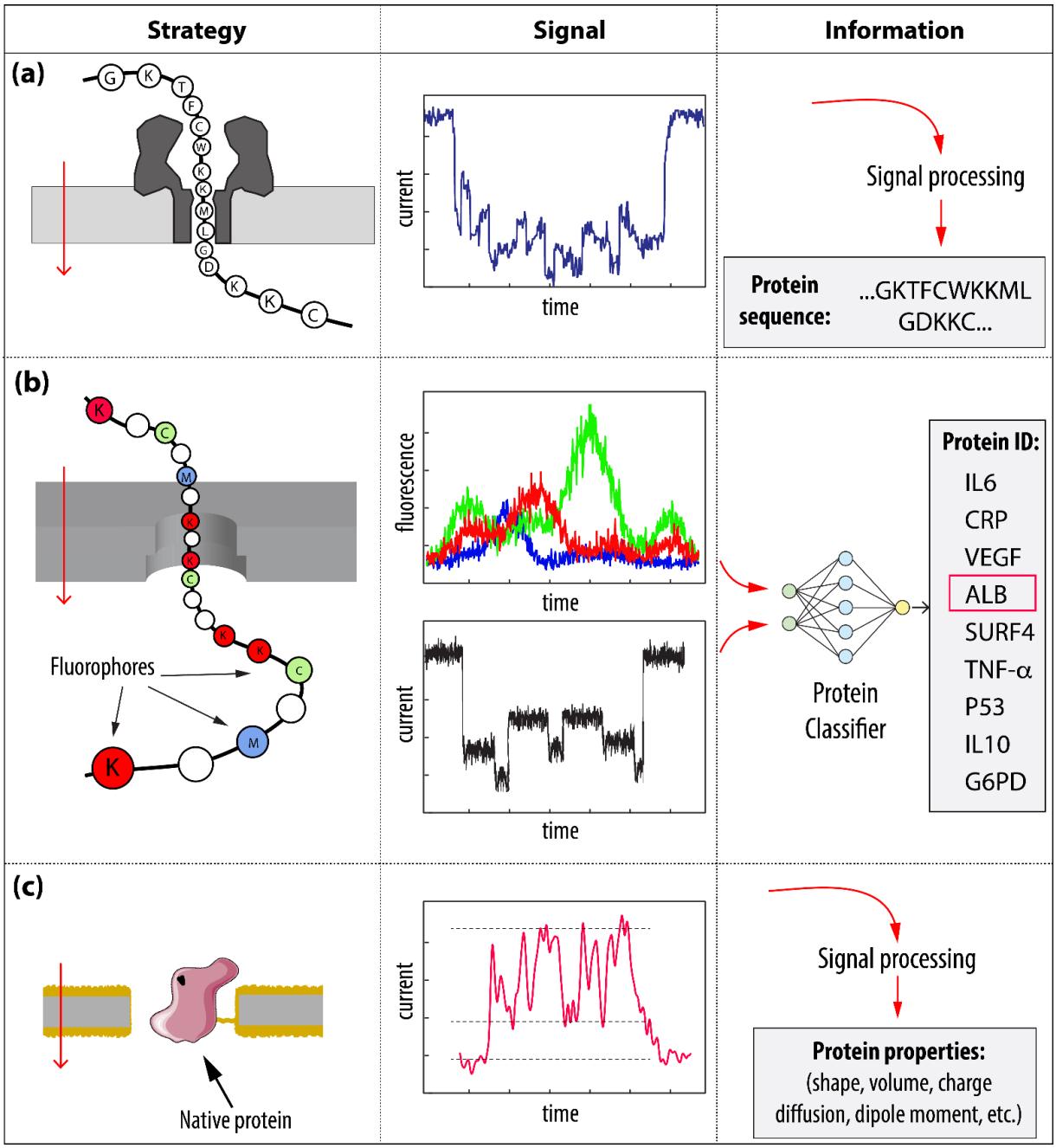

Figure 5: Chemistry for protein sequencing