A beginner's guide to mass spectrometry–based proteomics

Ankit Sinha (Max Planck Institute of Biochemistry, Germany)
Matthias Mann (Max Planck Institute of Biochemistry, Germany and Novo Nordisk Foundation Center for Protein Research, Denmark)

Mass spectrometry (MS)-based proteomics is the most comprehensive approach for the quantitative profiling of proteins, their interactions and modifications. It is a challenging topic as a firm grasp requires expertise in biochemistry for sample preparation, analytical chemistry for instrumentation and computational biology for data analysis. In this short guide, we highlight the various components of a mass spectrometer, the sample preparation process for conversion of proteins into peptides, and quantification and analysis strategies. The advancing technology of MS-based proteomics now opens up opportunities in clinical applications and single-cell analysis.

Genes are the unit of heredity, but they only come to life when they are translated to proteins – the primary functional actors in biology. They perform an incredible range of functions, from biochemical reactions, signalling and transport to structural support. The proteome is the collection of proteins present in biofluids, cells and tissues and reflects the functional state of the biological system. Proteomics is the quantitative study of the proteome and is often used for contrasting different cellular conditions. As a contemporary example, proteomic differences between virus-infected and uninfected cells would highlight cellular pathways and proteins needed for viral infection and replication. Drugs developed to target these proteins could slow down the infection. Proteomics is well-suited for unravelling the underlying biochemical mechanisms in an unbiased way as it directly characterizes all proteins at once. Here, we focus on the system-wide characterization of the proteome using mass spectrometry (MS), and more specifically on bottom-up proteomics where proteins are digested into smaller pieces called peptides, which are analysed by MS.

The basics of mass spectrometry

Since their inception in 1912, mass spectrometers have undergone continuous development, and these sophisticated bioanalytical instruments have now reached unrivalled detection limits, speed and diversity in applications. They detect the presence and abundance of peptides (or other biomolecules such as metabolites, lipids and proteins) using fundamental properties of molecules, such as mass, and net charge. When peptides obtain a net charge (usually through gain of protons), they are referred to as peptide ions.

All mass spectrometers have three fundamental components: an ion source, mass analyser and detector (Figure 1A). As mass spectrometers can only analyse gaseous ions, methods such as electrospray ionization (ESI) are needed to convert peptides from the liquid phase to gaseous ions. The liquid containing the peptides is pumped through a micrometre-sized orifice held at a high voltage (2–4 kV). Upon reaching this emitter, the steady stream of liquid disintegrates into extremely small, highly charged and rapidly evaporating charged droplets, leaving peptide ions in the gas phase. Even 20 years after John Fenn received the Nobel Prize for this discovery, the exact mechanisms are not completely understood. We know that the abundance of gaseous peptide ions is proportional to their original concentration, making it beneficial to use the lowest flow rates possible, thereby maximizing sensitivity. It is common in proteomics to separate peptide mixtures using high-performance liquid chromatography (HPLC) systems with flow rates of only a few hundred nanolitres per minute rather than millilitres in conventional HPLC.

The principal role of a mass analyser is to separate ions by their mass-to-charge ratios (m/z). Fundamentally, all ions are separated by modulating their trajectories in electrical fields. Mass analysers differ in the principle they use for separating ions, and this defines their preferred application areas. Quadrupoles, usually combined with time-of-flight (TOF) or Orbitrap analysers, are the most common in proteomics. Quadrupole mass analysers separate ions using an oscillating electrical field between four cylindrical rods in a parallel arrangement, where each pair of rods produces a radio frequency electrical field with a phase offset. The resulting electrical fields define a pseudo-potential surface that is configured to allow the transmission of all ions, or to selectively transmit ions of a specific m/z window.

TOF mass analysers separate ions based on the differences in velocities after acceleration to about 20 kV and subsequent different arrival times at the detector. A TOF can measure mass differences of one part per million (ppm) by detecting time differences of sub-microseconds. In contrast, the Orbitrap mass analyser distinguishes ions based on their oscillation frequencies. Ions are tangentially injected and then trapped in the Orbitrap, and they move along the length axis of a central metal spindle (Figure 1B). Although an

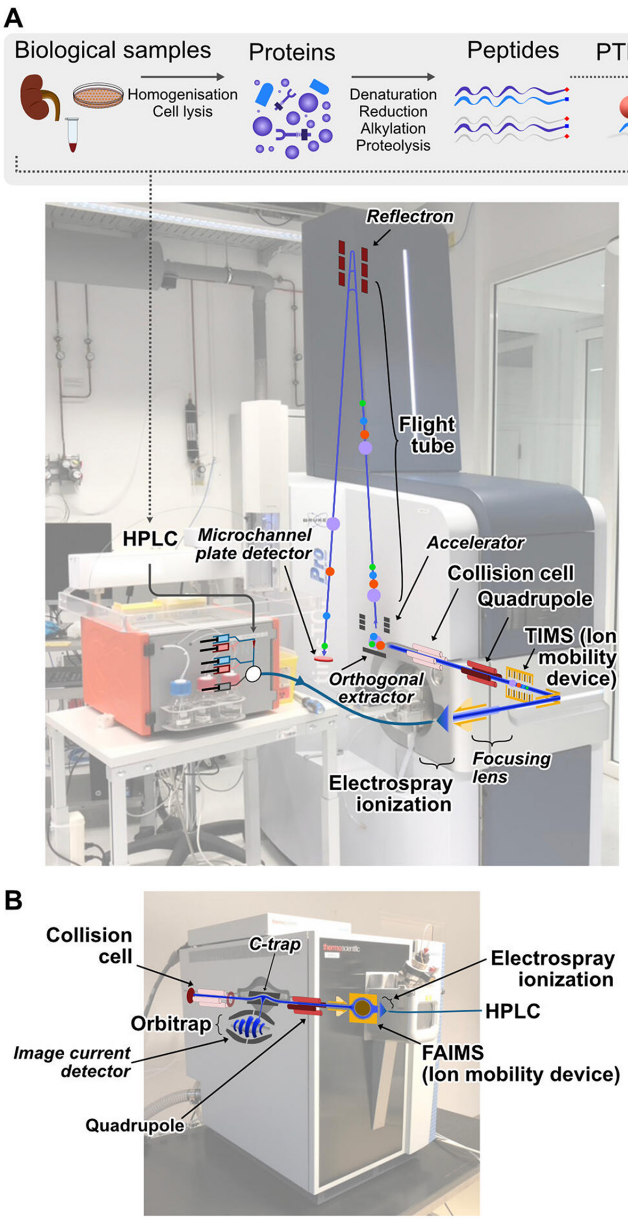

Figure 1. Overview of sample preparation and instrumentation used in MS-based proteomics. **(A)** Proteins are digested into peptides using sequence-specific proteases. Optionally, post-translational modification (PTM)-containing peptides can be enriched using beads with specific surface chemistry or coupled antibodies. High-performance liquid chromatography (HPLC) separates peptides based on hydrophobicity, and they are subsequently analysed by a TOF mass spectrometer. **(B)** Alternatively, peptides can be analysed by an Orbitrap mass spectrometer, which is a mainstream instrument in proteomics.

Orbitrap is only a few centimetres long, the ions can rapidly travel up to several kilometres, enabling very high resolution (typically tens of thousands) and low ppm mass accuracy.

In proteomics, the quadrupole element is normally followed by a 'collision cell', which is a quadrupole where the ions can be fragmented. Either intact peptide ions or fragment ions enter the final stage that also contains the detector – the resulting spectra are called MS^1 or precursor ion spectra in the former case and MS^2 or product or MS/MS spectra in the latter. TOF instruments have microchannel plate (MCP) detectors, where each individual ion ejects electrons from a surface that are then amplified. Individual ions can be readily measured with MCPs, but this exquisite sensitivity comes with the caveat that the detector can easily saturate in case of high signals. In Orbitrap analysers, the 'image current' induced by the rapidly oscillating ions is measured, and it represents a quantitative readout of the strength of the individual ion packages. The current is recorded in the time domain and is converted into the frequency domain using Fourier transformation. Advances

in signal processing algorithms have repeatedly doubled the achievable resolution with a given transient time of the signal, but these are still orders of magnitude slower than those of TOF analysers (tens to hundreds of milliseconds vs typically 100 microseconds for a single TOF pulse).

How do the MS instruments sequence or identify peptides? Precursor ions with a specific m/z are first isolated by the quadrupole and fragmented through collision with inert gases such as N_2, He or Ar. This causes them to break apart at the lowest energy bonds – typically, some of the amide bonds (peptide bonds) connecting the amino acid residues – and leaves MS/MS spectra with incomplete ladders of peaks differing by amino acid masses. This information is incredibly specific and is used for identification of the peptide sequence. A sequence of just a few amino acids and the flanking masses – a peptide sequence tag – is sufficient for identifying a peptide from the entirety of human proteome. More usually, database identification involves generating all possible fragmentation spectra and then statistically scoring them against the experimental spectra.

The chromatographic retention time is an important level of information when matching a dataset against a previous measurement and is key to 'targeted proteomics' technologies. Furthermore, ion mobility analysis, an additional dimension of separation of peptide ions, has recently become mainstream. Ions are either filtered based on their cross-section (FAIMS, field asymmetric ion mobility spectrometry) or actually separated during their analysis (T-Wave or TIMS, trapped ion mobility spectrometry). TIMS is the basis of the parallel accumulation–serial fragmentation (PASEF) technology, which multiplies sequencing speed 10-fold while improving sensitivity.

Sample preparation and specific enrichment

MS-based proteomics can analyse the protein content of any material. Apart from mainstream sources such as cell lines, this encompasses clinically important, archived formalin-fixed paraffin-embedded (FFPE) biopsy tissues and even fossils that are hundreds of thousands of years old. This is because proteins are very durable biomolecules, especially compared to RNA. Often, proteins are isolated after a biochemical enrichment procedure appropriate to the question at hand, such as cellular fractionation, affinity enrichment or proximity assays.

Proteomic sample preparation is challenging and can be considered an art as much as a science. The overall aim of sample preparation is the controlled digestion of proteins into peptides (Figure 1A). To this end, extracted and solubilized proteins are digested with a sequence-specific protease. Trypsin is a favourite because it specifically cleaves C-terminally to arginines and lysines, thereby leaving a positively charged amino acid at the newly created C-termini, favouring ionization and fragmentation. Throughout the sample preparation procedure, polymers and detergents must be avoided as they outcompete ionization of peptide ions. The sample preparation ends with hundreds of thousands of purified peptides produced from tens of thousands of proteins, with a million-fold concentration differences or more.

Monitoring post-translational modifications

More often than not, proteins contain modifications to their primary sequence, and these post-translational modifications (PTMs) are efficient and elegant control mechanisms to change the activity or even function of proteins. Charting the extent, nature and temporal progression of PTMs has arguably been the most substantial contribution of MS-based proteomics to biology. These modifications are generally sub-stoichiometric – meaning that only a fraction of the given protein is modified – and hence, are challenging to capture and detect. Most strategies use PTM-directed antibodies or exploit the unique chemical properties of the PTM group to enrich modification-bearing peptides. Phosphorylation is the most studied PTM, and titanium dioxide–based beads are frequently used for enriching phosphopeptides with a high specificity. Remarkably, standard workflows can detect more than 10,000 sites with single amino acid resolution and a broad array of intracellular signalling networks in a single 2-hour experiment, an achievement totally unthinkable before the advent of MS-based proteomics. Today, proteomics is routinely used for unravelling the role of ubiquitination, sumoylation, acetylation and glycosylation in biological processes. However, analysis of less-common PTMs – especially those without highly specific antibodies – still remains challenging.

Data acquisition and quantification strategies

At any given time during an MS acquisition, hundreds of peptides are ionized, and they simultaneously enter the mass spectrometer. Until recently, they had been analysed invariably by data-dependent acquisition (DDA), meaning that the mass spectrometer follows a set of user-defined rules (such as m/z, charge, intensity and cross-section) to select as many peptides as possible for acquiring MS/MS spectra (Figure 2A). However, this selection is partly stochastic as there are more peptides than analysis time, and it generates missing values. In data-independent acquisition (DIA) methods, the quadrupole instead continuously cycles across the entire mass range while selecting relatively large m/z values (20–40 m/z)

(Figure 2B). This leads to very complex MS/MS spectra since they contain the superimposed fragmentation patterns from co-isolated peptide ions. Modern software can deconvolute the spectra to identify the multiple peptides, usually by comparison to a previously acquired 'peptide library', but increasingly also without. Novel scan modes are still being developed to address the 'dynamic range problem': the challenge of detecting very low abundance proteins in the presence of much more abundant ones. For example, the cytokines in blood can be 12 orders of magnitude less abundant than albumin.

Quantification strategies for peptides can be divided into two broad classes, label free and label based. In label-free quantification (LFQ), the MS signals of the peptides (usually at the MS^1 level) are extracted from the raw data, normalised and compared between the proteomic conditions of interest. It is experimentally the most straightforward and usually the most economical approach, providing great flexibility in project design. However, this strategy has higher quantification variance, and differences in peptide purity and instrument performance may impact comparisons between individual samples if sufficient care is not taken.

Label-based approaches use stable isotopes to encode different proteome states – the beauty is that the resulting peptides have exactly the same physiochemical behaviour, but have predictable differences in mass. The isotopes can be metabolically introduced, which also allows determination of protein turnover; however, chemical labelling with 'readout' at the MS/MS level is now much more common. The latter is referred to as isobaric labelling and involves a clever trick in which the mass of the tag remains the same, but the distribution of isotopes in the tag is revealed after fragmentation. Within a set of 6–16 different tags, quantification variance is typically lower than in LFQ if samples are consistently and reproducibly labelled and combined. A major caveat of the typically

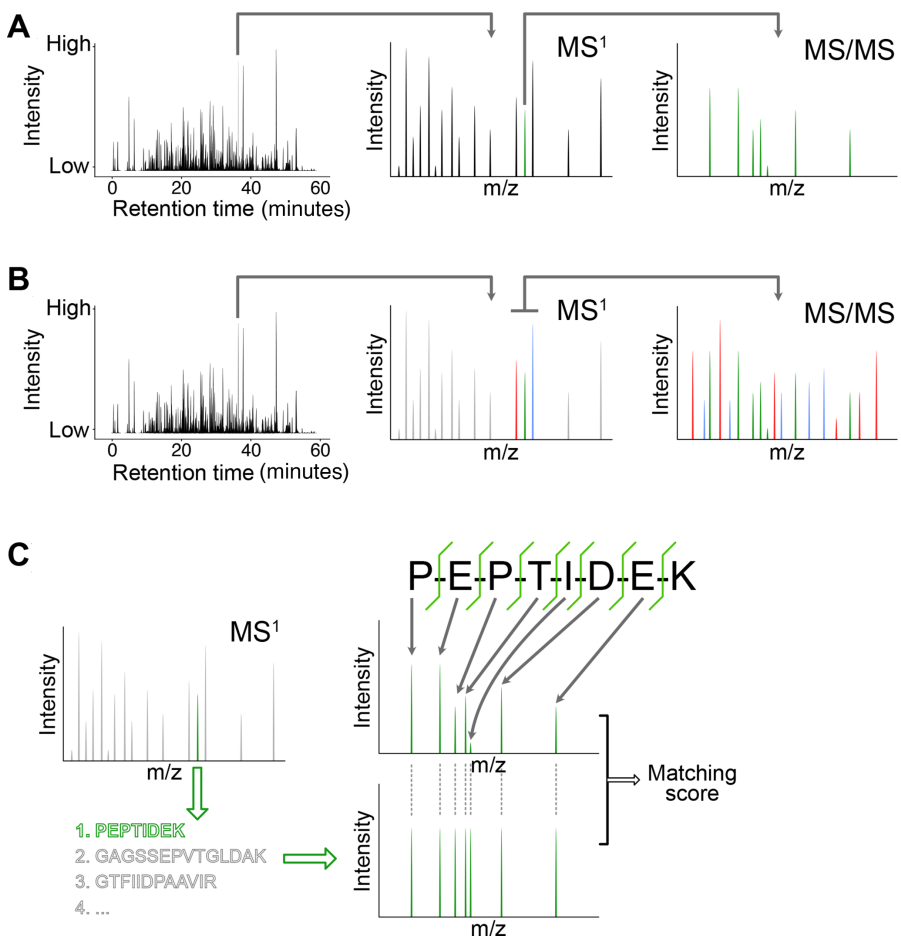

Figure 2. DDA and DIA are the common data acquisition strategies in shotgun proteomics. **(A)** In DDA, a peptide ion is selected from many ions available in the MS^1 data at a given retention (chromatography) time. The peptide ion is fragmented, and the data are recorded as MS/MS. **(B)** In DIA, multiple peptide ions are selected based on m/z window at a given retention (chromatography) time. The peptide ions are fragmented, and they inherently produce a complex MS/MS spectra. **(C)** Peptide sequence assignment involves *in silico* shortlisting of peptides and subsequent construction of their *in silico* MS/MS spectra. The fragment ion information is transferred from the best matching *in silico* MS/MS.

used TMT (tandem mass tag) isobaric labelling method is that co-fragmented peptides can suppress quantitative differences ('ratio compression').

Regardless of the quantification and scan modes, the output from mass spectrometers always contains MS[1] and MS/MS spectra. A multitude of open or closed software programs handle the processing of these data from finding the signals ('feature finding') to search engines that match the MS/MS spectra to peptide sequences in the database, algorithms to assemble the peptides back to proteins ('protein inference problem') and finally the quantification at peptide or protein levels (Figure 2C).

At the simplest, the output is a matrix with a list of proteins and their corresponding abundances in the respective samples, filtered using false-discovery rate cut-offs. Recent efforts extend this functionality by incorporating standard or proteomic-specific bioinformatics pipelines, including machine learning, and by integrating the proteomic data with other omics-type data such as various flavours of next-generation sequencing (NGS).

Multidimensional readout of the functional cell states

Advances in MS have now reached a state where a multitude of conceptually novel applications have become feasible in proteome identification and quantification, protein–protein interactions (interactomics), organellar proteomics, PTM detection and many more (Figure 3). It is now poised to make a major contribution in translational medicine, particularly in the identification and routine use of biomarkers. MS-based proteomics is a more complex technology than antibody-based methods, but its exquisite specificity of detection and global nature more than make up for this.

Generally, proteomics bridges the gap between genotype and phenotype as aberrations in the genetic information may or may not be functionally consequential to the cell. Proteomics can evaluate the consequences of genomic aberrations on protein functions, which can provide more specific biomarkers for disease subtypes or new therapeutic targets.

Dramatic advances in the sensitivity of MS workflows are now opening up the vista of single-cell proteomics. The added benefits of proteomics will be that single cells can be studied while retaining the full spatial information of the cellular environment. Furthermore, there are many more protein copies compared to their corresponding mRNAs, making single-cell proteomics inherently more robust. MS-based single-cell proteomics will directly reveal intercellular dynamics such as receptor–ligand interactions between cells and their microenvironment.

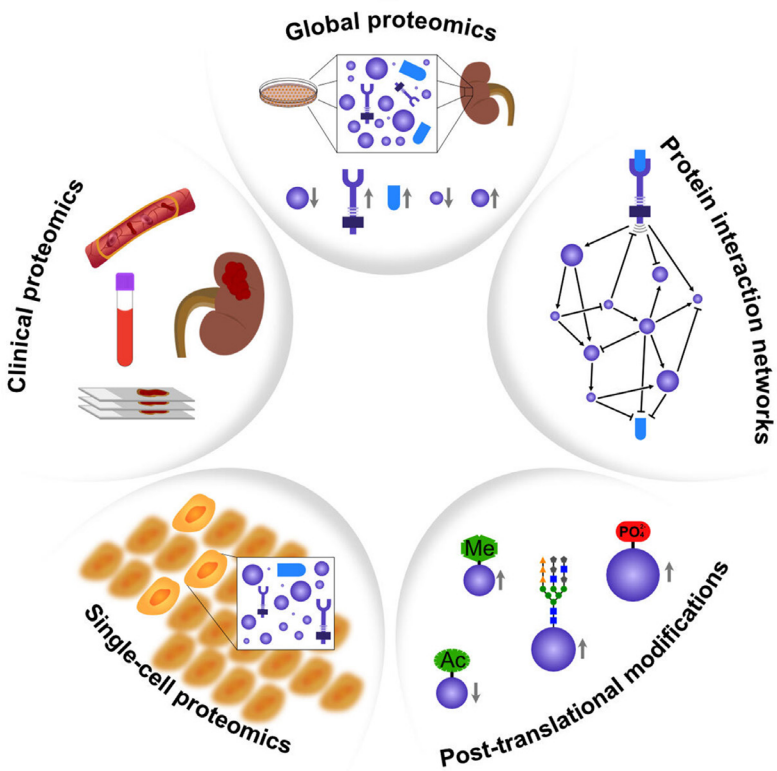

Figure 3. Some common applications of MS-based proteomics in biology. The analysis of global proteomes, interaction networks and post-translational modifications are examples of common applications. Recent innovations have expanded MS to clinical and single-cell proteomics. The grey arrow indicates changes in protein abundances.

In conclusion, we hope to have shed light on some of the basic concepts in MS-based proteomics. Proteins are multifaceted biomolecules as their functions are not just dictated by their abundances. Fortunately, MS-based proteomics is equally multifaceted and can readily adapt to study the various facets of proteins involved in biological functions. ■

Further reading

- Ludwig, C., Gillet, L., Rosenberger, G. et al. (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial, *Mol. Sys. Biol*. **14**, e8126, 10.15252/msb.20178126
- McLafferty, F.W. (2011) A century of progress in molecular mass spectrometry, *Annu. Rev. Anal. Chem*. **4**, 1–22, 10.1146/annurev-anchem-061010-114018
- Aebersold, R. and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355, 10.1038/nature19949
- Geyer, P.E., Holdt, L.M., Teupser, D. and Mann, M. (2017) Revisiting biomarker discovery by plasma proteomics. *Mol. Sys. Biol*. **13**, 942, https://doi.org/10.15252/msb.20156297
- Lundberg, E. and Borner, G.H. (2019) Spatial proteomics: a powerful discovery tool for cell biology, *Nat. Rev. Mol. Cell Biol*. **20**, 285–302, 10.1038/s41580-018-0094-y
- Budayeva, H.G. and Kirkpatrick, D.S.(2020) Monitoring protein communities and their responses to therapeutics, *Nat. Rev. Drug Discov*. **19**, 414–426, 10.1038/s41573-020-0063-y
- Zubarev, R.A. and Markov, A. (2013) Orbitrap mass spectrometry, *Anal. Chem*. **85**, 5288–5296, 10.1021/ac4001223
- Boresl U. (2012) Time-of-flight mass spectrometry: Introduction to the basics, *Mass Spectrom. Rev*. **36**, 86–109, 10.1002/mas.21520
- Müller, J.B., Geyer, P.E., Calaço, A.R. et al. (2020) The proteome landscape of the kingdoms of life, *Nature* **582**, 592–596, 10.1038/s41586-020-2402-x
- Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags, *Anal. Chem*. **66**, 4390–4399, 10.1021/ac00096a002

Ankit Sinha is an EMBO long-term fellow in the laboratory of Matthias Mann. He received his PhD in Medical Biophysics at the University of Toronto. His doctoral research focused on evaluating the performance of prognostic biomarkers developed from multi-omics data and identification of novel serological biomarkers of ovarian cancer. As part of Matthias Mann's team, he is currently researching the expansion of ion mobility mass spectrometry to cancer and heart disease.

Matthias Mann obtained his PhD in chemical engineering at Yale, contributing to the Nobel Prize for his supervisor John Fenn for electrospray ionization. He is a director at the Max Planck Institute of Biochemistry, Munich and also director of the Department of Proteomics, Novo Nordisk Foundation Center for Protein Research at the University of Copenhagen. He has obtained numerous prizes and is one of the most cited scientists with an h-factor of 232 and 250,000 total citations. He has pioneered advances in sample preparation, chromatography, mass spectrometry and computer algorithms, making mass spectrometry applicable to molecular biology. Apart from proteomics technology development, the team works on biological questions in systems biology. Today, the main focus is on clinically relevant topics, especially the analysis of the blood plasma proteome and cancer tissues at the single-cell level. Email: mmann@biochem.mpg.de