Published in final edited form as:

Trends Biochem Sci. 2020 January ; 45(1): 76–89. doi:10.1016/j.tibs.2019.09.005.

Strategies for Development of a Next-Generation Protein Sequencing Platform

Nicholas Callahan¹, Jennifer Tullman¹, Zvi Kelman^{1,2}, John Marino¹

¹Institute for Bioscience and Biotechnology Research (IBBR), National Institute of Standards and Technology (NIST), and University of Maryland (UMD); 9600 Gudelsky Drive, Rockville, MD 20850, USA

²Biomolecular Labeling Laboratory, IBBR, 9600 Gudelsky Drive, Rockville, MD 20850, USA

Abstract

Proteomic analysis can be a critical bottleneck in cellular characterization. The current paradigm relies primarily on mass spectrometry of peptides and affinity reagents (i.e. antibodies), both of which require a *priori* knowledge of the sample. A non-biased protein sequencing method, with a dynamic range that covered the full range of protein concentrations in proteomes, would revolutionize the field of proteomics, allowing a more facile characterization of novel gene products and subcellular complexes. To this end, several new platforms based on single-molecule protein-sequencing approaches have been proposed. This review summarizes four of these approaches, highlighting advantages, limitations and challenges for each method towards advancing as a core technology for next-generation protein sequencing.

Keywords

Single-molecule analysis; Peptide sequencing; Proteomics

Proteomics lags behind Genomics and Transcriptomics

The central dogma of life science is that cells encode genetic information in DNA, use DNA to transcribe messenger RNA (mRNA), and use mRNA to translate genetic information into proteins. Because cells alter their RNA, protein, and metabolite levels in quick response to stimuli, no single step in this process contains all the relevant information about a cell's health and active pathways [1-4]. Rather, combined data from **genomics** (see Glossary), **transcriptomics**, **proteomics**, and **metabolomics** are needed to understand cellular pathways. In turn, in-depth understanding of cellular pathways opens up new potentials in drug discovery, personalized medicine, and **synthetic biology**.

Genomics and transcriptomics have been accelerated in the past decade by high-throughput, low cost sequencing technologies; these allow for the rapid and parallel sequencing of tens of thousands of unique, immobilized DNA and RNA molecules [5,6]. These sensitive

techniques also allow direct quantification of gene copy numbers and transcription levels [7]. Similar technologies have yet to emerge for rapid, sequencing based identification and quantification of proteins. The effect of this lag is exemplified by the so-called “missing proteome” which refers to the predicted open reading frames of the human genome for which no gene product has been identified [8]. Although resolving the missing proteome is one of the stated aims of the Human Proteome Project, an international collaboration started in 2001 to catalog the protein content of all human tissues [8], forty-four percent of computationally predicted human genes still have no assigned full-length protein product, nearly two decades after the first human genome sequence became available [9].

In this review, we first briefly introduce the present limits of proteomics followed by a discussion of four potential platforms for next-generation protein sequencing. We focus on how each platform proposes to read out a (currently limited) subset of amino acids from isolated single peptides derived from proteome samples. This will not be comprehensive, but is meant to compare the breadth of proposed technologies with potential for high-resolution protein sequencing. The surveyed methods include translocation of proteins through ångström-scale pores, amino acid sensing using electron tunneling spectroscopy, fluorescent imaging of protein digestion with ClpXP protease, and fluorescent imaging of the Edman degradation of immobilized peptides. We discuss the capabilities and limitations of each method and also touch on their possible future developments.

Current limitations of proteomics

Characterizing low-abundance proteins

Protein copy numbers have been reported as low as 10 to 100 molecules/cell for human [10], *Escherichia coli* [11], and *Saccharomyces cerevisiae* [12], and as high as 10^{11} , 10^6 , and 10^7 molecules/cell, respectively. Low-abundance proteins are difficult to characterize not only because of the absolute low quantity, but also because the limited dynamic range of existing methods lets low abundance species be masked by more abundant ones. To identify low-abundance species, current high-resolution proteomics largely relies on mass spectrometry of purified biomarkers and/or detection of antibodies binding to known targets [8,13].

While effective, these methods rely on prior knowledge of the target protein. Such techniques must also contend with a large number of possible sequence modifications, which are themselves of experimental interest. These include, but are not limited to, introns/exons [14] and inteins/exteins [15], activation by proteolysis [16], amino acid addition to a terminal [17], mistranslation at the ribosome [18], and post-translational modifications (PTMs, i.e. phosphorylation, acetylation, or glycosylation) [19]. In any of these scenarios, a region of the protein sequence can be altered without necessarily affecting the most easily detected peptide fragments or an antibody epitope, presenting false homogeneity.

Current typical sequencing workflow

In a current typical protein sequencing workflow (Figure 1), protein samples are cleaved into peptides by different proteases [20], then the peptides are characterized by a combination of high-pressure liquid chromatography (HPLC) and mass spectrometry (MS) [20,21]. The

sequence of individual, cleaved peptides are identified by fragments from **tandem MS/MS** [22]. Overlapping peptide sequences from parallel digestions are then used to assemble the full protein sequence [23] or the peptide sequence is mapped directly onto a predicted gene [24].

In principle, MS can accurately detect molecules at attomolar concentrations [25]. However, because different peptides can have similar masses, rigorous separation is often necessary for resolution [22,26]. As a result, MS-based peptide mapping approaches typically require picomole amounts of purified proteomic samples [21,27] and have difficulty with low-abundance proteins from limited tissue samples. In genomics and transcriptomics, the problem of low abundance is solved by sample amplification with polymerase chain reactions (PCR), which can exponentially copy targeted oligonucleotides, as well as insert useful modified nucleotides [28,29] for separation and immobilization. For peptides, no analogous form of enzymatic amplification has been identified; rather, as discussed above, they must often be enriched via use of antibodies.

Proteomic efforts would be aided by new methods for sequencing heterogeneous proteome samples at less than attomole (10^{-18} moles) quantities in the context of a typical range of protein copy numbers (10^1 to 10^{10}). While complete sequencing using such methods would be ideal to replace current proteomic methods, partial sequence data (that is, sequence reads-outs limited to a subset of the natural amino acids) would still be useful for quantitation and for **protein fingerprinting**, as has been predicted computationally [30].

Sub-nanoscale Pores

After establishing an electric circuit by placing a pair of electrodes on opposing sides of a **microfluidic cell**, the cell can be partitioned such that the circuit depends on a single pore in the partition [31]. When analytes move through the pore, analyte-pore interactions alter the effective volume of the pore, thus changing the current of the circuit and allowing for modeling of the analyte volume [31]. The total number of analytes passing through the pore can be counted using the number of events in the current over time, allowing for quantification of similar but distinct species. This principle is the basis for recent commercial nanopore-based DNA sequencing devices [29]. In comparison to DNA, proteins are more challenging analytes because of their variable charge, local structure, diffusion behavior, and small side-chain volumes [32-34].

Nonetheless, early experiments showed that when a protein was denatured and impelled through an α -hemolysin pore by an **unfoldase**, consistent current changes could be assigned to known point mutations and side-chain modifications [35]. The sensitivity of these nanopores was such that they could be used to distinguish between very similar peptides via established electric fingerprints [36,37]. Hypothetically, this should allow *de novo* identification of a protein sequence from residue volume. A number of studies have shown applications of nanopores in protein analysis [31,38-40], and the theoretical and technical challenges of nanopore protein sequencing have been extensively reviewed [41,42]. Here, we highlight advances towards application in amino acid resolved protein sequencing.

In an example using silicon nitride pores, the observed current while a protein passes through a <1 nm pore under denaturing conditions was able to partially resolve amino acid sequences [43]. Pores are created by sputtering silicon nitride with an electron beam, which results in a biconical shape with a sub-nanometer region at the “waist” where two cones meet [43]. In this experimental set-up (Figure 2), a variety of proteins and polypeptides were denatured in sodium dodecyl-sulfate (SDS) and β -mercaptoethanol (BME), then electrophoresed through pores. SDS binds to proteins and imparts a uniform negative charge, causing denaturation and uniform movement under current [44]. A follow-up experiment, measuring the force of electrophoresis using **atomic force microscopy** (AFM), demonstrated that SDS was displaced when the protein entered the pore and bound again in the opposite chamber, confirming that the protein chain alone could be modeled in the pore [45].

Using this approach, the number of distinct current changes were found to correspond to the number of amino acids in the denatured protein, with reported errors that range from 5 to 10% of the sequence length. This result suggested that interaction of individual residues with the waist dictated the transduction rate; however, the current changes could only be modeled as the combined volume of five amino acid stretches, since current is affected by any change in volume of the entire biconical structure [43,46]. Additionally, the reported errors in amino acid identification suggest 70 to 90% accuracy in sequence read-out. Despite this, a comparison of the current shifts from a series of histones which differed by known point mutations showed that sensitivity to volume could resolve even 1 Å³ differences [45]. The volume of a methyl group, for context, is 25.95 Å³ [47], meaning that theoretically even small amino acids such as alanine and serine could be resolved from each other.

A major challenge with this approach is the inefficiency of protein translocation across the pore. In the reported data, a micromolar solution of analyte produced only few dozen single-molecule events over the course of hours [43]. It would take a prohibitively long time for this technique to resolve heterogeneity without prior separation. Improvement in the efficiency of translocation of peptides is therefore required and one approach to achieve this end would be covalent attachment of a short oligonucleotide to the N-terminal amine [48], which would make peptides more sensitive to electromotive forces.

Moreover, as resolving amino acids is limited by the effective change in volume of the nanopore, then higher resolution could be achieved, in principle, by using thinner substrates. By lowering the number of amino acids occupying the pore at a given step, the relative change in effective volume upon step-wise translocation becomes greater. One speculated improvement [49] is to use a molybdenum sulfate membrane (6 Å thick) instead of silicon nitride (10 Å thick). Computational modeling of such a system predicts current shifts corresponding to two-to-three residues instead of five [49]. Another proposed improvement is the attachment of tags to reactive residues to create more pronounced current changes. This approach has been used to distinguish between three peptides passing through a silicon nitride membrane, which vary by the position of a cysteine modified with a Flamma 496 dye [50].

It will also be important to make sub-nanoscale pore manufacturing easier and more consistent. It may be possible to create sub-nanometer gaps using α -hemolysin, a pore-forming protein used to create biological nanopores. Because α -hemolysin is a protein, engineered mutations could also be used to customize the pore geometry. Proteins pores have been shown to be sensitive to single amino acid differences in peptide length [51] and computational models predict that translocation time could be used to predict the volume of the amino acid occupying the most constricted region of α -hemolysin [52]. Models also predict that current could be used to measure the hydrophobicity of the amino acids occupying the pore [53].

Recognition tunneling

Quantum tunneling is the phenomenon of a subatomic particle passing through an energetically unfavorable region. An example of this is an electron passing through non-conductive material, such as oxidized metal species or organic molecules bonded to a metal surface. This is called electron tunneling, and electric currents generated by it can be used to characterize materials in the space between a conductive probe and substrate [54]. Electron tunneling spectroscopy uses current fluctuations as electrons tunnel through non-conductive molecules to model the bond vibrations of said molecules [55]. Recognition tunneling is a variation of electron tunneling spectroscopy that uses probes and substrates functionalized with small molecules to detect single-molecule binding events of the target analyte [56,57].

A commercially-available scanning tunneling microscope (STM) with a palladium probe staged over a palladium substrate, both functionalized with 4(5)substituted-1-H-imidazole-2-carboxamide (ICA) [58] via thiol chemistry, was able to differentiate amino acids by side chain and by chirality for eight different amino acids [59]. ICA was initially used to measure hydrogen bond formation with nucleotides in ssDNA [58], then later shown to be an effective recognition molecule for amino acids [59] and carbohydrates [60].

Single-molecule amino acid resolution was achieved by using probes with small defects in their insulation (determined by **transmission electron microscopy**) so that only a <10 nm surface area was functionalized with ICA [59]. A computational model supported by the results suggests that when the amino acid is bound by ICA on the probe and the substrate at the same time, features that appear in the current versus time trace correspond to thermal fluctuations of intermolecular hydrogen bonds [61], which are unique to each amino acid. Recognition tunneling produces different results for free amino acids and for those in a polypeptide [59]. Samples are identified using a characteristic profile trained from multiple scans of known amino acids, meaning that the eight published examples do not reflect an inherent limitation, but rather the time spent to build profiles by supervised machine learning.

Recognition tunneling is an appealing basis for potential *de novo* sequencing technology because of its extremely high accuracy (>99%) for amino acid identification [59], but the technique is limited in that it only works for free amino acids. It has been suggested that a recognition tunneling device could be placed downstream of an exopeptidase microreactor -- porous material containing immobilized proteases -- and that fractions of the liberated amino

acids could be flowed over the sensor (Figure 3) [59]. A population of proteins could enter the microreactor, where fractions of liberated amino acids are generated sequence-wise. Recognition tunneling could then identify unique amino acids in each fraction and provide what percent of the total population each species represents. In this way, post-translation modifications and mistranslations could be quantified. However, an exopeptidase microreactor has yet to be demonstrated. Moreover, since STM is sensitive to vibration, this raises the question of how well such a sensor device would perform under flow and what alternative designs could accommodate it.

Electron tunneling spectroscopy has been demonstrated using gold wire imbedded in an insulating inorganic multilayer, with the wire broken by a nanogap to form two electrodes [62-64]. This wire-break could potentially be functionalized with a recognition molecule, such as ICA, in the same way a larger STM device already has been modified. This represents a potential path for miniaturization of recognition tunneling, for use in a parallel high-throughput method.

Image-Based ClpXP Digestion

Many proteases target specific side-chains or motifs, making them useful for controlled protein degradation. A particular class of protease, the ATPase method. Associated with diverse cellular Activities (AAA+) exopeptidases, cleave the bond connecting a terminal amino acid to the rest of the peptide and are noteworthy for completely degrading proteins sequentially, from one terminus to the other [65]. These large, tubular complexes bind to a targeting molecule or sequence tag and pull the attached protein through themselves using mechanical transduction, driven by the hydrolysis of ATP [65]. This mechanical force is how an AAA+ complex denatures native proteins [66]. The unfolded chain moves via an internal pore from the ATPase to the proteolytic complex, where the exposed peptide bonds are cleaved by nucleophilic attack from multiple serine residues [67,68].

Identification of peptides has been demonstrated by using the ClpXP complex, a AAA+ protease from *E. coli*, reconstituted on a quartz chip using a biotin-streptavidin attachment (Figure 4) [69]. The ClpXP complex is composed of a hexameric ATPase (ClpX) and a tetradecameric protease (ClpP). In this example, a **Förster resonance energy transfer** (FRET) donor was attached to the inner wall of ClpP and FRET acceptors were linked to cysteine and lysine residues and the N-terminus of the protein. The NHS ester and maleimide chemistry is 95% efficient for on-target linkage and has no detachable off-target activity, making it dependable for the labeling of lysines and cysteines, respectively [69]. As the protein was broken down by the ClpXP complex, starting from the C-terminus, the appearance and disappearance of FRET signals indicated when an acceptor-labeled amino acid moved into the ClpP subcomplex, then was cleaved off. The number of amino acids between each signal was estimated based on the known dwell time of the protease, with errors that amount to 20-30% of the sequence length. The process of labeling the amino acids was sufficient to denature any secondary structure that might stall the ATPase activity. This was confirmed by showing that labeled titin (also known as connectin) was digested with the same efficiency as a known unfolded mutant of titin [69].

Although not *de novo*, the high frequency of lysine residues makes this technique applicable to nearly all known proteins. The read-through of a long protein by the ClpXP means that this method should be able to deconvolute heterologous protein mixtures and provide the relative abundance of distinct protein species, with no need for separation. That a labeled N-terminal residue can still pass through the protease is also advantageous because it simplifies the process for labeling the lysine residues of the peptide, as no protection of the N-terminus is required.

The main limitation to this protease-based approach is that the C-terminus must be labeled with a small stable RNA A (ssrA) peptide tag to initiate digestion. In *E. coli*, this C-terminal tag is added at the ribosome, during a stalled translation event [70], to induce degradation of the incomplete protein. Because of the chemical similarity of polypeptide terminal groups and common side chains, a protection scheme for lysine and aspartate/glutamate side-chains would be required to add this tag by established peptide bond synthesis to proteins purified from biological samples. Moreover, the proof-of-principle experiments have used substrate concentrations which would suggest the current version of the method would require nanoliter volumes to maintain less-than-femtomole samples at a nanomolar concentration. This is not necessarily a unique problem to this approach, but rather a problem best demonstrated by this method.

Image-Based Edman Degradation

Classic Edman degradation chemistry breaks the peptide bond between the N-terminal amino acid and the N-1 amino acid of a peptide without affecting the rest of the chain. In this process, phenylisothiocyanate (PITC) covalently binds the N-terminal amine [27]. When the sample is then heated under acidic, anhydrous conditions, the PITC and the terminal amino acid cyclize, breaking the peptide bond in the process and converting the second residue in the sequence into the new N-terminus [27]. Additionally, immobilizing peptides on a solid substrate for Edman degradation is an established workflow [71].

A protein sequencing method that couples immobilized Edman degradation with quantitative total internal reflectance fluorescence (TIRF) microscopy (Figure 5, scheme A) enables observation of the loss of an amino acid with a covalently added fluorescent tag [72]. This method was successfully demonstrated with fluorescently-tagged cysteine, lysine, and phosphor-serine residues [73]. As mentioned above, the chemistry for dye attachment is both efficient and specific. To increase the discriminatory power of this method, there has been further work to improve the efficiency of fluorescent tagging chemistry for aromatic and carboxylic amino acids as well [74]. In demonstrating this approach, test peptides were attached to a slide by amide bond formation at a high picomolar concentrations to achieve ideal spacing for microscopy while capturing 95% of the peptides [73]. With microliter-scale fluid handling, this technique would be appropriate for sub-femtomolar samples and has the potential for parallel reading of a heterogeneous population of short peptides, such as proteome digestion products. Thus, this technique is a way to get limited sequence features (useful for database searches and abundance determination) from short heterogeneous fragments, when given very small amounts of purified starting materials.

The precision of this method, however, is hampered by the degradation of the fluorescent tags under the conditions necessary for Edman degradation, creating the false appearance of residue cleavage [73]. The extent to which the false assignment rate might limit sample heterogeneity and throughput is unclear, and false calls from one population might overlap with the true calls from another.

An alternative to fluorescently labeling side chains would be to attach a sensitive label to the Edman degradation reagent. The information in a U.S. patent describes using a hydroxymethyl rhodamine green (HMRG) dye conjugated to an isothiocyanate (ITC) moiety as the Edman reagent [75], resulting in a HMRG tag on only the N-terminal amino acid. HMRG's emission wavelength is known to be sensitive to the proximity of hydrophobic (blue shift) and acidic amino acid side chains (red shift) [76]. Measurement of fluorescence after ITC attachment but before heating and acidification could be used, in principle, to identify the N-terminal amino acid.

Another proposed alternative is using a fluorescently-labeled N-terminal amino acid binder (NAAB), a peptide-binding protein whose k_{off} and/or k_{on} is sensitive to the N-terminal residue (Figure 5, scheme B) [77]. This approach could minimize dye degradation as a source of error, allow the method to be used in tandem with other probes, and remove the reliance on reactive residues. A mathematical model for determining the N-terminal amino acid based on a low-specificity NAAB has been suggested as part of a sequencing workflow [77]. A recent study has detailed the engineering of *Agrobacterium tumefaciens* Ndegron adaptor protein ClpS2 for use as a NAAB [78]. In addition, patents have been filed for potential methods using NAABs based on tRNA synthetases [79] and the *E. coli* ClpS protein [80].

A second major limitation of this approach is the efficiency of Edman degradation. If the PITC-peptide intermediate fails to cyclize during the elimination step, it would produce a "skip" in the sequencing read. Combined with the destruction of dyes by Edman conditions, this has been shown to produce a 10% overall chance of mis-assignment of chain position [73]. The longer the read, the higher the likelihood of a failed elimination having occurred, further skewing position assignment.

Moreover, the Edman degradation reaction is the most time-consuming part (~ 1.5 hours/ reaction) of the method [73]. A new means of catalyzing Edman degradation would therefore greatly improve the prospects of this sequencing technique. Already, an engineered "Edmanase" has been reported (as part of tRNA synthetase-based NAAB scheme) [81], but this Edmanase shows low turnover and high reagent specificity. Further improvements will clearly be needed to bring this to practical application.

Further considerations - sequencing errors and sample handling

Single-molecule experiments make use of high sensitivity to resolve small populations that would be dominated by larger populations if averaged together in a bulk measurement. But the potential for false positives and negatives means that no single-molecule measurement

can be taken as reliably true. For example, single-molecule DNA sequencing sets a threshold of individual counts a sequence must meet before being considered real [82].

Understanding the source of error is an important part of modeling this threshold, and differs between experimental methods. For error comparison, we divide protein sequencing into three broad steps: capturing molecules, assigning amino acid position, and assigning side-chain identity (Table I). Although the means of side-chain identification still need development, the greatest hurdle in all scenarios seems to be sample delivery and preparation.

As noted, many proposed single-molecule sequencing schemes rely on chemistry and molecular interactions that function in the micromolar to nanomolar range. To achieve this concentration range for attomole or less samples, the reaction volume must be kept on the microliter scale or below. Fortunately, well-established microfluidic devices (discussed below) could meet this requirement. Separation and control of small volumes will also be necessary to make use of parallelization in a similar way to commercial DNA sequencing [83]. When combined with small-scale separation techniques, parallel microfluidics allow heterogeneous, low-abundance samples to be characterized by an array of sensors.

Microfluidic technologies allow isolation of single cells and subcellular complexes [13]. Microscale western blotting [84] and isoelectric focusing [85] have been implemented for separating the contents of a single cell. The protein content of sub-cellular structures can be isolated in microliter-scale using parafilm-assisted dissection [86], liquid microdrops [87], and absorbent hydrogels [88]. On-chip mixing and heating also makes it possible to carry about chemical reactions at the microfluidic scale [89]. Similarly, microfluidic HPLC allows for separation of peptides and reactive species, making it realistic to employ chemical modifications that require high concentrations as part of a sequencing methodology [90].

Concluding remarks

While a next-generation protein-sequencing platform has yet to emerge, a number of technologies are actively being pursued to achieve this goal. In this review, the four single-molecule approaches highlighted are all currently limited in what subset of natural amino acids can be resolved, as would be needed in a true *de novo* protein sequencing method. Nonetheless, each method could still be used in tandem with genomic and transcriptomic information for protein identification and quantitation.

Looking ahead towards realizing the full potential of these platforms for protein sequencing, there remains broad unmet component requirements in nanofabrication methods, protein engineering, side-chain tagging schemes, dye design, and degradation chemistry (see Outstanding Questions). As these technical hurdles cover such a wide array of disciplines, it would be difficult for a single effort to address them all. Instead, advances in protein engineering, synthetic chemistry, nanofabrication, and microfluidics must converge to generate working solutions.

Despite the formidable technical challenges that still lay ahead, there is no doubt as to the importance of generating tools for next-generation protein sequencing and the impact single-

molecule proteomics could have in addressing key questions about cellular signaling, cellular organization, drug targets and ultimate cell fate. At this point, it is premature to predict if any method has a decided advantage over the others or if each will find unique applications. Thus, while it is tempting to think of next-generation peptide sequencing as an all-or-nothing endeavor, this is a time when even modest advances could have great utility and wide implementation.

Acknowledgements

We would like to thank the editor and anonymous reviewers for their constructive suggestions.

Glossary

Atomic force microscopy (AFM)

technique that measures the force exerted on a nanoscale tip by transduction through a cantilever. This allows for the detection of nanoscale surface features and of forces exerted on single molecules.

Förster resonance energy transfer (FRET)

through-space transfer of excitation from a donor fluorophore to an acceptor fluorophore. The acceptor fluorescence can be used to calculate the distance between the two dyes.

Genomics

detection and quantification of genes

Metabolomics

detection and quantification of small molecules involved in cellular metabolism

Microfluidics

manipulating and controlling fluids at the scale of microliters and less

Protease/peptidase

enzyme that catalyzes the breaking of peptide bonds

Protein/peptide fingerprinting

using a database of characterized peptides to identify proteins from physical and/or partial sequence information.

Proteomics

detection and quantification of cellular protein

Synthetic biology

engineering cellular functions to overproduce target molecules

Tandem MS/MS

technique that first separates ions by mass over charge, then fragments ions by collision and detects the fragmentation products by mass over charge. This is commonly used to calculate the amino acid composition of peptides

Transcriptomics

detection and quantification of cellular RNA

Transmission electron microscopy

technique that uses the penetration of a beam of electrons through an object to visualize its density

Unfoldase

enzyme that releases energy from ATP to unfold proteins

References

1. D'Alessandro A and Zolla L (2013) Meat science: From proteomics to integrated omics towards system biology. *J. Proteomics* 78, 558–577 [PubMed: 23137709]
2. Fukushima A et al. (2009) Integrated omics approaches in plant systems biology. *Curr. Opin. Chem. Biol* 13, 532–538 [PubMed: 19837627]
3. Zhang W et al. (2010) Integrating multiple “omics” analysis for microbial biology: Application and methodologies. *Microbiology* 156, 287–301 [PubMed: 19910409]
4. Palsson B (2002) In silico biology through “omics.” *Nat. Biotechnol* 20, 649–650 [PubMed: 12089538]
5. Gawad C et al. (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet* 17, 175–88 [PubMed: 26806412]
6. Hrdlickova R et al. (2017) RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA* 8, e1364
7. Bray NL et al. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol* 34, 525–7 [PubMed: 27043002]
8. Omenn GS (2014) The strategy, organization, and progress of the HUPO Human Proteome Project. *J. Proteomics* 100, 3–7 [PubMed: 24145142]
9. Wilhelm M et al. (2014) Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587 [PubMed: 24870543]
10. Kopylov AT et al. (2016) The size of the human proteome: the width and depth. *Int. J. Anal. Chem* 2016, 1–6
11. Schmidt A et al. (2016) The quantitative and condition-dependent *Escherichia coli* proteome. *Nat. Biotechnol* 34, 104–10 [PubMed: 26641532]
12. Ho B et al. (2018) Unification of protein abundance datasets yields a quantitative *saccharomyces cerevisiae* proteome. *Cell Syst.* 6, 192–205.e3 [PubMed: 29361465]
13. Liu Y et al. (2019) Advancing single-cell proteomics and metabolomics with microfluidic technologies. *Analyst* 144, 846–858 [PubMed: 30351310]
14. Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* 72, 291–336 [PubMed: 12626338]
15. Paulus H (2000) Protein splicing and related forms of protein autoprocesing. *Annu. Rev. Biochem.* 69, 447–496 [PubMed: 10966466]
16. Salvesen GS and Dixit VM (1997) Caspases: intracellular signaling by proteolysis. *Cell* 91, 443–446 [PubMed: 9390553]
17. Lai ZW et al. (2015) Protein amino-terminal modifications and proteomic approaches for N-terminal profiling. *Curr. Opin. Chem. Biol* 24, 71–79 [PubMed: 25461725]
18. Ribas de Pouplana L et al. (2014) Protein mistranslation: Friend or foe? *Trends Biochem. Sci* 39, 355–362 [PubMed: 25023410]
19. Witze ES et al. (2007) Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* 4, 798–806 [PubMed: 17901869]
20. Steen H and Mann M (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol* 5, 699–711 [PubMed: 15340378]

21. Scheffler K et al. (2018) High resolution top-down experimental strategies on the Orbitrap platform. *J. Proteomics* 175, 42–55 [PubMed: 28385662]
22. Medzihradszky KF and Chalkley RJ (2015) Lessons in *de novo* peptide sequencing by tandem mass spectrometry. *Mass Spectrom. Rev* 34, 43–63 [PubMed: 25667941]
23. Sinitcyn P et al. (2018) Computational methods for understanding mass spectrometry-based shotgun proteomics data. *Annu. Rev. Biomed. Data Sci* 1, 207–234
24. Muth T et al. (2018) A Potential Golden Age to Come—current tools, recent use cases, and future avenues for *de novo* sequencing in proteomics. *Proteomics* 18, 1–14
25. Tsedilin AM et al. (2015) How sensitive and accurate are routine NMR and MS measurements? *Mendeleev Commun.* 25, 454–456
26. Lesur A and Domon B (2015) Advances in high-resolution accurate mass spectrometry application to targeted proteomics. *Proteomics* 15, 880–890 [PubMed: 25546610]
27. Hunkapiller MW and Hood LE (1983) Protein Sequence Analysis : Automated Microsequencing. *Science* (80-.). 219, 650–654
28. Vanguilder HD et al. (2008) Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* 44, 619–626 [PubMed: 18474036]
29. Mardis ER (2017) DNA sequencing technologies : 2006 – 2016. *Nat. Protoc* 12, 213–218 [PubMed: 28055035]
30. Yao Y et al. (2015) Single-molecule protein sequencing through fingerprinting: Computational assessment. *Phys. Biol* 12, 055003 [PubMed: 26266455]
31. Dekker C (2007) Solid-state nanopores. *Nat. Nanotechnol* 2, 209–215 [PubMed: 18654264]
32. Restrepo-Pérez L et al. (2017) SDS-assisted protein transport through solid-state nanopores. *Nanoscale* 9, 11685–11693 [PubMed: 28776058]
33. Rodriguez-Larrea D and Bayley H (2013) Multistep protein unfolding during nanopore translocation. *Nat. Nanotechnol* 8, 288–295 [PubMed: 23474543]
34. Plesa C et al. (2013) Fast Translocation of Proteins through Solid State Nanopores. *Nano Lett.* 13, 658–663 [PubMed: 23343345]
35. Nivala J et al. (2013) Unfoldase-mediated protein translocation through an α -hemolysin nanopore. *Nat. Biotechnol.* 31, 247–250 [PubMed: 23376966]
36. Rosen CB et al. (2014) Single-molecule site-specific detection of protein phosphorylation with a nanopore. *Nat. Biotechnol* 32, 179–181 [PubMed: 24441471]
37. Sampath G (2019) Protein fingerprinting with digital sequences of linear protein subsequence volumes: a computational study. *J. Biosci* 44, 1–11 [PubMed: 30837353]
38. Oukhaled A et al. (2012) Sensing proteins through nanopores: Fundamental to applications. *ACS Chem. Biol* 7, 1935–1949 [PubMed: 23145870]
39. Ma L and Cockroft SL (2010) Biological nanopores for single-molecule biophysics. *ChemBioChem* 11, 25–34 [PubMed: 19938028]
40. Varongchayakul N et al. (2018) Single-molecule protein sensing in a nanopore: a tutorial. *Chem. Soc. Rev* 47, 8512–8524 [PubMed: 30328860]
41. Restrepo-Pérez L et al. (2018) Paving the way to single-molecule protein sequencing. *Nat. Nanotechnol* 13, 786–796 [PubMed: 30190617]
42. Chinappi M and Cecconi F (2018) Protein sequencing via nanopore based devices: A nanofluidics perspective. *J. Phys. Condens. Matter* 30, 204002 [PubMed: 29595524]
43. Kennedy E et al. (2016) Reading the primary structure of a protein with 0.07 nm 3 resolution using a subnanometre-diameter pore. *Nat. Nanotechnol* 11, 968–976 [PubMed: 27454878]
44. Gallagher SR (2012) One-dimensional SDS gel electrophoresis of proteins. *Curr. Protoc. Protein Sci* 75, 10–12
45. Dong Z et al. (2017) Discriminating Residue Substitutions in a Single Protein Molecule Using a Sub-nanopore. *ACS Nano* 11, 5440–5452 [PubMed: 28538092]
46. Kolmogorov M et al. (2017) Single-molecule protein identification by subnanopore sensors. *PLoS Comput. Biol* 13, e1005356 [PubMed: 28486472]
47. Richards FM (1974) The interpretation of protein structures: Total volume, group volume distributions and packing density. *J. Mol. Biol* 82, 1–14 [PubMed: 4818482]

48. Biswas S et al. (2015) Click addition of a DNA thread to the N-Termini of peptides for their translocation through solid-state nanopores. *ACS Nano* 9, 9652–9664 [PubMed: 26364915]
49. Chen H et al. (2018) Protein translocation through a MoS₂ nanopore: a molecular dynamics study. *J. Phys. Chem. C* 122, 2070–2080
50. Yu JS et al. (2019) Differentiation of selectively labeled peptides using solid-state nanopores. *Nanoscale* 11, 2510–2520 [PubMed: 30672547]
51. Piguet F et al. (2018) Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. *Nat. Commun* 9, 966 [PubMed: 29511176]
52. Asandei A et al. (2017) Protein Nanopore-Based Discrimination between Selected Neutral Amino Acids from Polypeptides. *Langmuir* 33, 14451–14459 [PubMed: 29178796]
53. Asandei A et al. (2018) Single-molecule dynamics and discrimination between hydrophilic and hydrophobic amino acids in peptides, through controllable, stepwise translocation across nanopores. *Polymers (Basel)*. 10, 885
54. Della Pia A and Costantini G (2013) Scanning tunneling microscopy In *Springer Series in Surface Sciences* (Gianangelo Bracco BH, ed), Springer-Verlag
55. Wolf EL (2012) *Principles of Electron Tunneling Spectroscopy*, Oxford University Press.
56. Lindsay S et al. (2010) Recognition tunneling. *Nanotechnology* 21, 262001 [PubMed: 20522930]
57. Huang S et al. (2010) Identifying single bases in a DNA oligomer with electron tunnelling. *Nat. Nanotechnol* 5, 868–873 [PubMed: 21076404]
58. Liang F et al. (2012) Synthesis, physicochemical properties, and hydrogen bonding of 4(5)-substituted 1-H-imidazole-2-carboxamide, a potential universal reader for DNA sequencing by recognition tunneling. *Chemistry (Easton)*. 18, 5998–6007
59. Zhao Y et al. (2014) Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol* 9, 466–473 [PubMed: 24705512]
60. Im JO et al. (2016) Electronic single-molecule identification of carbohydrate isomers by recognition tunnelling. *Nat. Commun* 7, 13868 [PubMed: 28000682]
61. Krsti P et al. (2015) Physical model for recognition tunneling. *Nanotechnology* 26, 084001 [PubMed: 25650375]
62. Ohshiro T et al. (2014) Detection of post-translational modifications in single peptides using electron tunnelling currents. *Nat. Nanotechnol* 9, 835–840 [PubMed: 25218325]
63. Tsutsui M et al. (2011) Single-molecule sensing electrode embedded inplane nanopore. *Sci. Rep* 1, 46 [PubMed: 22355565]
64. Morikawa T et al. (2017) Fast and low-noise tunnelling current measurements for single-molecule detection in an electrolyte solution using insulator-protected nanoelectrodes. *Nanoscale* 9, 4076–4081 [PubMed: 28116385]
65. Aubin-Tam ME et al. (2011) Single-molecule protein unfolding and translocation by an ATP-fueled proteolytic machine. *Cell* 145, 257–267 [PubMed: 21496645]
66. Hanson PI and Whiteheart SW (2005) AAA+ proteins: Have engine, will work. *Nat. Rev. Mol. Cell Biol* 6, 519–529 [PubMed: 16072036]
67. Maurizi MR et al. (1990) Sequence and structure of Clp P, the proteolytic component of the ATP-dependent Clp protease of *Escherichia coli*. *J. Biol. Chem* 265, 12536–12545 [PubMed: 2197275]
68. Thompson MW et al. (1994) Processive degradation of proteins by the ATP-dependent Clp protease from *Escherichia coli*: Requirement for the multiple array of active sites in ClpP but not ATP hydrolysis. *J. Biol. Chem* 269, 18209–18215 [PubMed: 8027082]
69. van Ginkel J et al. (2018) Single-molecule peptide fingerprinting. *Proc. Natl. Acad. Sci* 115, 3338–3343 [PubMed: 29531063]
70. Karzai AW et al. (2000) The SsrA-SmpB system for protein tagging, directed degradation and ribosome rescue. *Nat. Struct. Biol* 7, 449–455 [PubMed: 10881189]
71. L'Italien JJ and Strickler JE (1982) Application of high-performance liquid chromatographic peptide purification to protein microsequencing by solid-phase Edman degradation. *Anal. Biochem* 127, 198–212 [PubMed: 7165088]
72. Swaminathan J et al. (2015) A Theoretical Justification for Single Molecule Peptide Sequencing. *PLoS Comput. Biol* 11, e1004080 [PubMed: 25714988]

73. Swaminathan J et al. (2018) Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol* 36, 1076–1082
74. Hernandez ET et al. (2017) 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
75. Emili A (2018) Patent App: US 2018/0299460 A1.
76. Iwatate RJ et al. (2016) Asymmetric rhodamine-based fluorescent probe for multicolour *in vivo* imaging. *Chemistry (Easton)*. 22, 1696–1703
77. Rodrigues S et al. (2019) A theoretical analysis of single molecule protein sequencing via weak binding spectra. *bioRxiv* 14, e0212868
78. Tullman J et al. (2019) Engineering ClpS for selective and enhanced N-terminal amino acid binding. *Appl. Microbiol. Biotechnol* 103, 2621–2633 [PubMed: 30675637]
79. Havranek JJ and Borgo B (2017) Patent App: US 2017/0052194 A1.
80. Emili A et al. (2017) Patent: US 9,566,335 B1,
81. Borgo B and Havranek JJ (2015) Computer-aided design of a catalyst for Edman degradation utilizing substrate-assisted catalysis. *Protein Sci.* 24, 571–579
82. Georgiou G et al. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol* 32, 158–168 [PubMed: 24441474]
83. Lu H et al. (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics Bioinforma.* 14, 265–279
84. Hughes AJ et al. (2014) Single-cell western blotting. *Nat. Methods* 11, 749–755 [PubMed: 24880876]
85. Tentori AM et al. (2016) Detection of Isoforms Differing by a Single Charge Unit in Individual Cells. *Angew. Chemie - Int. Ed* 55, 12431–12435
86. Quanico J et al. (2015) Parafilm-assisted microdissection: A sampling method for mass spectrometry-based identification of differentially expressed prostate cancer protein biomarkers. *Chem. Commun* 51, 4513–4722
87. Wisztorski M et al. (2016) Spatially-resolved protein surface microsampling from tissue sections using liquid extraction surface analysis. *Proteomics* 16, 1622–1632 [PubMed: 26929135]
88. Rizzo DG et al. (2017) Enhanced Spatially Resolved Proteomics Using On-Tissue Hydrogel-Mediated Protein Digestion. *Anal. Chem* 89, 2948–2955 [PubMed: 28193007]
89. Elvira KS et al. (2013) The past, present and potential for microfluidic reactor technology in chemical synthesis. *Nat. Chem* 5, 905–915 [PubMed: 24153367]
90. Lazar IM et al. (2006) Microfluidic liquid chromatography system for proteomic applications and biomarker screening. *Anal. Chem* 78, 5513–5524 [PubMed: 16878890]

Outstanding Questions

- Are there chemical or biochemical means of constructing subnanometer pores? How reliable and uniform is the creation of subnanometer pores?
- How could electron tunneling recognition be placed in-line with a protease? Are current hypothetical models practical to build?
- A tag on the N- or C-terminus is necessary to recruit sample peptides to proteases and surfaces. Are there tagging schemes which do not rely on synthetic protecting groups to target a terminus? Are there enzymes which can be exploited to controllably target peptide termini?
- Are there more efficient alternatives to Edman degradation? Are there alternatives that would be less destructive to fluorescent dyes? Can Edman degradation be catalyzed?
- Will NAAB proteins work as predicted in a modified Edman scheme? What is the practical limit of NAAB affinity? Are there yet unconsidered NAAB reagents?
- Can these sequencing techniques be used to accurately detect low abundant proteins in a heterogeneous mixture? Can these sequencing techniques provide information on quantitation of proteins? What is the tolerance for error in useful quantitation?

Highlights

- Advancements in high-throughput technologies has enabled rapid and parallel sequencing of genomes and transcriptomes.
- High-resolution sequencing of attomole amounts of protein would revolutionize characterization of gene products (the proteome) and sub-cellular structures.
- Advancements in nanopore sensors and single-molecule fluorescence methods can now detect specific residues on a peptide, but not yet all twenty.
- Controlling peptide location and orientation is a major technical hurdle in the development of next-generation protein sequencing modalities.
- Present work in next-generation protein sequencing suggests future compatibility with multiplexing and microfluidic sampling.

Figure 1.

Current protein sequencing paradigm. After a protein of interest (grey) is purified, separated samples are digested with different proteases to yield a collection of peptides (Step 1). The peptides are then identified using a combination of HPLC (Step 2) and mass spectrometry (Step 3). The sequences of the digestion products are then used to computationally assemble the full-length protein sequence (Step 4).

Figure 2.

Proposed sub-nanogap sequencing. The protein sample (grey) is denatured in SDS (green circles; Step 1) and injected into a microfluidic cell for electrophoresis (Step 2, right). The current drives the denatured protein (grey squares with single letter amino acid abbreviations and N- and C-termini labeled) through a biconical pore structure (blue) that is 10 Å thick (Step 2, left). Each residue in the protein chain transiently interacts with the “waist” of the biconical pore, creating a unique step in the current over time (Step 3). The magnitude of each step is determined by the combined volume of amino acids in the pore.

Figure 3.

Proposed use of recognition tunneling in sequencing. Proteins (grey) are digested sequentially by either chemical degradation or a peptidase (Step 1), and cleaved residues are collected by flow (Step 2). Each fraction is then analyzed by recognition tunneling spectroscopy (Step 3). In this process, the free amino acids (grey squares) pass through a palladium-plated probe and a substrate, interacting with the 4(5) substituted-1-H-imidazole-2-carboxamide (ICA, green sphere) functional groups on the probe and substrate. The binding event generates a unique current trace from the interaction of the amino acid with the ICA; fractions are described by collections of current traces (Step 4). The sub-populations of current traces in each fraction allow quantification of residue-level sub-populations in the protein sample.

Figure 4.

Proposed image-based ClpXP sequencing. The purified protein sample (grey) is labeled with FRET acceptors (orange) at the N-terminal amine and at lysine and cysteine residues (Step 1), and a ClpX initiation tag is added to the C-terminus. Labeled proteins are flowed over a slide with donor-tagged ClpXP (green and blue), attached to the slide by a biotin-streptavidin bond (Step 2, right). When the acceptor on the lysine or the N-terminus enters ClpP, a FRET signal is observed due to the acceptor nearing the donor (green sphere; Step 2, left). After the protein is digested, the total time between signals is used to estimate the chain length and protein sequence (Step 3).

Figure 5.

Proposed image-based Edman sequencing. (Scheme A) A protein sample (grey) is digested (Step 1), then labeled with fluorescent tags (red and green spheres, respectively) at lysine and cysteine residues (Step 2). The labeled peptides are immobilized on an amine-functionalized slide by a peptide bond (Step 3). The fluorescent signal from each peptide is quantified prior to a round of Edman degradation. A step down in quantified signal indicates the elimination of a lysine or cysteine (Step 4). (Scheme B) In an alternative protocol, peptides are not directly labeled before attachment to the slide (Steps 1-2). Instead, labeled N-terminal amino-acid binding (NAAAB) proteins are used to identify the N-terminal amino acid between each round of degradation (Step 3). Please expand on what the colors may mean here.

Table 1:

Accuracy/efficiency of each step of the sequencing methods

Method	Molecular Capture	Position Assignment	Side-chain assignment	Ref.
Sub-nanogap sensor	Electrophoresis along constrained path: <i>Not studied</i>	Estimated from the number of steps in the current trace: <i>90 to 95% accuracy</i>	Prediction of current blockade from modeling of 5-mer: <i>70 to 90% accuracy</i>	43
Recognition Tunneling	Undeveloped	Undeveloped	Computational pattern recognition of current trace: <i>>99% accuracy</i>	59
Image-Based ClpXP Digestion	Undeveloped	Estimation of chain length from dwell time of bound protein: <i>70 to 80% efficiency</i>	Attachment of fluorescent dyes to Cysteine and Lysine: <i>95% efficiency</i>	69
Image-Based Edman Degradation	Reaction of carboxylic acid groups with amide on surface: <i>95% efficiency</i>	Elimination of fluorescence by cleavage of residue from chain: <i>90% efficiency</i>		73