A Practical Beginner's Guide to Proteomics

This manuscript (permalink) was automatically generated from jessegmeyerlab/proteomics-tutorial@4065c4a on May 31, 2022.

Authors

- **Muralidharan Vanuopadath**
  0000-0002-9364-917X · vanuopadathmurali · V_MuraleeDhar
  School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam-690 525, Kerala, India

- **Amit Kumar Yadav**
  0000-0002-9445-8156 · aky · theoneamit
  Translational Health Science and Technology Institute · Funded by Grant BT/PR16456/BID/7/624/2016 (Department of Biotechnology, India); Grant Translational Research Program (TRP) at THSTI funded by DBT

- **Devasahayam Arokia Balaya Rex**
  0000-0002-9556-3150 · ArokiaRex · rexprem
  Center for Systems Biology and Molecular Medicine, Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore 575018, India

- **Dina Schuster**
  0000-0001-6611-8237 · dschust-r · dina_sch
  Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland; Department of Biology, Institute of Molecular Biology and Biophysics, ETH Zurich, Zurich 8093, Switzerland; Laboratory of Biomolecular Research, Division of Biology and Chemistry, Paul Scherrer Institute, Villigen 5232, Switzerland

- **Emma H. Doud**
  0000-0003-0049-0073 · edoud1 · fireinlab
  Center for Proteome Analysis, Indiana University School of Medicine, Indianapolis, Indiana, USA

- **Martín L. Mayta**
  0000-0002-7986-4551 · martinmayta · MartinMayta2
  School of Medicine and Health Sciences, Center for Health Sciences Research, Universidad Adventista del Plata, Libertador San Martín 3103, Argentina; Molecular Biology Department, School of Pharmacy and Biochemistry, Universidad Nacional de Rosario, Rosario 2000, Argentina

- **Benjamin A. Neely**
  0000-0001-6120-7695 · neely · neely615
  Chemical Sciences Division, National Institute of Standards and Technology, NIST Charleston · Funded by NIST

- **Jesse G. Meyer**
  0000-0003-2753-3926 · jessegmeyerlab · j_my_sci
  Department of Computational Biomedicine, Cedars Sinai Medical Center · Funded by Grant R21 AG074234; Grant R35 GM142502

Abstract

Proteomics is the large scale study of protein structure and function from biological systems. “Shotgun proteomics” or “bottom-up proteomics” is the prevailing strategy, in which proteins are hydrolyzed into peptide that are analyzed by mass spectrometry. Proteomics studies can be applied to diverse studies ranging from simple protein identification to studies of protein-protein interactions, absolute and relative protein quantification, post-translational modifications, and protein stability. To enable this range of different experiments, there are diverse strategies for proteome analysis. The nuances of how proteomic workflows differ may be difficult to understand for new practitioners. Here, we provide a comprehensive tutorial of different proteomics methods. Our tutorial covers all necessary steps starting from protein extraction and ending with biological interpretation. We expect that this work will serve as a basic resource for new practitioners of the field of shotgun or bottom-up proteomics.

Introduction

Proteomics is the large scale study of protein structure and function. Proteins are translated from mRNAs that are transcribed from the genome. Although the genome encodes potential cellular functions and states, the study of proteins is necessary to truly understand biology. Currently, proteomic studies are facilitated by mass spectrometry, although alternative methods are being developed.

Modern proteomics started around the year 1990 with the introduction of soft ionization methods that enabled, for the first time, transfer of large biomolecules into the gas phase without destroying them [1,2]. Shortly afterward, the first computer algorithm for matching peptides to a database was introduced [3]. Another major milestone that allowed identification of over 1000 proteins were actually improvements to chromatography [4]. As the volume of data exploded, methods for statistical analysis transitioned use from the wild west to modern informatics based on statistical models [5] and the false discovery rate [6].

Two strategies of mass spectrometry-based proteomics differ fundamentally by whether proteins are cleaved into peptides before analysis: “top-down” and “bottom-up”. Bottom-up proteomics (also refered to as shotgun proteomics) is defined by the hydrolysis of proteins into peptide pieces [7]. Therefore, bottom-up proteomics does not actually measure proteins, but must infer their presence [5]. Sometimes proteins are infered from only one peptide sequence representing a small fraction of the total protein sequence predicted from the genome. In contrast, top-down proteomics attempts to measure all proteins intact [8]. The potential benefit of top-down proteomics is the ability to measure proteoforms [9]. However, due to analytical challenges, the depth of protein coverage that is achievable by top-down proteomics is less than the depth that is achievable by bottom-up proteomics.

In this tutorial we focus on the bottom-up proteomics workflow. The most common version of this workflow is generally comprised of the following steps. First, proteins in a biological sample must be extracted. Usually this is done by denaturing and solubilizing the proteins while disrupting DNA and tissue. Next, proteins are hydrolyzed into peptides, usually using a protease like trypsin. Peptides from proteome hydrolysis must be purified. Most often this is done with reversed phase chromatography cartridges or tips. The peptides are then almost always separated by liquid chromatography before they are ionized and introduced into a mass spectrometer. The mass spectrometer then collects precursor and fragment ion data from those peptides. The data analysis is usually the rate limiting step. Peptides must be identified, and proteins are inferred and quantities are assigned. Changes in proteins across conditions are determined with statistical tests, and results must be interpreted in the context of the relevant biology.

There are many variations on this workflow. The wide variety of experimental goals that are achievable with proteomics technology leads to a wide variety of potential proteomics workflows. Even choice is important and every choice will affect the results. In this tutorial, we cover all of the required steps in detail to serve as a tutorial for new proteomics practioners.

1. Types of experiments enabled by proteomics
2. Protein extraction
3. proteolysis
4. Isotopic Labeling
5. Enrichments
6. Peptide purification
7. Mass Spectrometry
8. Peptide Ionization

9. Data Acquisition
10. Basic Data Analysis
11. Biological Interpretation
12. Experimental considerations and design

Types of Experiments

A wide range of questions are addressable with proteomics technology, which translates to a wide range of variations of proteomics workflows. Sometimes identifying what proteins are present is desired, and sometimes the quantities of as many proteins as possible are desired. Proteomics experiments can be both qualitative and quantitative.

Qualitative experiments

- Identifying proteins
- Identifying post translational modifications
- Identifying protein isoforms

Quantitative experiments

- Protein abundance changes
- Phosphoproteomics
- Glycoproteomics
- Structural techniques (XL-MS, HDX-MS, FPOP, protein-painting, LiP-MS, radical footprinting, ion mobility)
- Protein stability and small molecule binding (Thermal proteome profiling, TPP, or cellular thermal shift assay, CETSA)
- Protein-protein interactions (PPIs): AP-MS, APEX, BioID

Protein Extraction

Protein extraction from the sample of interest is the initial phase of any mass spectrometry-based proteomics experiment. Thought should be given to any planned downstream assays, specific needs of proteolysis (LiP-MS, post translational modification enrichments, enzymatic reactions, glycan purification or hydrogen-deuterium exchange experiments) long term project goals (reproducibility, multiple sample types, low abundance samples) as well as to the initial experimental question (coverage of a specific protein, subcellular proteomics, global proteomics, protein-protein interactions or immune or affinity enrichment of a specific classes of modifications.) The 2009 version of Methods in Enzymology: guide to Protein Purification [10] serves as a deep dive into how molecular biologists and biochemists traditionally thought about protein extraction. Any change in extraction conditions should be expected to create potential changes in downstream results. Optimize this step first and stick with a protocol that works for your needs. If a collaborator is attempting to reproduce your results, make sure they begin with the same extraction protocols.

Buffer choice

General proteomics

A common question to proteomics core facilities is, “What is the best buffer for protein extraction?” Unfortunately, there is no one correct answer. For global proteomics experiments, a buffer of neutral pH (50-100 mM PBS, Tris, HEPES, Ammonium Bicarbonate, triethanolamine bicarbonate; pH 7.5-8.5) is used in conjunction with a chaotrope or surfactant to denature and solubilize proteins (e.g., 8 M urea, 6 M guanidine, 5% SDS) [11,12]. Often other salts like 50-150 mM NaCl are also added. If intact protein separations are planned (based on size or isoelectric point) choose a denaturant compatible with those methods, such as SDS. Compatibility with protease (typically trypsin) and peptide cleanup steps will need to be considered. 8 M urea must be diluted to 2 M or less for trypsin and chymotrypsin digestions, while guanidine and SDS should be removed either through protein precipitation, through filter-assisted sample preparation (FASP), or similar solid phase digestion techniques. Note that some buffers can potentially introduce modifications onto proteins such as carbamylation from urea at high temperatures {[13]}. Newer mass spectrometry compatible detergents are also useful for protein extraction and ease of downstream processing – including Rapigest® (Waters), N-octyl-β-glucopyranoside, Azo (Ge lab Wisconsin), PPS silent surfactant. AVOID the use of tween-20, triton-X, NP-40, and PEGs as these compounds are challenging to remove after digestion.

Protein-protein interactions

Denaturing conditions will efficiently extract proteins – but they will denature/disrupt most protein-protein interactions. If you are working on an immune- or affinity purification of a specific protein and expect to analyze enzymatic activity, structural features, and/or protein-protein interactions, a non-denaturing lysis buffer should be utilized. Check the calculated pI and hydrophobicity for a good idea of starting pH/conductivity, but you may need to perform a stability screen. In general, the buffer will still be close to neutral pH with 50-250 mM NaCl. A low percent of mass spec compatible detergent may also be used.

Optional additives

For denaturing buffer conditions, additional additives may not be necessary for successful extraction and to prevent proteolysis or PTM modifications throughout the extraction process. Protease, phosphatase and deubiquitinase inhibitors are optional additives in less denaturing conditions or in experiments focused on specific post-translational modifications. Keep in mind that protease

inhibitors may impact digestion conditions and will need to be diluted or removed prior to trypsin addition. For extraction of DNA or RNA binding proteins, addition of a small amount of benzonase might be useful for degradation of any bound nucleic acids and result in a more consistent digestion.

Mechanical or Sonic Disruption

Cell lysis

One typical lysis buffer is 8 M urea in 100 mM Tris pH 8.5. Small mammalian cell pellets and exosomes will lyse almost instantly upon addition of this sort of denaturing buffer. Efficiency of extraction and degradation of nucleic acids can be improved using various sonication methods: 1) probe sonicator with ice; 2) water bath sonicator with ice or cooling; 3) bioruptor® sonication device 4) Adaptive focused acoustics (AFA®) [**PMID?** 21060726]. Key to these additional lysis techniques are to keep the temperature of the sample from rising significantly which can cause proteins to aggregate or degrade. Some cell types may require additional force for effective lysis (see below). For bacteria with cell wells, lysozyme is often added in the lysis buffer. Any added protein will be present in downstream results, however, so excessive addition of lysozyme is to be avoided unless tagged protein purification will occur.

Tissue/other lysis

Although small pieces of soft tissue can often be successfully extracted with the probe and sonication methods described above, larger/harder tissues as well as plants/yeast/fungi are better extracted with some form of additional mechanical force. If proteins are too be extracted from a large amount of sample, such as soil, feces, or other diffuse samples, one option is to use a dedicated blender and filter the sample after followed by centrifugation. If samples are tissue or more concise, cryo-homogenization is recommended. The simplest form of this is grinding the sample with liquid nitrogen and a mortar and pestle. Tools such as bead beaters are also used, where the sample is placed in a tube with appropriately sized beads and shaken rapidly. Cryo-mills are chambers where liquid nitrogen is applied around a vessel and large bead or beads. Cryo-fractionators homogenize samples in special bags that are frozen in liquid nitrogen and smashed with various degrees of force [**PMID?** 34002278]. After homogenization, samples can be sonicated by one of the methods above to fragment DNA and increase solubilization of proteins.

Efficiency of protein extraction

Following protein extraction, samples should be centrifuged (10-14,000 g for 10-30 min depending on sample type) to remove debris prior to calculating protein concentration. The amount of insoluble material remaining should be noted throughout an experiment as a large change may indicate protein extraction issues. Protein concentration can be calculated using a number of assays or tools; generally UV absorption methods are facile and affordable, such as Bradford or BCA assays. Protein can also be estimated by tryptophan fluorescence, which has the benefit of not consuming sample [14]. A nanodrop UV spectrophotometer may be used. Consistency in this method is important as each method will have inherent bias and error. Extraction buffer components will need to be compatible with any assay chosen; alternatively, buffer may be removed (see below) prior to protein concentration calculation.

Reduction and alkylation

Typically, disulfide bonds in proteins are reduced and alkylated prior to proteolysis in order to disrupt structures and simplify peptide analysis. This allows better access to all residues during proteolysis and removes the crosslinked peptides created by S-S inter peptide linkages. There are a variety of

reagent options for these steps. For reduction, the typical agents used are 5-15 mM concentration of TCEP/DTT/2BME. For the following alkylation step, a slightly higher 10-20mM concentration, alkylating agent such as Chloroacetamide/iodoacetamide, n-ethyl maleimide can be used [PMID: 29019370; [**PMID?**]: 15351294; [**PMID?**]: 28539326]. It is also possible to alklylate free cysteines with one reagent, reduce di-sulfide bonds (or other Cysteine modifications) and alkylate with a different reagent in order to monitor which residues are linked/modified in a protein. Reduction reactions are generally carried out in the dark at room temperature to avoid excessive off target alklation.

Removal of buffer/interfering small molecules

If extraction must take place in a buffer which is incompatible for efficient proteolysis (check the guidelines for the protease of choice), then protein cleanup should occur prior to digestion. This is generally performed through precipitation of proteins. The most common types are 1) acetone, 2) trichloroacetic acid (TCA), and 3) methanol/chloroform/water. Proteins are generally insoluble in most pure organic solvents, so cold ethanol or methanol are sometimes used. Pellets should be washed with organic solvent for complete removal especially of detergents. Alternatively, solid phase based digestion methods such as S-trap, FASP, and on column/bead can allow for proteins to be applied to a solid phase and buffers removed prior to proteolysis [**PMID?** 29754492]. Specialty detergent removal columns exist (Pierce/Thermo Fisher Scientific) but add expense and time consuming steps to the process. Additionally these should be checked for efficiency prior to implementing in a workflow with many samples as avoiding detergent contamination in the LC/MS is very important.

Summary

Often you will be given protein extraction conditions from molecular biologists or biochemistry which you will have to make work with downstream mass spectrometry applications. For bottom-up proteomics, the overreaching goal is efficient and consistent extraction and digestion.

<–! test edit example –>

Proteolysis

Proteolysis is the defining step that differentiates bottom-up or shotgun proteomics from top-down proteomics. Hydrolysis of proteins is extremely important because it defines the population of potentially identifyable peptides. Generally peptides between a length of 7-35 amino acids are considered useful for mass spectrometry analysis. Peptides that are too long are difficult to identify by tandem mass spectrometry, or may be lost during sample preparation due to irreversible binding with solid-phase extraction sorbents. Peptides that are too short are also not useful because they may match to many proteins during protein inference. There are many choices of enzymes and chemicals that hydrolyze proteins into peptides. This section summarizes potential choices and their strengths and weaknesses.

Trypsin is the most common choice of protease for proteome hydrolysis [15]. Trypsin is favorable because of its specificity, availability, efficiency and low cost. Trypsin cleaves at the C-terminus of basic amino acids, Arg and Lys. Many of the peptides generated from trypsin are short in length (less than ~ 20 amino acids), which is ideal for chromatographic separation, MS-based peptide fragmentation and identification by database search. The main drawback of trypsin is that majority (56%) of the tryptic peptides are ≤ 6 amino acids, and hence using trypsin alone limits the observable proteome [16,17,18]. This limits the number of identifiable protein isoforms and post-translational modifications.

3. theoretical studies of proteolysis enzymes [19]
4. Challenges associated with alternative enzyme choices (non-specific and semi-specific enzymes)

Many alternative proteases are available with different specificities that complement trypsin to reveal different proteomic sequences [16,20], which can help distinguish protein isoforms [21]. The enzyme choice mostly depends on the application. In general, for a mere protein identification mostly trypsin is the choice due to the reasons aforementioned. However, alternative enzymes can facilitate *de novo* assembly when the genomic data information is limited in the public database repositories [22,23,24,25,26]. Use of multiple proteases for proteome digestion also can improve the sensitivity and accuracy of protein quantification [27]. Moreover, by providing an increased peptide diversity, the use of multiple proteases can expand sequence coverage and increase the probability of finding peptides which are unique to single proteins [19,28,29]. A multi-protease approach can also improve the identification of N-Termini and signal peptides for small proteins [30]. Overall, integrating multiple-protease data can increase the number of proteins identified [31,32], the number of identified post-translational modifications detected [28,29,33] and decrease the ambiguity of the protein group list [28].

Lysyl endopeptidase (Lys-C) obtained from *Lysobacter enzymogenesis* is a serine protease involved in cleaving carboxyl terminus of Lys [17]. Like trypsin, the optimum pH range required for its activity is from 7 to 9. A major advantage of Lys-C is its resistance to denaturing agents, including 8 M urea - a chaotrope commonly used to denature proteins *prior* to digestion [21]. Trypsin is less efficient at cleaving Lys than Arg, which could limit the quality of quantitation from tryptic peptides. Hence, to achieve complete protein digestion with minimal missed cleavages, Lys-C is often used simultaneously with trypsin digestion [35].

Alpha-lytic protease (aLP) is also secreted by the soil bacterial *Lysobacter enzymogenesis* [36]. Wild-type aLP (WaLP) and an active site mutant of aLP, M190A (MaLP), have been used to expand proteome coverage [29]. Based on observed peptide sequences from yeast proteome digestion, WaLP showed a specificity for small aliphatic amino acids like alanine, valine, and glycine, but also threonine and serine. MaLP showed specificity for slightly larger amino acids like methionine, phenylalanine, and

surprisingly, a preference for leucine over isoleucine. The specificity of WaLP for threonine enabled the first method for mapping endogenous human SUMO sites [37].

Glutamyl peptidase I, commonly known as Glu-C or V8 protease, is a serine protease obtained from *Staphyloccous aureus* [38]. Glu-C cleaves at the C-terminus of glutamate, but also after aspartate [38,39].

Peptidyl-Asp metallopeptidase, commonly known as Asp-N, is a metalloprotease obtained from *Pseudomonas fragi* [40]. Asp-N catalyzes the hydrolysis of peptide bonds at the N-terminal of aspartate residues. The optimum activity of this enzyme occurs at a pH range between 4 and 9. As with any metalloprotease, chelators like EDTA should be avoided for digestion buffers when using Asp-N. Studies also suggest that Asp-N cleaves at the amino terminus of glutamate when a detergent is present in the proteolysis buffer [40]. Asp-N often leaves many missed cleavages [21].

Chymotrypsin or chymotrypsinogen A is a serine protease obtained from porcine or bovine pancreas with an optimum pH range from 7.8 to 8.0 [41]. It cleaves at the C-terminus of hydrphobic amino acids Phe, Trp, Tyr and barely Met and Leu residues. Since the transmembrane region of membrane proteins commonly lacks tryptic cleavage sites, this enzyme works well with membrane proteins having more hydrophobic residues [21,42,43]. The chymotryptic peptides generated after proteolysis will cover the proteome space orthogonal to that of tryptic peptides both in a quantitative and qualitative manner [43,44,45]

Clostripain, commonly known as Arg-C, is a cysteine protease obtained from *Clostridium histolyticum* [46]. It hydrolyses mostly the C-terminal Arg residues and sometimes Lys residues, but with less efficiency. The peptides generated are generally longer than that of tryptic peptides. Arg-C is often used with other proteases for improving qualitative proteome data and also for investigating PTMs [17].

LysargiNase, also known as Ulilysin, is a recently discovered protease belonging to the metalloprotease family. It is a thermophilic protease derived from *Methanosarcina acetivorans* that specifically cleaves at the N-terminus of Lys and Arg residues [47]. Hence, it enabled discovery of C-terminal peptides that were not observed using trypsin. In addition, it can also cleave modified amino acids such as methylated or dimethylated Arg and Lys [47].

Peptidyl-Lys metalloendopeptidase, or Lys-N, is an metalloprotease obtained from *Grifola frondosa* [48]. It cleaves N-terminally of Lys and has an optimal activity at pH 9.0. Unlike trypsin, Lys-N is more resistant to denaturing agents and can be heated up to 70 °C [17]. Reports suggest that the peptides generated after Lys-N digestion produces more of c-type ions in a ETD-based mass spectrometer [49]. Hence this can be used for analysing PTMs, identification of C-terminal peptides and also for *de novo* sequencing strategies [49,50].

Pepsin A, commonly known as pepsin, is an aspartic protease obtained from bovine or porcine pancreas [51]. Pepsin was one of several proteins crystalized by John Northrop, who shared the 1946 Nobel prize in chemistry for this work [52,53,54,55/]. Pepsin works at an optimum pH range from 1 to 4 and specifically cleaves Trp, Phe, Tyr and Leu [17]. Since it possess high enzyme activity and broad specificity at lower pH, it is preferred over other proteases for MS-based disulphide mapping [56,57]. Pepsin is also used extensively for structural mass spectrometry studies with hydrogen-deuterium exchange (HDX) because the rate of back exchange of the amide deuteron is minimized at low pH [58,59].

Proteinase K was first isolated from the mold *Tritirachium album* Limber [60]. The epithet 'K' is derived from its ability to efficiently hydrolyse keratin [60]. It is a member of the subtilisin family of proteases and is relatively unspecific with a preference for proteolysis at hydrophobic and aromatic amino acid

residues [61]. The optimal enzyme activity is between pH 7.5 and 12. Proteinase K is used at low concentrations for limited proteolysis (LiP) and the detection of protein structural changes in the eponymous technique LiP-MS [62].

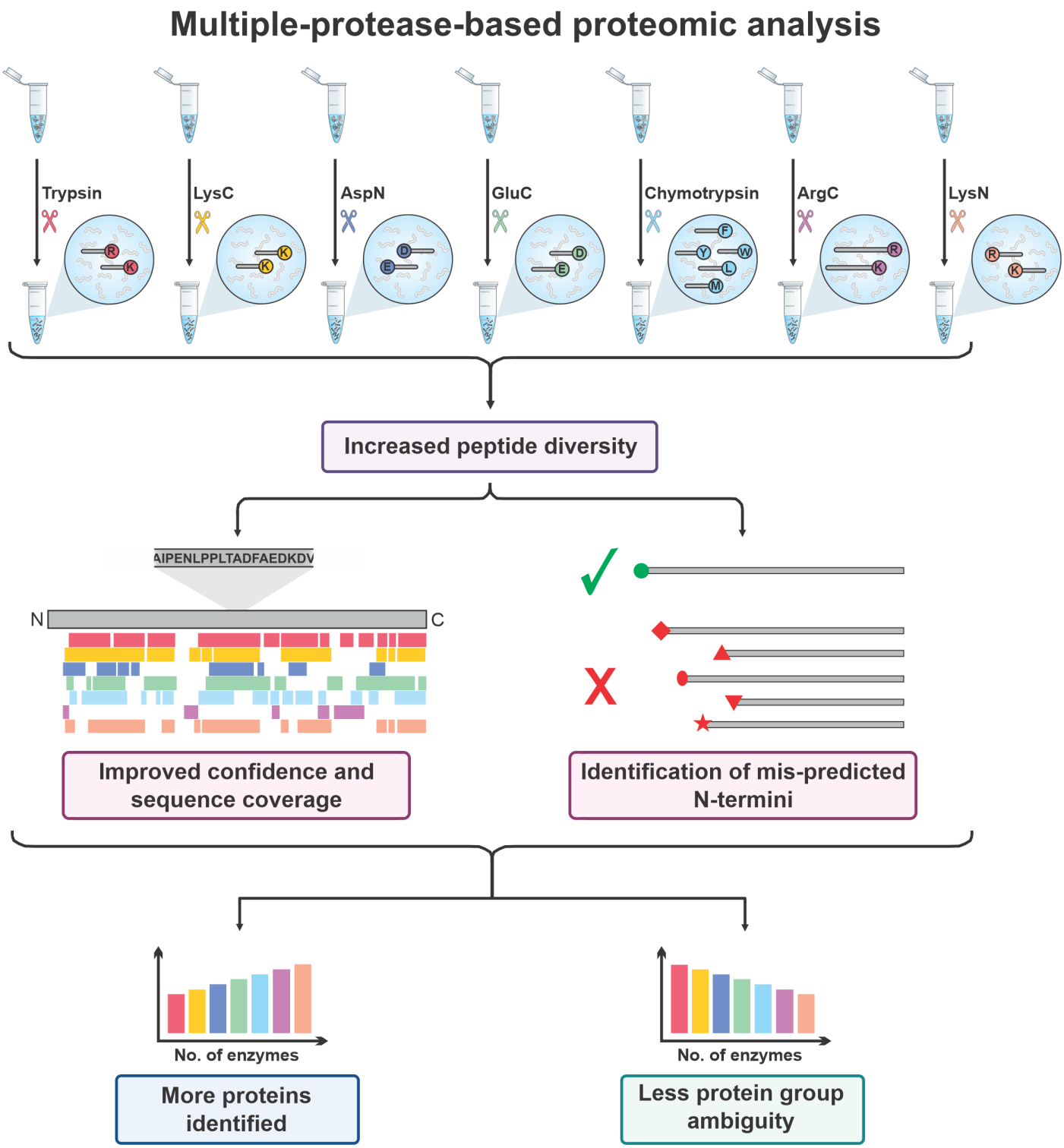

Figure 1: Multiple protease proteolysis improves protein inference The use of other proteases beyond Trypsin such as Lysyl endopeptidase (Lys-C), Peptidyl-Asp metallopeptidase (Asp-N), Glutamyl peptidase I, (Glu-C), Chymotrypsin, Clostripain (Arg-C) or Peptidyl-Lys metalloendopeptidase (Lys-N) can generate a greater diversity of peptides. This improves protein sequence coverage and allows for the correct identification of their N-termini. Increasing the number of complimentary enzymes used will increase the number of proteins identified by single peptides and decreases the ambiguity of the assignment of protein groups. Therefore, this will allow more protein isoforms and post-translational modifications to be identified than using Trypsin alone.

Peptide and Protein Labeling

Discussion of methods to isotopically label peptides or proteins that enable quantification

1. SILAC/SILAM
2. iTRAQ
3. TMT
4. dimethyl labeling

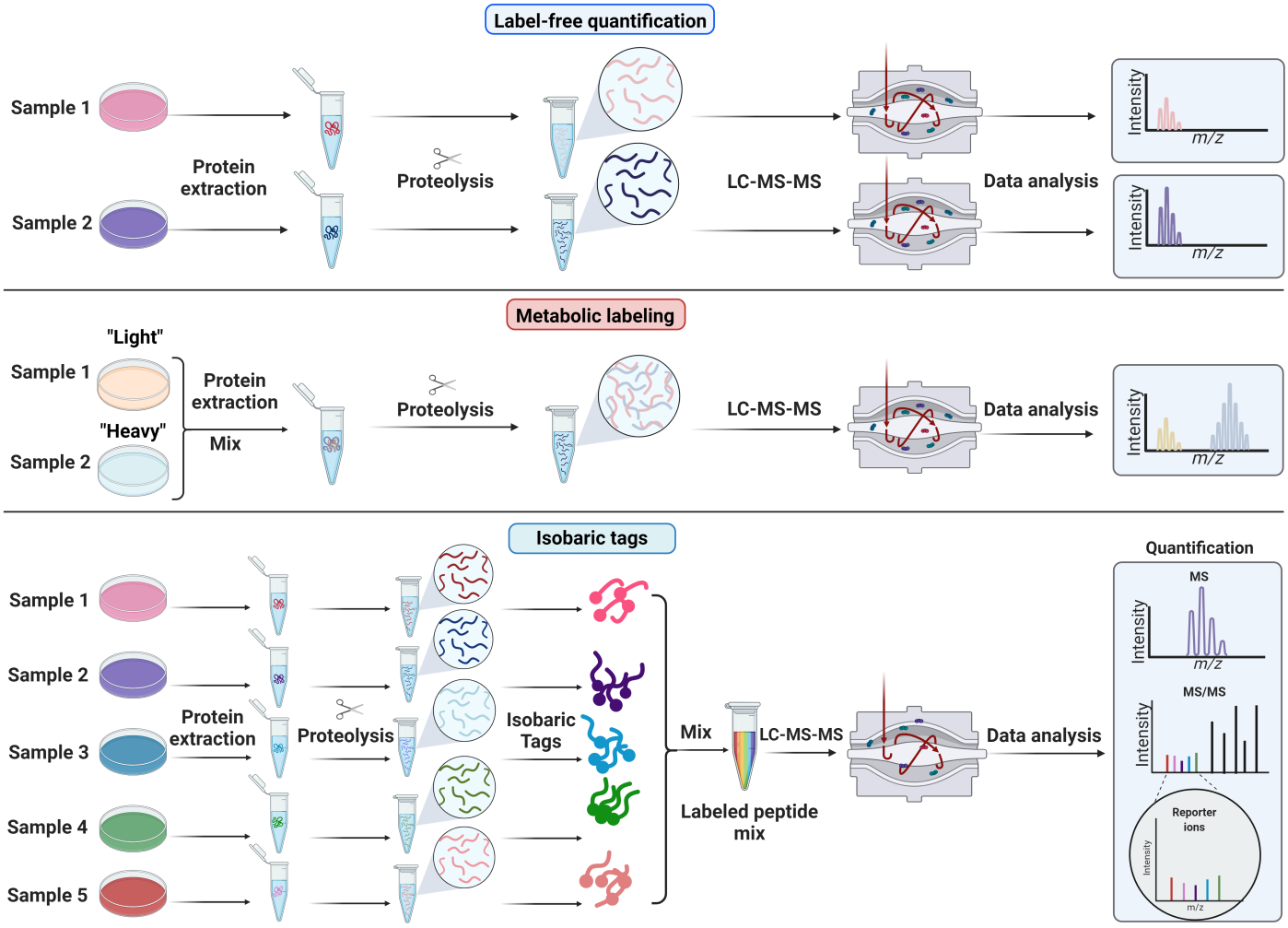

Figure 1: Quantitative strategies commonly used in proteomics. A) Label-free quantitation. Proteins are extracted from samples, enzymatically hydrolyzed into peptides and analyzed by mass spectrometry. Chromatographic peak areas from peptides are compared across samples that are analyzed sequentially. B) Metabolic labelling. Stable isotope labeling with amino acids in cell culture (SILAC) is based on feeding cells stable isotope labeled amino acids (“light” or “heavy”). Samples grown with heavy or light amino acids are mixed before cell lysis. The relative intensities of the heavy and light peptide are used to compute protein changes between samples. C) Isobaric or chemical labelling. Proteins are isolated separately from samples, enzymatically hydrolyzed into peptides, and then chemically tagged with isobaric stable isotope labels. These isobaric tags produce unique reporter mass-to-charge (m/z) signals that are produced upon fragmentation with MS/MS. Peptide fragment ions are used to identify peptides, and the relative reporter ion signals are used for quantification.

Peptide or Protein Enrichment

Protein enrichment (e.g. for protein protein interactions)

- coIP

- APEX
- bioID
- bioplex

Peptide enrichment

- antibody enrichments of modifications, e.g. lysine acetylation [63].
- TiO2 and Fe enrichment of phosphorylation
- Glycosylation
- SISCAPA

Methods for Peptide Purification

1. Reverse phase including tips and cartridges
2. stage tips
3. in stage tip (iST)
4. SP2, SP3
5. s traps

Types of Mass Spectrometers used for Proteomics

1. QQQ
2. Q-TOF
3. Q-Orbitrap
4. LTQ-Orbitrap
5. TOF/TOF
6. FT-ICR
7. types of ion mobility

- SLIM
- FAIMS
- traveling wave
- tims

Peptide Ionization

Until the early 1990s, peptides analysis by mass spectrometry was challenging. Hard ionization techniques in use at the time, like fast atom bombardment, were not directly applicable to peptides without destroying or breaking them. The soft ionization tehniques however, revolutionized the proteomics field and it became possible to routinely ionize and analyze peptides using MALDI and ESI techniques at high-throughput scale. These two techniques were so impactful that the 2002 Nobel Prize in Chemistry was co-awarded to John Fenn (ESI) and Koichi Tanaka (MALDI) “for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules” [64/].

MALDI

The term, Matrix-assisted LASER desorption ionization (MALDI), was coined by Hillenkamp and Karas in 1985[65]. Karas and Hillenkamp discovered the MALDI technique first, although a similar ionization method was shown by Koichi Tanaka in 1988 [2]. A few months later, Karas and Hillenkamp also demonstrated MALDI applied to protein ionization [66]. It also created a controversy that the widely used method of MALDI from these two people had been overlooked, and the Nobel prize was awarded to Tanaka, whose system was rarely used[67].

MALDI first requires the peptide sample to be co-crystallized with a matrix molecule, which is usually a volatile, low molecular-weight, organic aromatic compound. Some examples of such compounds are cyno-hydroxycinnamic acid, dihyrobenzic acid, sinapinic acid, alpha-hydroxycinnamic acid, ferulic acid etc [68/]. Subsequently, the analyte is placed in a vacuum chamber in which it is irradiated with a LASER, usually at 337nm [69]. This laser energy is absorbed by the matrix, which then transfers that energy along with its free protons to the co-crystalized peptides without significantly breaking them. The matrix and co-crystallized sample generate plumes, and the volatile matrix imparts its protons to the peptides as it gets ionized first. The weak acidic conditions used as well as the acidic nature of the matrix allows easy exchange of protons for the peptides to get ionized and fly under the electrical field in the mass spectrometer. These ionized peptides generally form the metastable ions, most of them will fragment quickly [70]. However, it can take several milliseconds and the mass spectrometry analysis can be performed before this time. Peptides ionized by MALDI almost always take up a single charge and thus observed and detected as [M+H]+ species.

MALDI Mechanism

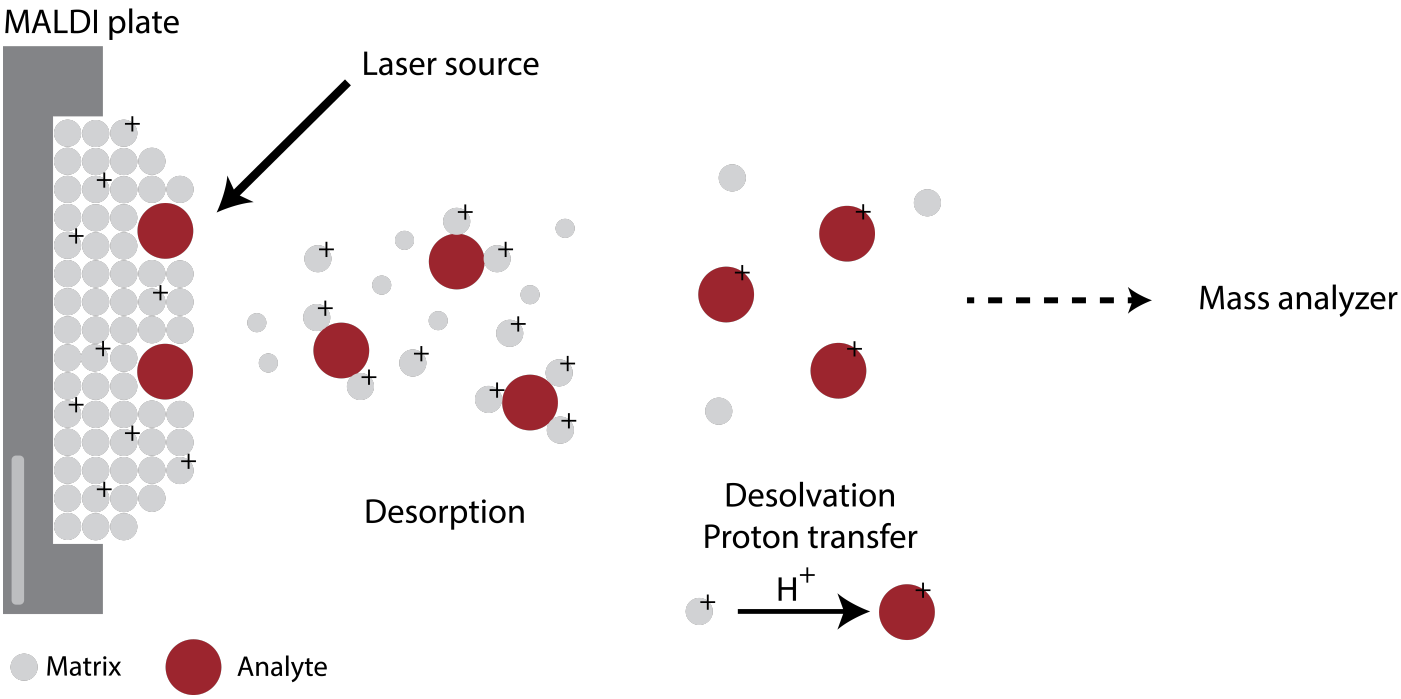

Figure 1: MALDI The analyte-matrix mixture is irradiated by a laser source, leading to ablation. Desorption and proton transfer ionize the analyte molecules that can then be accelerated into a mass spectrometer.

Electrospray Ionization

ESI was first applied to peptides by John Fenn and coworkers in 1989 [1]. Concepts related to electrospray Ionization (ESI) were published at least as early as 1882, when Lord Rayleigh described the number of charges that could assemble on the surface of a droplet [71]. ESI is usually coupled with reverse-phase liquid-chromatography of peptides directly interfaced to a mass spectrometer. A high voltage (~ 2 kV) is applied between the spray needle and the mass spectrometer. As solvent exits the needle, it forms droplets that take on charge at the surface, and through a debated mechanism, those charges are imparted to peptide ions. The liquid phase is generally kept acidic to help impart protons easily to the analytes.

Tryptic peptides ionized by ESI usually carry one charge one the side chain of their c-terminal residue (Arg or Lys) and one charge at their n-terminal amine. Peptides can have more than one charge if they have a longer peptide backbone, have histidine residues, or have missed cleavages leaving extra Arg and Lys. In most cases, peptides ionized by ESI are observed at more than one charge state. Evidence suggests that the distribution of peptide charge states can be maniupulated through chemical additives [72].

Electrospray Mechanism

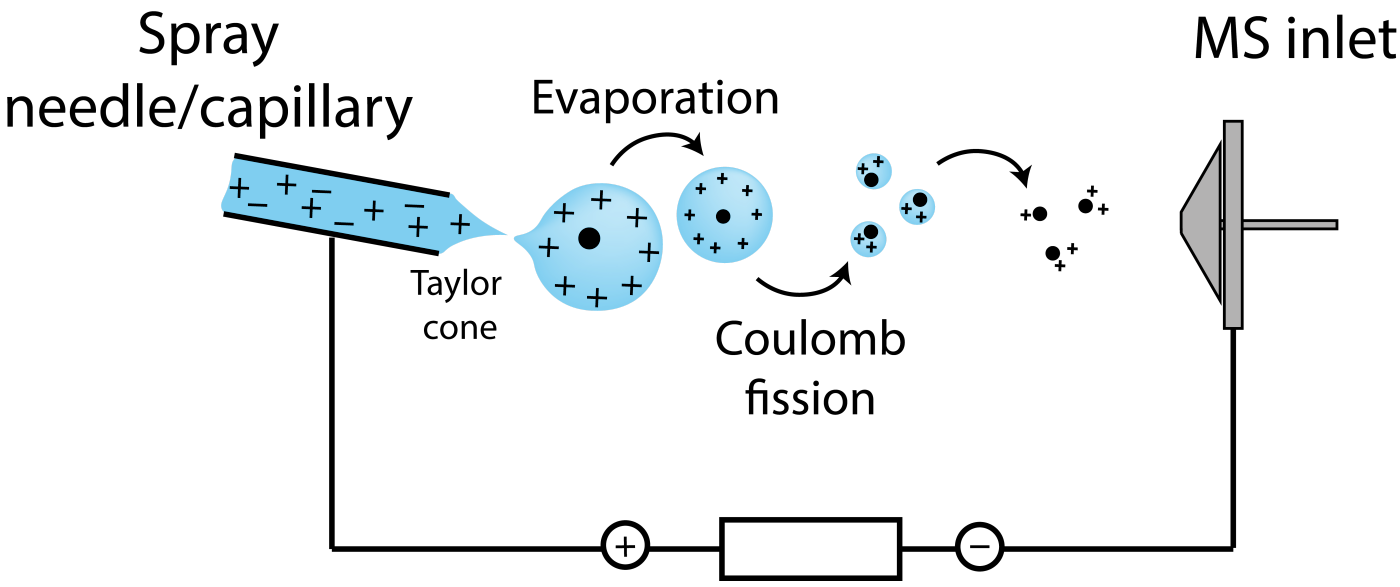

Figure 1: Electrospray Ionization Charged droplets are formed, their size is reduced due to evaporation until charge repulsion leads to Coulomb fission and results in charged analyte molecules.

The main goal of ESI is the production of gas-phase ions from electrolyte ions in solution. During the process of ionization, the solution emerging from the electrospray needle or capillary is distorted into a Taylor cone and charged droplets are formed. The charged droplets subsequently decrease in size due to solvent evaporation. As the droplets shrink, the charge density and Coulombic repulsion increase. This process destabilizes the droplets until the repulsion between the charges is higher than the surface tension and they fission (Coulomb explosion) [73] [74]. Typical bottom-up proteomics experiments make use of acidic analyte solutions which leads to the formation of positively charged analyte molecules due to an excess presence of protons.

Data Acquisition

Data acquisition strategies for proteomics fall generally within targeted or untargeted, and they can depend on the data (data dependent acquisition or DDA) or be data independent (data-independent acquisition or DIA).

DDA

Targeted DDA

Untargeted DIA

DIA

Targeted DIA

Untargeted DIA

Analysis of Raw Data

The goal of basic data analysis is to convert raw spectral data into identities and quantities of peptides and proteins that can be used for biologically-focused analysis. This step may often include measures of quality control, cross-run data normalization, quantification on different levels (precursor, peptide, protein), protein inference, PTM (post translational modification) localization and also first steps of data analysis, such as statistical hypothesis tests.

In typical bottom-up proteomics experiments, proteins are digested into peptides and further analyzed with LC-MS/MS systems. Peptides can have different PTMs and ionize differently depending on their length and amino acid distributions. Therefore, mass spectrometers often record different charge and modification states of one single peptide. The entity that is recorded on a mass spectrometer is usually referred to as a precursor ion (peptide with its modification and charge state). This precursor ion is fragmented and the precursor or peptide sequences are obtained though spectral matching. The quantity of a precursor is estimated with various methods. The measured precursor quantities are combined to generate a peptide quantity. Peptides are also often combined into a protein group through protein inference, which combines multiple peptide identifications into a single protein identification [75] [76]. Protein inference is still a challenge in bottom-up proteomics.

Due to the inherent differences in the data structures of DDA and DIA measurements, there exist different types of software that can facilitate the steps mentioned above. The existing software for DDA and DIA analysis can be further divided into freeware and non-freeware:

DDA freeware

Name	Publication	Website
MaxQuant	Cox and Mann, 2008[77]	MaxQuant
MSFragger	Kong et al., 2017[78]	MSFragger
Mascot	Perkins et al., 1999[79]	Mascot
MS-GF+	Kim et al., [80]	MS-GF+
X!Tandem	Craig et al., [81,82]	GPMDB

DIA freeware:

Name	Publication	Website
MaxDIA	Cox and Mann, 2008[77]	MaxQuant
Skyline	MacLean et al., 2010[83]	Skyline
DIA-NN	Demichev et al., 2019[84]	DIA-NN

Targeted proteomics freeware:

Name	Publication	Website
Skyline	MacLean et al., 2010[83]	Skyline

DDA non-freeware:

Name	Publication	Website
ProteomeDiscoverer		ProteomeDiscoverer
Mascot	Perkins et al., 1999[79]	Mascot
Spectromine		Spectromine
PEAKS	Tran et al., 2018[85]	PEAKS

DIA non-freeware:

Name	Publication	Website
Spectronaut	Bruderer et al., 2015[86]	Spectronaut
PEAKS	Tran et al., 2018[85]	PEAKS

Data Summary and Interpretation

Name	Publication	Website
Peptide Shaker	Vaudel *et al.*, 2015[87,88]	PeptideShaker, Peptide Shaker Online

Analysis of DDA data

DDA data analysis either directly uses the vendor proprietary data format directly with a proprietary search engine like Mascot, Sequest (through Proteome Discoverer), Paragon (through Protein Pilot), or it can be processed through one of the many freely available search engines or pipelines, for example, MaxQuant, MSGF+, X!Tandem, Morpheus, MSFragger, and OMSSA. Tables 1 and 4 give weblinks and citations for these software tools. For analysis with freeware, raw data is converted to either text-based MGF (mascot generic format) or into a standard open XML format like mzML [89] [[90]][91]. The appropriate FASTA file containing proteins predicted from that organism's genome is chosen as a reference database to search the experimental spectra. All search parameters like peptide and fragment mass errors (i.e. MS1 and MS2 tolerances), enzyme specificity, number of missed cleavages, chemical artefacts (fixed modifications) and potential biological modifications (variable/dynamic modifications) are specified before executing the search. The search algorithm scores each query spectrum against its possible peptide matches [92]. A spectrum and its best scoring candidate peptide are called a peptide spectrum match (PSM). The scores reflect a *goodness-of-fit* between an experimental spectrum and a theoretical one and do not necessarily depict the correctness of the peptide assignment.

For evaluating the matches, a decoy database is preferred as a null model for peptide matching. A randomized or reversed version of target database is used as a nonparametric null model. The decoy database can be searched separate from the target database (Kall's method)[93] or it can be combined with the target database before search (Elias and Gygi method)[94]. Using either separate method or concatenated database search method, an estimate of false hits can be calculated which is used to estimate the false discovery rate (FDR) [95]. The FDR denotes the proportion of false hits in the population accepted as true. For Kall's method: the false hits are estimated to be the number of decoys above a given threshold. It is assumed that the number of decoy hits that pass a threshold are the false hits. A similar number of target population may also be false. Therefore, the FDR is calculated as:

$$FDR = \frac{DecoyPSMs}{TargetPSMs}$$

For Elias and Gygi Method, the target population in which FDR is estimated changes. The target and decoy hits coming from a joint database compete against each other. For any spectrum, either a target or a decoy peptide can be the best hit. It is argued that the joint target-decoy population has decoy hits as confirmed false hits. However, due to the joint database search, the target database may also have equal number of false hits. Thus, the number of false hits is multiplied by two for FDR estimation.

$$FDR = \frac{2 * DecoyPSMs}{Target + DecoyPSMs}$$

Strategies for analysis of DIA data

Targeted proteomics data analysis

Quality control

Statistical hypothesis testing

Databases

What are they and where do you get them?

Protein Database Sources and Types

Many mass spectrometry-based proteomic techniques use search algorithms that require a defined theoretical search space to identify peptide sequences based on precursor mass and fragmentation patterns, which are then used to infer the presence and abundance of a protein. The search space is calculated from the potential proteins in a sample, which includes the proteome (often a single species) and expected contaminants. This is called database searching and the flat file of protein sequences in FASTA format acts as a protein database. In this section, we will describe major resources for proteome FASTA files (protein sequence collections), how to retrieve them, and suggested best practices for preserving FASTA file provenance to improve reproducibility.

In general, FASTA sequence collections can be retrieved from three central clearing houses: UniProt, RefSeq, and Ensembl. These will be discussed separately below as they each have specific design goals, data products, and unique characteristics. It is important to learn the following three points for each resource: the source of the underlying data, canonical versus non-canonical sequences, and how versioning works. These points, along with general best practices, such as using a taxonomic identifier, are essential to understand and communicate search settings used in analyses of proteomic datasets. Finally, it is critical to understand that sequence collections from these three resources are not the same, nor do they offer the same sets of species.

Key terminology may vary between resources, so these terms are defined here. The term “taxon identifier” is used across resources and is based on the NCBI taxonomy database. Every taxonomic node has a number, e.g., Homo sapiens (genus species) is 9606 and Mammalia (class) is 40674. This can be useful when retrieving and describing protein sequence collections. Another term used is “annotation”, which has different meanings in different contexts. Broadly, a “genome annotation” is the result of an annotation pipeline to predict coding sequences, and often a gene name/symbol if possible. Two examples are MAKER [96] and the RefSeq annotation pipeline [97]. Alternatively, “protein annotation” (or gene annotation) often refers to the annotation of proteins (gene products) using names and ontology (i.e., protein names, gene names/symbols, functional domains, gene onotology, keywords, etc.). Protein annotation is termed “biocuration” and described in detail by UniProt [98]. Lastly, there are established minimum reporting guidelines for referring to FASTA files established in MIAPE: Mass Spectrometry Informatics that are taxon identifier and number of sequences [99,100]. The FASTA file naming suggestions below are not official but are suggested as a best practice.

UniProt

The Universal Protein Resource (UniProt) [101], has three different products: UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The numerous resources and capabilities associated with the UniProt are not explored in this section, but these are well described on UniProt’s website. UniProtKB is the source of proteomes across the Tree of Life and is the resource we will be describing herein. There are broadly two types of proteome sequence collections: Swiss-Prot/TrEMBL and designated proteomes. The Swiss-Prot/TrEMBL type can be understood by discussing how data is integrated into UniProt. Most protein sequences in UniProt are derived from coding sequences submitted to EMBL-Bank, GenBank and DDBJ. These translated sequences are initially imported into TrEMBL database, which is why TrEMBL is also termed “unreviewed”. There are other sources of protein sequences, as described by UniProt [102]. These

include the Protein Data Bank (PDB), direct protein sequencing, sequences derived from the literature, gene prediction (from sources such as Ensembl) or in-house prediction by UniProt itself. Protein sequences can then be manually curated into the Swiss-Prot database using multiple outlined steps (described in detail by UniProt here [103]) and is why Swiss-Prot is also termed “reviewed”. Note that more than one TrEMBL entry may be removed and replaced by a single Swiss-Prot entry during curation. A search of “organism:9606” at UniProtKB will retrieve both the Swiss-Prot/reviewed and TrEMBL/unreviewed sequences for Homo sapiens. The entries do not overlap, so users often either use just Swiss-Prot or Swiss-Prot combined with TrEMBL, the latter being the most exhaustive option. With ever-increasing numbers of high-quality genome assemblies processed with robust automated annotation pipelines, TrEMBL entries will contain higher quality protein sequences than in the past. In other words, if a mammal species has 20 000 to 40 000 entries in UniProtKB and many of these are TrEMBL, users should be comfortable using all the protein entries to define their search space (more on this later when discussing proteomes at UniProtKB). Determining the expected size of a well-annotated proteome requires additional knowledge, but tools to answer these questions continue to improve. As more and more genome annotations are generated, the backlog of manual curation continues to increase. However, automated genome annotations are also rapidly improving, blurring the line between Swiss-Prot and TrEMBL utility.

The second type of protein sequence collections available at UniProtKB are designated proteomes, with subclasses of “proteome”, “reference proteome” or “pan-proteome”. As defined by UniProt, a proteome is the set of proteins derived from the annotation of a completely sequenced genome assembly (one proteome per genome assembly). This means that a proteome will include both Swiss-Prot and TrEMBL entries present in a single genome annotation, and that all entries in the proteome can be traced to a single complete genome assembly. This aids in tracking provenance as assemblies change, and metrics of these assemblies are available. These metrics include Benchmarking Universal Single-Copy Ortholog (BUSCO) score, and “Completeness” as Standard, Close Standard or Outlier based on the Complete Proteome Detector (CPD). Given the quality of genome annotation pipelines, using a proteome as a FASTA file for a species is the preferred method of defining search spaces now. Outside of humans, no higher eukaryotic Swiss-Prot sequence collections are complete enough for use in proteomics analyses, but this does not mean that the available Swiss-Prot plus TrEMBL protein sequence collection precludes accurate proteomic data analysis. Lastly, the difference between reference proteome and proteome is used to highlight model organisms or organisms of interest, but not to imply improved quality. UniProt also has support for the concept of “pan proteomes” (consensus proteomes for a closely related set of organisms) but this is mostly used for bacteria (e.g., strains of a given species will share a pan proteome).

When retrieving protein sequence collections as Swiss-Prot/TrEMBL or designated proteomes, there is an option of downloading “FASTA (canonical)” or “FASTA (canonical & isoform)”. The later includes additional manually annotated isoforms for Swiss-Prot sequences. Each Swiss-Prot entry has one canonical sequence chosen by the manual curator. Any additional sequence variants (mostly from alternative slicing) are annotated as differences with respect to the canonical sequence. Specifying “canonical” will select only one protein sequence per Swiss-Prot entry while specifying “canonical & isoforms” will download additional protein sequences by including isoforms for Swiss-Prot entries. Recently, an option to “download one protein sequence per gene (FASTA)” has been added. These FASTA files include Swiss-Prot and TrEMBL sequences to number about 20 000 protein sequences for a wide range of higher eukaryotic organisms.

The number of additional isoforms varies considerably by species. In the human, mouse, and rat proteomes of the total number of entries, 26 %, 40 % and 72 % are canonical, respectively. The choice of including isoforms is related to the search algorithm and experimental goals. For instance, if differentiating isoforms is relevant, they should be included otherwise they will not be detected. In cases where isoforms are present in the FASTA (evident by shared protein names) but these cannot be removed prior to downloading (e.g., California sea lion, Zalophus californianus, proteome

UP000515165, release 2022_01), non-redundant FASTA files can be manually generated (i.e., “remove_duplicates.py” via [104]). If possible, retrieving canonical protein sequences via proteomes is the most straight forward approach and in general appropriate for most search algorithms, versus the method of searching and downloading Swiss-Prot and/or TrEMBL entries.

Though FASTA files are the typical input of many search algorithms, UniProt also offers an XML and GFF format download. In contrast to the flat FASTA file format, the XML format includes sequence information as well as associated information like PTMs, which is used in some search algorithms like MetaMorpheus [105].

Once a protein sequence collection has been selected and retrieved, there is the evergreen question of how to name and report this to others in a way that allows them to reproduce the retrieval. The minimum reporting information is the taxon identified and number of sequences used [99,100]. The following naming format (and those below) augments this and is suggested for UniProtKB FASTA files (the use of underscores or hyphens is not critical): [common or scientific name]-[taxon id]-uniprot-[swiss-prot/trembl/proteome]-[UP# if used]-[canonical/canonical plus isoform]-[release] example of a Homo sapiens (human) protein fasta from UniProtKB:

Human-9606-uniprot-proteome-UP000005640-canonical-2022_01.fasta

The importance of the taxon identifier has already been described above and is a consistent identifier across time and shared across resources. The choices of Swiss-Prot and TrEMBL in some combination was discussed above, and Proteome can be “proteome”, “reference proteome” or “pan-proteome”. The proteome identifier (‘UP’ followed by 9 digits) is conserved across releases, and release information should also be included. A confusing issue to newcomers is what the term “release” means. This is a year_month format (e.g., 2022_01), but it is not the date a FASTA file was downloaded or created, nor does it imply there are monthly updates. This release “date” is a traceable release identifier that is listed on UniProt’s website. Including all this information ensures that the exact provenance of a FASTA file is known and allows the FASTA file to be regenerated.

RefSeq

NCBI is a clearing house of numerous types of data and databases. Specific to protein sequence collections, NCBI Reference Sequence Database (RefSeq) provides annotated genomes across the Tree of Life. The newly developed NCBI Datasets portal [106] is the preferred method for accessing the myriad of NCBI data products, though protein sequence collections can also be retrieved from RefSeq directly[107,108]. Like UniProt described above, most of the additional functionality and information available through NCBI Datasets and RefSeq will not be described here, although the Eukaryotic RefSeq annotation dashboard [109] is a noteworthy resource to monitor the progress of new or re-annotations. We recommend exploring the resources available from NCBI [110], utilizing their tutorials and help requests.

RefSeq is akin to the “proteome” sequence collection from UniProtKB, where a release is based on a single genome assembly. If a more complete genome assembly is deposited or additional secondary evidence (e.g., RNA sequencing) is deposited, RefSeq can update the annotation with a new annotation release. Every annotation release will have an annotation report that contains information on the underlying genome assembly, the new genome annotation, secondary evidence used, and various statistics about what was updated. The current annotation release is referred to as the “reference annotation”, but each annotation is numbered sequentially starting at 100 (the first release). Certain species are on scheduled re-annotation, like human and mouse, while other species are updated as needed based on new data and community feedback (ex. release 100 of taxon 9704 was in 2018, but a more contiguous genome assembly resulted in re-annotation to release 101 in 2020). This general process for new and existing species is described in Heck and Neely [111].

Since RefSeq is genome assembly-centric, its protein sequence collections are retrieved for each species. This contrasts with being able to use a higher-level taxon identifier like 40674 (Mammalia) in UniProt to retrieve a single FASTA. To accomplish this same search in NCBI Datasets requires a Mammalia search, followed by browsing all 2083 genomes and then filtering the results to reference genomes with annotations, and those resulting 188 could be bulk downloaded, though this will still be 188 individual FASTA files. It is possible to download a single FASTA from an upper-level taxon identifier using the NCBI Taxonomy Browser, though this service may be redundant with the new NCBI Datasets portal. Given the constant development of NCBI Datasets, these functionalities may change, but the general RefSeq philosophy of single species FASTA should be kept in mind. Likewise, when retrieving genome annotations there is no ability to specify canonical entries only, but it is possible to use computational tools to remove redundant entries ("remove_duplicates.py" from [104]).

Similar to the UniProtKB FASTA file naming suggestion, the following naming format is suggested for RefSeq protein sequence collection FASTA (the use of underscores or hyphens is not critical): [common or scientific name]-[taxon id]-refseq-[release number] example of a Equus caballus (horse) protein FASTA from RefSeq: Equus_caballus-9796-refseq-103.fasta The release number starts at 100 and is consecutively numbered. Note, the human releases only recently began following this consecutive numbering for Release 110, and previously had a much longer number to be included (e.g., NCBI Release 109.20211119). Also, in a few species (Human and Chinese hamster, currently), there is a reference and an alternate assembly, both with an available annotation. In these cases, including the underlying assembly identifier would be needed. Note that when you retrieve the protein FASTA from NCBI it will include two more identifiers that aren't required in the file name since it can be determined from the taxon identifier and release number. These are the genome assembly used (this is generated by the depositor and follows no naming scheme) and the RefSeq identifier (GCF followed by a number string). These aren't essential for FASTA naming, but are for comparing between UniProt, RefSeq and Ensembl when the same underlying assembly is used (or not, indicating how up to date one is versus the other).

Ensembl

There are two main web portals for Ensembl sequence collections: the Ensembl genome browser [112] has vertebrate organisms and the Ensemble Genome project [113] has specific web portals for different non-vertebrate branches of the Tree of Life. This contrasts with NCBI and UniProt where all branches are centrally available. Recently, Ensembl has created a new portal "Rapid Release" focusing on quickly making annotations available (replacing the "Pre-Ensemble" portal), albeit without the full functionality of the primary Ensembl resources. Overall, Ensembl provides diverse comparative and genomic tools that should be explored, but, specific to this discussion, they provide species-specific genome annotation products similar to RefSeq.

To retrieve a protein sequence collection from Ensemble at any of the portals, a species can be searched using a name, which will then have taxon identifier displayed (but searching by identifier is not readily apparent). From the results you can select your species and follow links for genome annotation. Caution should be used when browsing the annotation products since the protein coding sequence (abbreviated "cds") annotations are nucleic acid sequences (a useable via 3-frame translation if using certain software), while actual translated peptide sequences are in the "pep" folders. The pep folders contain file names with "ab initio" and "all" in the FASTA file names (file extensions are "fa" for FASTA and "gz" indicating gzip compression algorithm), while there may only be one pep product for certain species in the "Rapid Release" portal. The "ab initio" FASTA files contain mostly predicted gene products. The "all" FASTA files are the usable protein sequence collections. Ensembl FASTA files usually have some protein sequence redundancy.

Ensembl provides a release number for all the databases within each portal. Similar to the UniProt file naming suggestion, the following naming format is suggested for Ensembl protein sequence collection

FASTA (the use of underscores or hyphens is not critical):

[common or scientific name]-[taxon id]-ensembl-[abinitio/all]-[rapid]-[release number]

example of a Sus scrofa (pig) protein FASTA from Ensembl:

Pig-9823-ensembl-all-106.fasta

Similar to the FASTA download from RefSeq, the downloaded file name can include additional identifying information related to the underlying genome assembly. Again, this is not required for labeling, but is useful to easily compare assembly versions.

Since much of the data from Ensembl is also regularly processed into UniProt, using UniProt sequence collections instead may be preferred. That said, they are not on the same release schedule nor will the FASTA files contain the same proteins. Ensembl sequences still must go through the established protein sequence pipeline at UniProt to remove redundancy and conform to UniProt accession and FASTA header formats. Moreover, the gene-centric and comparative tools built into Ensembl may be more experimentally appropriate and using an Ensembl protein sequence collection can better leverage those tools.

Other resources

There are other locations of protein sequence collections, and these will likewise have different FASTA file formatting; sequences may have unusual characters, and formats of accessions and FASTA header lines may need to be reformatted to be compatible with search software. These alternatives include institutes like the Joint Genome Institute's microbial genome clearing house, species-specific community resource (e.g., PomBase, FlyBase, WormBase, TryTrypDB, etc.), and one-off websites tenuously hosting in-house annotations. It is preferred to use protein sequence collection from the main three sources described here, since provenance can be tracked, and versions maintained. It is beyond the scope of this discussion to address other genome annotation resources, how they are versioned, or the best way to describe FASTA files retrieved from those sources. In these cases, defaulting to the minimum requirements of listing number of entries and supplying the FASTA along with data are necessary.

Contaminants

Samples are rarely comprised of only proteins from the species of interest. There can be protein contamination during sample collection or processing. This may include proteins from human skin, wool from clothing, particles from latex, or even porcine trypsin itself, all of which contain proteins that can be digested along with the intended sample and analyzed in the mass spectrometer. Avoiding unwanted matching of mass spectra originating from contaminant proteins to the cellular proteins due to sequence similarities is important to the identification and quantitation of as many cellular proteins as possible. To avoid random matching, repositories of supplementary sequences for contaminant proteins have been added to a reference database for MS data searches. Appending a contaminants database to the reference database allows the identification of peptides that are not exclusive to one species. Peptides that are exclusive to the organism of interest are used to calculate abundance to avoid inflated quantitative results due to potential contaminant peptides.

As early as 2004, The Global Proteome Machine was providing a protein sequence collection of these common Repository of Adventitious Proteins (cRAP), while another contaminant list was published in 2008 [114]. The current cRAP version (v1.0) was described in 2012 [115] and is still widely in use today. cRAP is the contaminant protein list used in nearly all modern database searching software, though

the documentation, versioning or updating of many of these “built-in” contaminant sequence collections is difficult to follow. There is also another contaminant sequence collection distributed with MaxQuant. Together, the cRAP and MaxQuant contaminant protein sequence collections are found in some form across most software, including MetaMorpheus and Philosopher (available in FragPipe) [116]. This list of known frequently contaminating proteins can either be automatically included by the software or can be retrieved as a FASTA to be used along with the primary search FASTA(s). Recently the Hao Lab has revisited these common contaminant sequences in an effort to update the protein sequences, test their utility on experimental data, and add or remove entries [117].

In addition to these environmentally unintended contaminants, there are known contaminants that also have available protein sequence collections (or can be generated using the steps above) and should be included in the search space. These can include the media cells were grown in (e.g., fetal bovine serum [118,119], food fed to cells/animals (e.g., Caenorhabditis elegans grown on Escherichia coli) or known non-specific binders in affinity purification (i.e., CRAPome [120]). The common Repository of Fetal Bovine Serum Proteins (cRFP)[121] are protein lists of common protein contaminants and fetal serum bovine sequences used to reduced the number of falsely identified proteins in cell culture experiments. Cells washed or cultured in contaminant free media before harvest or the collection of secreted proteins depletes most high abundance contaminant proteins but the sequence similarity between contaminant and secreted proteins can cause false identifications and overestimation of the true protein abundance leading to wasted resources and time on validating false leads. As emphasized throughout this section, accurately defining the search space is essential for accurate results and, especially in the case of contaminants, requires knowledge of the experiment and sample processing to adequately define possible background proteins.

Choosing the right database

Proteomics data analysis requires carefully matching the search space (defined by the database choice) with the expected proteins. A properly chosen database will minimize false positives and false negatives. Choosing a database that is too large will increase the number of false positives, or decoy hits, which in turn will reduce the total number of identifiable proteins. For this reason it is ill advised to search against all possible protein sequences ever predicted from any genomic sequence. On the other hand, choosing a database that is too small may increase false negatives, or missed protein identifications, because in order for a protein to be identified it must be present in the database. Thus, proteomics practitioners must do their best to predict the proteins that might be in their sample before they analyze their data.

Proteomics data analysis requires carefully aligning the search space with the expected proteome and the statistical approach of the search algorithm. Search algorithms can self-correct when a database is overly large such that higher identity thresholds are required for identification to minimize false positives (e.g., Mascot), while smaller experiment-specific search spaces (also referred to as “subsets”) can have unintended effects on false positives if not managed appropriately [122,123,124] or may even improve protein identifications [125]. Whether to employ a search space that is sample-specific (i.e., subset), species-specific (with only canonical proteins, described below), exhaustive species-specific (including all isoforms), or even larger clade-level protein sequence set (e.g., the over 14 million protein sequences associated with Fungi, taxon identifier 4751) is a complex issue that is experiment and software dependent. Moreover, in cases where no species-specific protein sequence collection exists, homology-based searching can be used (as described in [111]). In each of these cases, proteomics practitioners must understand their specific experimental sample and search algorithm in order to know how to best define the search space, which is essential to yielding accurate results. See more discussion of database choice in the following section.

Biological Interpretation

1. term enrichment analysis (KEGG, GO)
2. network analysis methods
3. structure analysis
4. isoform analysis
5. follow-up experiments

Methods for protein or peptide fractionation

Protein fractionation * SDS-PAGE

Peptide fractionation * bRP

Experiment Design

This section should discuss trade offs and balancing them to design an experiment. 1. constraints: Each experiment will have different constraints, which may include the number of samples needed for analysis, or desire to quantify a specific subset of proteins within a sample. 2. sample size 3. statistics 4. costs

Acknowledgements

The authors thank Phil Wilmarth for helpful input. Identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

References

1. **Electrospray Ionization for Mass Spectrometry of Large Biomolecules**
John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, Craig M Whitehouse
Science (1989-10-06) https://doi.org/cq2q43
DOI: 10.1126/science.2675315 · PMID: 2675315

2. **Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry**
Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, T Matsuo
Rapid Communications in Mass Spectrometry (1988-08) https://doi.org/ffbwwr
DOI: 10.1002/rcm.1290020802

3. **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.**
JK Eng, AL McCormack, JR Yates
Journal of the American Society for Mass Spectrometry (1994-11)
https://www.ncbi.nlm.nih.gov/pubmed/24226387
DOI: 10.1016/1044-0305(94)80016-2 · PMID: 24226387

4. **An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics**
Dirk A Wolters, Michael P Washburn, John R Yates
Analytical Chemistry (2001-10-25) https://doi.org/bn4kq6
DOI: 10.1021/ac010617e · PMID: 11774908

5. **A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry**
Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, Ruedi Aebersold
Analytical Chemistry (2003-07-15) https://doi.org/b2xv45
DOI: 10.1021/ac0341261 · PMID: 14632076

6. **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**
Joshua E Elias, Steven P Gygi
Nature Methods (2007-03) https://doi.org/djz7fz
DOI: https://doi.org/10.1038/nmeth1019

7. **Mass-spectrometric exploration of proteome structure and function**
Ruedi Aebersold, Matthias Mann
Nature (2016-09) https://doi.org/f83zqm
DOI: 10.1038/nature19949 · PMID: 27629641

8. **High-throughput quantitative top-down proteomics**
Kellye A Cupp-Sutton, Si Wu
Molecular Omics (2020) https://doi.org/gnx98p
DOI: 10.1039/c9mo00154a · PMID: 31932818 · PMCID: PMC7529119

9. **Proteoforms as the next proteomics currency**
Lloyd M Smith, Neil L Kelleher
Science (2018-03-09) https://doi.org/gn6p4x
DOI: 10.1126/science.aat1884 · PMID: 29590032 · PMCID: PMC5944612

10. **Guide to protein purification**

Richard R Burgess, Murray P Deutscher
Elsevier/Academic Press (2009)
ISBN: 9780123745361

11. **Effective interactions between chaotropic agents and proteins.**
Giovanni Salvi, Paolo De Los Rios, Michele Vendruscolo
Proteins (2005-11-15) https://www.ncbi.nlm.nih.gov/pubmed/16152629
DOI: 10.1002/prot.20626 · PMID: 16152629

12. **A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin.**
Jennifer L Proc, Michael A Kuzyk, Darryl B Hardie, Juncong Yang, Derek S Smith, Angela M Jackson, Carol E Parker, Christoph H Borchers
Journal of proteome research (2010-10-01) https://www.ncbi.nlm.nih.gov/pubmed/20722421
DOI: 10.1021/pr100656u · PMID: 20722421 · PMCID: PMC2996461

13. **Inhibition of protein carbamylation in urea solution using ammonium-containing buffers.**
Shisheng Sun, Jian-Ying Zhou, Weiming Yang, Hui Zhang
Analytical biochemistry (2013-10-23) https://www.ncbi.nlm.nih.gov/pubmed/24161613
DOI: 10.1016/j.ab.2013.10.024 · PMID: 24161613 · PMCID: PMC4072244

14. **Fast and Sensitive Total Protein and Peptide Assays for Proteomic Analysis**
Jacek R Wiśniewski, Fabienne Z Gaugaz
Analytical Chemistry (2015-04-09) https://doi.org/f3nsk2
DOI: 10.1021/ac504689z · PMID: 25837572

15. **Getting intimate with trypsin, the leading protease in proteomics**
Elien Vandermarliere, Michael Mueller, Lennart Martens
Mass Spectrometry Reviews (2013-06-15) https://doi.org/gn64qb
DOI: 10.1002/mas.21376 · PMID: 23775586

16. **Value of using multiple proteases for large-scale mass spectrometry-based proteomics.**
Danielle L Swaney, Craig D Wenger, Joshua J Coon
Journal of proteome research (2010-03-05) https://www.ncbi.nlm.nih.gov/pubmed/20113005
DOI: 10.1021/pr900863u · PMID: 20113005 · PMCID: PMC2833215

17. **Proteomics beyond trypsin.**
Liana Tsiatsiani, Albert JR Heck
The FEBS journal (2015-04-14) https://www.ncbi.nlm.nih.gov/pubmed/25823410
DOI: 10.1111/febs.13287 · PMID: 25823410

18. Jesse G Meyer
ISRN computational biology (2014-04-22) https://www.ncbi.nlm.nih.gov/pubmed/30687733
DOI: 10.1155/2014/960902 · PMID: 30687733 · PMCID: PMC6347401

19. **In SilicoProteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage**
Jesse G Meyer
ISRN Computational Biology (2014-04-22) https://doi.org/gb6s2r
DOI: 10.1155/2014/960902 · PMID: 30687733 · PMCID: PMC6347401

20. **Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS.**
Gargi Choudhary, Shiaw-Lin Wu, Paul Shieh, William S Hancock
Journal of proteome research https://www.ncbi.nlm.nih.gov/pubmed/12643544

DOI: 10.1021/pr025557n · PMID: 12643544

21. **Six alternative proteases for mass spectrometry-based proteomics beyond trypsin.**
Piero Giansanti, Liana Tsiatsiani, Teck Yew Low, Albert JR Heck
Nature protocols (2016-04-28) https://www.ncbi.nlm.nih.gov/pubmed/27123950
DOI: 10.1038/nprot.2016.057 · PMID: 27123950

22. **Combination of Proteogenomics with Peptide**
B Blank-Landeshammer, I Teichert, R Märker, M Nowrousian, U Kück, A Sickmann
mBio (2019-10-15) https://www.ncbi.nlm.nih.gov/pubmed/31615963
DOI: 10.1128/mbio.02367-19 · PMID: 31615963 · PMCID: PMC6794485

23. **Precision**
Hao Yang, Yan-Chang Li, Ming-Zhi Zhao, Fei-Lin Wu, Xi Wang, Wei-Di Xiao, Yi-Hao Wang, Jun-Ling Zhang, Fu-Qiang Wang, Feng Xu, … Ping Xu
Molecular & cellular proteomics : MCP (2019-01-08)
https://www.ncbi.nlm.nih.gov/pubmed/30622160
DOI: 10.1074/mcp.tir118.000918 · PMID: 30622160 · PMCID: PMC6442358

24. **Mass spectrometry-assisted venom profiling of Hypnale hypnale found in the Western Ghats of India incorporating de novo sequencing approaches.**
Muralidharan Vanuopadath, Nithin Sajeev, Athira Radhamony Murali, Nayana Sudish, Nithya Kangosseri, Ivy Rose Sebastian, Nidhi Dalpatraj Jain, Amit Pal, Dileepkumar Raveendran, Bipin Gopalakrishnan Nair, Sudarslal Sadasivan Nair
International journal of biological macromolecules (2018-07-07)
https://www.ncbi.nlm.nih.gov/pubmed/29990557
DOI: 10.1016/j.ijbiomac.2018.07.016 · PMID: 29990557

25. **Venomics and antivenomics of Indian spectacled cobra (Naja naja) from the Western Ghats**
Muralidharan Vanuopadath, Dileepkumar Raveendran, Bipin Gopalakrishnan Nair, Sudarslal Sadasivan Nair
Acta Tropica (2022-04) https://doi.org/gpbzf7
DOI: 10.1016/j.actatropica.2022.106324 · PMID: 35093326

26. **Sequencing-Grade De novo Analysis of MS/MS Triplets (CID/HCD/ETD) From Overlapping Peptides**
Adrian Guthals, Karl R Clauser, Ari M Frank, Nuno Bandeira
Journal of Proteome Research (2013-05-30) https://doi.org/f47kqd
DOI: 10.1021/pr400173d · PMID: 23679345 · PMCID: PMC4591044

27. **Multiple-Enzyme-Digestion Strategy Improves Accuracy and Sensitivity of Label- and Standard-Free Absolute Quantification to a Level That Is Achievable by Analysis with Stable Isotope-Labeled Standard Spiking.**
Jacek R Wiśniewski, Christine Wegler, Per Artursson
Journal of proteome research (2018-10-30) https://www.ncbi.nlm.nih.gov/pubmed/30336047
DOI: 10.1021/acs.jproteome.8b00549 · PMID: 30336047

28. **Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data**
Rachel M Miller, Robert J Millikin, Connor V Hoffmann, Stefan K Solntsev, Gloria M Sheynkman, Michael R Shortreed, Lloyd M Smith
Journal of Proteome Research (2019-08-04) https://doi.org/gpqmp2
DOI: 10.1021/acs.jproteome.9b00330 · PMID: 31378069 · PMCID: PMC6733628

29. **Expanding Proteome Coverage with Orthogonal-specificity α-Lytic Proteases**

Jesse G Meyer, Sangtae Kim, David A Maltby, Majid Ghassemian, Nuno Bandeira, Elizabeth A Komives
Molecular & Cellular Proteomics (2014-03) https://doi.org/f5vgcg
DOI: 10.1074/mcp.m113.034710 · PMID: 24425750 · PMCID: PMC3945911

30. **Multi-protease Approach for the Improved Identification and Molecular Characterization of Small Proteins and Short Open Reading Frame-Encoded Peptides**
Philipp T Kaulich, Liam Cassidy, Jürgen Bartel, Ruth A Schmitz, Andreas Tholey
Journal of Proteome Research (2021-03-24) https://doi.org/gpqmpz
DOI: 10.1021/acs.jproteome.1c00115 · PMID: 33760615

31. **A Multiple Protease Strategy to Optimise the Shotgun Proteomics of Mature Medicinal Cannabis Buds**
Delphine Vincent, Vilnis Ezernieks, Simone Rochfort, German Spangenberg
International Journal of Molecular Sciences (2019-11-11) https://doi.org/gpqmp3
DOI: 10.3390/ijms20225630 · PMID: 31717952 · PMCID: PMC6888629

32. **Confetti: A Multiprotease Map of the HeLa Proteome for Comprehensive Proteomics**
Xiaofeng Guo, David C Trudgian, Andrew Lemoff, Sivaramakrishna Yadavalli, Hamid Mirzaei
Molecular & Cellular Proteomics (2014-06) https://doi.org/f56bwx
DOI: 10.1074/mcp.m113.035170 · PMID: 24696503 · PMCID: PMC4047476

33. **An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas**
Piero Giansanti, Thin Thin Aye, Henk van den Toorn, Mao Peng, Bas van Breukelen, Albert JR Heck
Cell Reports (2015-06) https://doi.org/gpqmpx
DOI: 10.1016/j.celrep.2015.05.029 · PMID: 26074081

34. **Use of endoproteinase Lys-C from Lysobacter enzymogenes in protein sequence analysis.**
PA Jekel, WJ Weijer, JJ Beintema
Analytical biochemistry (1983-10-15) https://www.ncbi.nlm.nih.gov/pubmed/6359954
DOI: 10.1016/0003-2697(83)90308-1 · PMID: 6359954

35. **Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion.**
Timo Glatter, Christina Ludwig, Erik Ahrné, Ruedi Aebersold, Albert JR Heck, Alexander Schmidt
Journal of proteome research (2012-10-16) https://www.ncbi.nlm.nih.gov/pubmed/23017020
DOI: 10.1021/pr300273g · PMID: 23017020

36. **The alpha-lytic protease gene of Lysobacter enzymogenes. The nucleotide sequence predicts a large prepro-peptide with homology to pro-peptides of other chymotrypsin-like enzymes.**
DM Epstein, PC Wensink
The Journal of biological chemistry (1988-11-15)
https://www.ncbi.nlm.nih.gov/pubmed/3053694
PMID: 3053694

37. **Site-specific identification and quantitation of endogenous SUMO modifications under native conditions.**
Ryan J Lumpkin, Hongbo Gu, Yiying Zhu, Marilyn Leonard, Alla S Ahmad, Karl R Clauser, Jesse G Meyer, Eric J Bennett, Elizabeth A Komives
Nature communications (2017-10-27) https://www.ncbi.nlm.nih.gov/pubmed/29079793
DOI: 10.1038/s41467-017-01271-3 · PMID: 29079793 · PMCID: PMC5660086

38. **Purification and properties of an extracellular protease of Staphylococcus aureus.**
GR Drapeau, Y Boily, J Houmard
The Journal of biological chemistry (1972-10-25)
https://www.ncbi.nlm.nih.gov/pubmed/4627743
PMID: 4627743

39. **Mildly acidic conditions eliminate deamidation artifact during proteolysis: digestion with endoprotease Glu-C at pH 4.5.**
Shanshan Liu, Kevin Ryan Moulton, Jared Robert Auclair, Zhaohui Sunny Zhou
Amino acids (2016-01-09) https://www.ncbi.nlm.nih.gov/pubmed/26748652
DOI: 10.1007/s00726-015-2166-z · PMID: 26748652 · PMCID: PMC4795971

40. **Specificity of endoproteinase Asp-N (Pseudomonas fragi): cleavage at glutamyl residues in two proteins.**
D Ingrosso, AV Fowler, J Bleibaum, S Clarke
Biochemical and biophysical research communications (1989-08-15)
https://www.ncbi.nlm.nih.gov/pubmed/2669754
DOI: 10.1016/0006-291x(89)90848-6 · PMID: 2669754

41. **Chymotrypsin: molecular and catalytic properties.**
W Appel
Clinical biochemistry (1986-12) https://www.ncbi.nlm.nih.gov/pubmed/3555886
DOI: 10.1016/s0009-9120(86)80002-9 · PMID: 3555886

42. **Mass-spectrometry-based draft of the human proteome.**
Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, … Bernhard Kuster
Nature (2014-05-29) https://www.ncbi.nlm.nih.gov/pubmed/24870543
DOI: 10.1038/nature13319 · PMID: 24870543

43. **Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics.**
Xiaofeng Guo, David C Trudgian, Andrew Lemoff, Sivaramakrishna Yadavalli, Hamid Mirzaei
Molecular & cellular proteomics : MCP (2014-04-02)
https://www.ncbi.nlm.nih.gov/pubmed/24696503
DOI: 10.1074/mcp.m113.035170 · PMID: 24696503 · PMCID: PMC4047476

44. **Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis.**
Teck Yew Low, Sebastiaan van Heesch, Henk van den Toorn, Piero Giansanti, Alba Cristobal, Pim Toonen, Sebastian Schafer, Norbert Hübner, Bas van Breukelen, Shabaz Mohammed, … Victor Guryev
Cell reports (2013-11-27) https://www.ncbi.nlm.nih.gov/pubmed/24290761
DOI: 10.1016/j.celrep.2013.10.041 · PMID: 24290761

45. **Protease bias in absolute protein quantitation.**
Mao Peng, Nadia Taouatas, Salvatore Cappadona, Bas van Breukelen, Shabaz Mohammed, Arjen Scholten, Albert JR Heck
Nature methods (2012-05-30) https://www.ncbi.nlm.nih.gov/pubmed/22669647
DOI: 10.1038/nmeth.2031 · PMID: 22669647

46. **Studies on the active site of clostripain. The specific inactivation by the chloromethyl ketone derived from -N-tosyl-L-lysine.**
WH Porter, LW Cunningham, WM Mitchell

The Journal of biological chemistry (1971-12-25)
https://www.ncbi.nlm.nih.gov/pubmed/4332560
PMID: 4332560

47. **LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification.**
Pitter F Huesgen, Philipp F Lange, Lindsay D Rogers, Nestor Solis, Ulrich Eckhard, Oded Kleifeld, Theodoros Goulas, FXavier Gomis-Rüth, Christopher M Overall
Nature methods (2014-11-24) https://www.ncbi.nlm.nih.gov/pubmed/25419962
DOI: 10.1038/nmeth.3177 · PMID: 25419962

48. **Proteomic analyses using Grifola frondosa metalloendoprotease Lys-N.**
Laura Hohmann, Carly Sherwood, Ashley Eastham, Amelia Peterson, Jimmy K Eng, James S Eddes, David Shteynberg, Daniel B Martin
Journal of proteome research (2009-03) https://www.ncbi.nlm.nih.gov/pubmed/19195997
DOI: 10.1021/pr800774h · PMID: 19195997 · PMCID: PMC2798736

49. **Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase.**
Nadia Taouatas, Madalina M Drugan, Albert JR Heck, Shabaz Mohammed
Nature methods (2008-04-20) https://www.ncbi.nlm.nih.gov/pubmed/18425140
DOI: 10.1038/nmeth.1204 · PMID: 18425140

50. **Cleavage specificities of the brother and sister proteases Lys-C and Lys-N.**
Reinout Raijmakers, Pieter Neerincx, Shabaz Mohammed, Albert JR Heck
Chemical communications (Cambridge, England) (2010-10-18)
https://www.ncbi.nlm.nih.gov/pubmed/20953479
DOI: 10.1039/c0cc02523b · PMID: 20953479

51. **A history of pepsin and related enzymes.**
Joseph S Fruton
The Quarterly review of biology (2002-06) https://www.ncbi.nlm.nih.gov/pubmed/12089768
DOI: 10.1086/340729 · PMID: 12089768

52. **CRYSTALLINE PEPSIN : I. ISOLATION AND TESTS OF PURITY.**
JH Northrop
The Journal of general physiology (1930-07-20)
https://www.ncbi.nlm.nih.gov/pubmed/19872561
DOI: 10.1085/jgp.13.6.739 · PMID: 19872561 · PMCID: PMC2141071

53. **CRYSTALLINE PEPSIN : II. GENERAL PROPERTIES AND EXPERIMENTAL METHODS.**
JH Northrop
The Journal of general physiology (1930-07-20)
https://www.ncbi.nlm.nih.gov/pubmed/19872562
DOI: 10.1085/jgp.13.6.767 · PMID: 19872562 · PMCID: PMC2141088

54. **CRYSTALLINE PEPSIN.**
JH Northrop
Science (New York, N.Y.) (1929-05-31) https://www.ncbi.nlm.nih.gov/pubmed/17758437
DOI: 10.1126/science.69.1796.580 · PMID: 17758437

55. **The Nobel Prize in Chemistry 1946**
NobelPrize.org
https://www.nobelprize.org/prizes/chemistry/1946/speedread/

56. **Protein disulfide bond determination by mass spectrometry.**
Jeffrey J Gorman, Tristan P Wallis, James J Pitt
Mass spectrometry reviews https://www.ncbi.nlm.nih.gov/pubmed/12476442

DOI: 10.1002/mas.10025 · PMID: 12476442

57. **Facilitating protein disulfide mapping by a combination of pepsin digestion, electron transfer higher energy dissociation (EThcD), and a dedicated search algorithm SlinkS.**
Fan Liu, Bas van Breukelen, Albert JR Heck
Molecular & cellular proteomics : MCP (2014-06-30)
https://www.ncbi.nlm.nih.gov/pubmed/24980484
DOI: 10.1074/mcp.o114.039057 · PMID: 24980484 · PMCID: PMC4189002

58. **Online, High-Pressure Digestion System for Protein Characterization by Hydrogen/Deuterium Exchange and Mass Spectrometry**
Lisa M Jones, Hao Zhang, Ilan Vidavsky, Michael L Gross
Analytical Chemistry (2010-01-22) https://doi.org/b993rm
DOI: 10.1021/ac902477u · PMID: 20095571 · PMCID: PMC2826105

59. **Hydrogen/deuterium exchange in mass spectrometry**
Yury Kostyukevich, Thamina Acter, Alexander Zherebker, Arif Ahmed, Sunghwan Kim, Eugene Nikolaev
Mass Spectrometry Reviews (2018-03-30) https://doi.org/gffzx8
DOI: 10.1002/mas.21565 · PMID: 29603316

60. **Proteinase K from Tritirachium album Limber.**
W Ebeling, N Hennrich, M Klockow, H Metz, HD Orth, H Lang
European journal of biochemistry (1974-08-15) https://www.ncbi.nlm.nih.gov/pubmed/4373242
DOI: 10.1111/j.1432-1033.1974.tb03671.x · PMID: 4373242

61. **Proteinase K**
W Saenger
Handbook of Proteolytic Enzymes (2013) https://doi.org/gkfkcz
DOI: https://doi.org/10.1016/b978-0-12-382219-2.00714-6 · ISBN: 9780123822192

62. **Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry.**
Simone Schopper, Abdullah Kahraman, Pascal Leuenberger, Yuehan Feng, Ilaria Piazza, Oliver Müller, Paul J Boersema, Paola Picotti
Nature protocols (2017-10-26) https://www.ncbi.nlm.nih.gov/pubmed/29072706
DOI: 10.1038/nprot.2017.100 · PMID: 29072706

63. **Simultaneous Quantification of the Acetylome and Succinylome by 'One-Pot' Affinity Enrichment**
Nathan Basisty, Jesse G Meyer, Lei Wei, Bradford W Gibson, Birgit Schilling
PROTEOMICS (2018-08-19) https://doi.org/gn4cmb
DOI: 10.1002/pmic.201800123 · PMID: 30035354 · PMCID: PMC6175148

64. **The Nobel Prize in Chemistry 2002**
NobelPrize.org
https://www.nobelprize.org/prizes/chemistry/2002/summary/

65. **Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules**
Michael Karas, Doris Bachmann, Franz Hillenkamp
Analytical Chemistry (1985-12-01) https://pubs.acs.org/doi/abs/10.1021/ac00291a042
DOI: 10.1021/ac00291a042

66. **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**
Michael Karas, Franz Hillenkamp

Analytical Chemistry (1988-10-15) https://doi.org/d577jp
DOI: 10.1021/ac00171a028 · PMID: 3239801

67. **The Scientist :: Nobel Prize controversy** (2007-05-17)
https://web.archive.org/web/20070517202246/http://cmbi.bjmu.edu.cn/news/0212/55.htm

68. **α-Cyano-4-hydroxycinnamic acid, sinapinic acid, and ferulic acid as matrices and alkylating agents for matrix-assisted laser desorption/ionization time-of-flight mass spectrometric analysis of cysteine-containing peptides**
Hongmei Yang, Debin Wan, Fengrui Song, Zhiqiang Liu, Shuying Liu
Rapid communications in mass spectrometry: RCM (2013-06-30)
https://pubmed.ncbi.nlm.nih.gov/23681820
DOI: 10.1002/rcm.6587

69. **The Desorption Process in MALDI**
Klaus Dreisewerd
Chemical Reviews (2003-01-24) https://doi.org/cpzqmq
DOI: 10.1021/cr010375i · PMID: 12580636

70. **Matrix Dependence of Metastable Fragmentation of Glycoproteins in MALDI TOF Mass Spectrometry**
Michael Karas, Ute Bahr, Kerstin Strupat, Franz Hillenkamp, Anthony Tsarbopoulos, Birendra N Pramanik
Analytical Chemistry (1995-02-01) https://doi.org/b54gnt
DOI: 10.1021/ac00099a029

71. **XX. On the equilibrium of liquid conducting masses charged with electricity**
Lord Rayleigh
The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science (1882-09)
https://doi.org/c6bp6h
DOI: 10.1080/14786448208628425

72. **Charge state coalescence during electrospray ionization improves peptide identification by tandem mass spectrometry.**
Jesse G Meyer, Elizabeth A Komives
Journal of the American Society for Mass Spectrometry (2012-05-18)
https://www.ncbi.nlm.nih.gov/pubmed/22610994
DOI: 10.1007/s13361-012-0404-0 · PMID: 22610994 · PMCID: PMC6345509

73. **Electrospray: from ions in solution to ions in the gas phase, what we know now.**
Paul Kebarle, Udo H Verkerk
Mass spectrometry reviews https://www.ncbi.nlm.nih.gov/pubmed/19551695
DOI: 10.1002/mas.20247 · PMID: 19551695

74. **Unraveling the mechanism of electrospray ionization.**
Lars Konermann, Elias Ahadi, Antony D Rodriguez, Siavash Vahidi
Analytical chemistry (2012-11-20) https://www.ncbi.nlm.nih.gov/pubmed/23134552
DOI: 10.1021/ac302789c · PMID: 23134552

75. **Interpretation of shotgun proteomic data: the protein inference problem.**
Alexey I Nesvizhskii, Ruedi Aebersold
Molecular & cellular proteomics : MCP (2005-07-11)
https://www.ncbi.nlm.nih.gov/pubmed/16009968
DOI: 10.1074/mcp.r500012-mcp200 · PMID: 16009968

76. **In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics**
Enrique Audain, Julian Uszkoreit, Timo Sachsenberg, Julianus Pfeuffer, Xiao Liang, Henning Hermjakob, Aniel Sanchez, Martin Eisenacher, Knut Reinert, David L Tabb, ... Yasset Perez-Riverol
Journal of Proteomics (2017-01) https://doi.org/f9r8r6
DOI: 10.1016/j.jprot.2016.08.002 · PMID: 27498275

77. **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification**
Jürgen Cox, Matthias Mann
Nature Biotechnology (2008-11-30) https://doi.org/crn24x
DOI: 10.1038/nbt.1511 · PMID: 19029910

78. **MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics**
Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, Alexey I Nesvizhskii
Nature Methods (2017-04-10) https://doi.org/f9z6p7
DOI: 10.1038/nmeth.4256 · PMID: 28394336 · PMCID: PMC5409104

79. **Probability-based protein identification by searching sequence databases using mass spectrometry data.**
DN Perkins, DJ Pappin, DM Creasy, JS Cottrell
Electrophoresis (1999-12) https://www.ncbi.nlm.nih.gov/pubmed/10612281
DOI: 10.1002/(sici)1522-2683(19991201)20:18<3551::aid-elps3551>3.0.co;2-2 · PMID: 10612281

80. **MS-GF+ makes progress towards a universal database search tool for proteomics**
Sangtae Kim, Pavel A Pevzner
Nature Communications (2014-10-31) https://doi.org/ggkdq8
DOI: 10.1038/ncomms6277 · PMID: 25358478 · PMCID: PMC5036525

81. **A method for reducing the time required to match protein sequences with tandem mass spectra**
Robertson Craig, Ronald C Beavis
Rapid Communications in Mass Spectrometry (2003) https://doi.org/b7bgb9
DOI: 10.1002/rcm.1198 · PMID: 14558131

82. **TANDEM: matching proteins with tandem mass spectra**
R Craig, RC Beavis
Bioinformatics (2004-02-19) https://doi.org/cthw6n
DOI: 10.1093/bioinformatics/bth092 · PMID: 14976030

83. **Skyline: an open source document editor for creating and analyzing targeted proteomics experiments**
Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, Michael J MacCoss
Bioinformatics (2010-02-09) https://doi.org/bqx9rq
DOI: 10.1093/bioinformatics/btq054 · PMID: 20147306 · PMCID: PMC2844992

84. **DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput**
Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, Markus Ralser
Nature Methods (2019-11-25) https://doi.org/gj9xgj
DOI: 10.1038/s41592-019-0638-x · PMID: 31768060 · PMCID: PMC6949130

85. **Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry**
Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, Ming Li
Nature Methods (2018-12-20) https://doi.org/gftvmn
DOI: 10.1038/s41592-018-0260-3 · PMID: 30573815

86. **Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues**
Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, Saša M Miladinović, Lin-Yang Cheng, Simon Messner, Tobias Ehrenberger, Vito Zanotelli, Yulia Butscheid, Claudia Escher, … Lukas Reiter
Molecular & Cellular Proteomics (2015-05) https://doi.org/f7b76h
DOI: 10.1074/mcp.m114.044305 · PMID: 25724911 · PMCID: PMC4424408

87. **PeptideShaker enables reanalysis of MS-derived proteomics data sets**
Marc Vaudel, Julia M Burkhart, René P Zahedi, Eystein Oveland, Frode S Berven, Albert Sickmann, Lennart Martens, Harald Barsnes
Nature Biotechnology (2015-01) https://doi.org/ggkds8
DOI: 10.1038/nbt.3109 · PMID: 25574629

88. **PeptideShaker Online: A User-Friendly Web-Based Framework for the Identification of Mass Spectrometry-Based Proteomics Data**
Yehia Mokhtar Farag, Carlos Horro, Marc Vaudel, Harald Barsnes
Journal of Proteome Research (2021-10-28) https://doi.org/gpdd85
DOI: 10.1021/acs.jproteome.1c00678 · PMID: 34709836 · PMCID: PMC8650087

89. **mzML—a Community Standard for Mass Spectrometry Data**
Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpp, Steffen Neumann, Angel D Pizarro, … Eric W Deutsch
Molecular & Cellular Proteomics (2011-01) https://doi.org/dxkg99
DOI: 10.1074/mcp.r110.000133 · PMID: 20716697 · PMCID: PMC3013463

90. **Mass spectrometer output file format mzML.**
Eric W Deutsch
Methods in molecular biology (Clifton, N.J.) (2010)
https://www.ncbi.nlm.nih.gov/pubmed/20013381
DOI: 10.1007/978-1-60761-444-9_22 · PMID: 20013381 · PMCID: PMC3073315

91. **File Formats Commonly Used in Mass Spectrometry Proteomics**
Eric W Deutsch
Molecular & Cellular Proteomics (2012-12) https://doi.org/ggkdvv
DOI: 10.1074/mcp.r112.019695 · PMID: 22956731 · PMCID: PMC3518119

92. **Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows**
Kenneth Verheggen, Helge Ræder, Frode S Berven, Lennart Martens, Harald Barsnes, Marc Vaudel
Mass Spectrometry Reviews (2020-05) https://doi.org/gbwkmf
DOI: 10.1002/mas.21543 · PMID: 28902424

93. **Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases**
Lukas Käll, John D Storey, Michael J MacCoss, William Stafford Noble

Journal of Proteome Research (2008-01) https://doi.org/fbxhxp
DOI: 10.1021/pr700600n · PMID: 18067246

94. **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.**
Joshua E Elias, Steven P Gygi
Nature methods (2007-03) https://www.ncbi.nlm.nih.gov/pubmed/17327847
DOI: 10.1038/nmeth1019 · PMID: 17327847

95. **False Discovery Rate Estimation in Proteomics**
Suruchi Aggarwal, Amit Kumar Yadav
Methods in Molecular Biology (2016) https://doi.org/f79mzp
DOI: 10.1007/978-1-4939-3106-4_7 · PMID: 26519173

96. **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.**
Carson Holt, Mark Yandell
BMC bioinformatics (2011-12-22) https://www.ncbi.nlm.nih.gov/pubmed/22192575
DOI: 10.1186/1471-2105-12-491 · PMID: 22192575 · PMCID: PMC3280279

97. **The NCBI Eukaryotic Genome Annotation Pipeline**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

98. **Biocuration in UniProt** https://www.uniprot.org/help/biocuration

99. https://www.psidev.info/sites/default/files/2018-03/MIAPE_MSI_1.1.pdf

100. **Guidelines for reporting quantitative mass spectrometry based experiments in proteomics.**
Salvador Martínez-Bartolomé, Eric W Deutsch, Pierre-Alain Binz, Andrew R Jones, Martin Eisenacher, Gerhard Mayer, Alex Campos, Francesc Canals, Joan-Josep Bech-Serra, Montserrat Carrascal, … Juan P Albar
Journal of proteomics (2013-03-14) https://www.ncbi.nlm.nih.gov/pubmed/23500130
DOI: 10.1016/j.jprot.2013.02.026 · PMID: 23500130

101. **UniProt: the Universal Protein knowledgebase.**
Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, … Lai-Su L Yeh
Nucleic acids research (2004-01-01) https://www.ncbi.nlm.nih.gov/pubmed/14681372
DOI: 10.1093/nar/gkh131 · PMID: 14681372 · PMCID: PMC308865

102. **Where do the UniProtKB protein sequences come from?**
https://www.uniprot.org/help/sequence_origin

103. **How do we manually annotate a UniProtKB entry?**
https://www.uniprot.org/help/manual_curation

104. **GitHub - pwilmart/fasta_utilities: Utilities for downloading and managing protein FASTA files.**
GitHub
https://github.com/pwilmart/fasta_utilities

105. **Global Identification of Protein Post-translational Modifications in a Single-Pass Database Search.**
Michael R Shortreed, Craig D Wenger, Brian L Frey, Gloria M Sheynkman, Mark Scalf, Mark P Keller, Alan D Attie, Lloyd M Smith

Journal of proteome research (2015-09-29) https://www.ncbi.nlm.nih.gov/pubmed/26418581
DOI: 10.1021/acs.jproteome.5b00599 · PMID: 26418581 · PMCID: PMC4642219

106. **NCBI Datasets**
NCBI
https://www.ncbi.nlm.nih.gov/datasets/

107. **Eukaryotic genomes annotated at NCBI**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/

108. **Prokaryotic RefSeq Genomes** https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

109. **Eukaryotic RefSeq Genome Annotation Status**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/status/

110. **Training & Tutorials - Site Guide - NCBI** https://www.ncbi.nlm.nih.gov/guide/training-tutorials/

111. **Proteomics in Non-model Organisms: A New Analytical Frontier.**
Michelle Heck, Benjamin A Neely
Journal of proteome research (2020-08-20) https://www.ncbi.nlm.nih.gov/pubmed/32786681
DOI: 10.1021/acs.jproteome.0c00448 · PMID: 32786681 · PMCID: PMC7874939

112. **Ensembl genome browser 106** https://uswest.ensembl.org/index.html

113. **Ensembl Genomes** http://ensemblgenomes.org/

114. **Interferences and contaminants encountered in modern mass spectrometry.**
Bernd O Keller, Jie Sui, Alex B Young, Randy M Whittal
Analytica chimica acta (2008-04-25) https://www.ncbi.nlm.nih.gov/pubmed/18790129
DOI: 10.1016/j.aca.2008.04.043 · PMID: 18790129

115. **cRAP protein sequences** https://www.thegpm.org/crap/

116. **Philosopher: a versatile toolkit for shotgun proteomics data analysis.**
Felipe da Veiga Leprevost, Sarah E Haynes, Dmitry M Avtonomov, Hui-Yin Chang, Avinash K Shanmugam, Dattatreya Mellacheruvu, Andy T Kong, Alexey I Nesvizhskii
Nature methods (2020-09) https://www.ncbi.nlm.nih.gov/pubmed/32669682
DOI: 10.1038/s41592-020-0912-y · PMID: 32669682 · PMCID: PMC7509848

117. **How do Protein Contaminants Influence DDA and DIA Proteomics**
Ashley M Frankenfield, Jiawei Ni, Mustafa Ahmed, Ling Hao
Cold Spring Harbor Laboratory (2022-04-28) https://doi.org/gp4xcp
DOI: 10.1101/2022.04.27.489766

118. **The minotaur proteome: avoiding cross-species identifications deriving from bovine serum in cell culture models.**
Jakob Bunkenborg, Guadalupe Espadas García, Marcia Ivonne Peña Paz, Jens S Andersen, Henrik Molina
Proteomics (2010-08) https://www.ncbi.nlm.nih.gov/pubmed/20641139
DOI: 10.1002/pmic.201000103 · PMID: 20641139

119. **Foetal bovine serum influence on in vitro extracellular vesicle analyses.**
Brandon M Lehrich, Yaxuan Liang, Massimo S Fiandaca
Journal of extracellular vesicles (2021-01-25) https://www.ncbi.nlm.nih.gov/pubmed/33532042
DOI: 10.1002/jev2.12061 · PMID: 33532042 · PMCID: PMC7830136

120. **The CRAPome: a contaminant repository for affinity purification-mass spectrometry data.**
Dattatreya Mellacheruvu, Zachary Wright, Amber L Couzens, Jean-Philippe Lambert, Nicole A St-Denis, Tuo Li, Yana V Miteva, Simon Hauri, Mihaela E Sardiu, Teck Yew Low, ... Alexey I Nesvizhskii
Nature methods (2013-07-07) https://www.ncbi.nlm.nih.gov/pubmed/23921808
DOI: 10.1038/nmeth.2557 · PMID: 23921808 · PMCID: PMC3773500

121. **Common Repository of FBS Proteins (cRFP) To Be Added to a Search Database for Mass Spectrometric Analysis of Cell Secretome.**
Jihye Shin, Yumi Kwon, Seonjeong Lee, Seungjin Na, Eun Young Hong, Shinyeong Ju, Hyun-Gyo Jung, Prashant Kaushal, Sungho Shin, Ji Hyun Back, ... Cheolju Lee
Journal of proteome research (2019-09-10) https://www.ncbi.nlm.nih.gov/pubmed/31475827
DOI: 10.1021/acs.jproteome.9b00475 · PMID: 31475827

122. **Accurately Assigning Peptides to Spectra When Only a Subset of Peptides Are Relevant.**
Andy Lin, Deanna L Plubell, Uri Keich, William S Noble
Journal of proteome research (2021-07-08) https://www.ncbi.nlm.nih.gov/pubmed/34236864
DOI: 10.1021/acs.jproteome.1c00483 · PMID: 34236864 · PMCID: PMC8489664

123. **Averaging Strategy To Reduce Variability in Target-Decoy Estimates of False Discovery Rate.**
Uri Keich, Kaipo Tamura, William Stafford Noble
Journal of proteome research (2019-01-03) https://www.ncbi.nlm.nih.gov/pubmed/30560673
DOI: 10.1021/acs.jproteome.8b00802 · PMID: 30560673 · PMCID: PMC6919216

124. **Mass spectrometrists should search only for peptides they care about.**
William Stafford Noble
Nature methods (2015-07) https://www.ncbi.nlm.nih.gov/pubmed/26125591
DOI: 10.1038/nmeth.3450 · PMID: 26125591 · PMCID: PMC4711994

125. **Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data.**
Dhirendra Kumar, Amit Kumar Yadav, Debasis Dash
Methods in molecular biology (Clifton, N.J.) (2017)
https://www.ncbi.nlm.nih.gov/pubmed/27975281
DOI: 10.1007/978-1-4939-6740-7_3 · PMID: 27975281